BIOINFORMATIC IDENTIFICATION OF GENOMIC ALTERATIONS IN BREAST CANCER

Henrik Edgren

Institute for Molecular Medicine Finland
University of Helsinki
and
Genome Scale Biology Research Program
Department of Medical Genetics
Faculty of Medicine
University of Helsinki

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Medicine of the University of Helsinki, for public examination in the Haartman hospital lecture hall, 4th floor, on September 14th 2012, at 12 noon.

Helsinki 2012

Supervised by
Olli Kallioniemi, MD, PhD
Professor, Director
Institute for Molecular Medicine Finland
University of Helsinki, Finland

Reviewed by
Minna Tanner, MD, PhD
Adjunct Professor in Clinical Oncology
Department of Oncology
Tampere University Hospital, Finland

And

Garry Wong, PhD
Professor of Molecular Bioinformatics
A.I. Virtanen Institute for Molecular Sciences
University of Eastern Finland, Finland

Official Opponent
Zoltan Szallasi, MD, PhD
Professor, Center for Biological Sequence Analysis
Department of Systems Biology
Technical University of Denmark, Denmark
and
Children's Hospital Informatics Program
Harvard-MIT Division of Health Sciences and Technology
Harvard Medical School, Boston, Massachusetts

ISBN 978-952-10-8150-7 (paperback)
ISBN 978-952-10-8151-4 (PDF)
http://ethesis.helsinki.fi
Helsinki University Printing House
Helsinki 2012

TABLE OF CONTENTS

1 LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following three publications (referred to in the text by their Roman numerals I-III).

I. Muggerud AA*, Edgren H*, Wolf M, Kleivi K, Dejeux E, Tost J, Sorlie T, Kallioniemi O. Data integration from two microarray platforms identifies bi-allelic genetic inactivation of *RIC8A* in a breast cancer cell line. BMC Med Genomics. 2009 May 11;2:26.

II. Edgren H*, Murumagi A*, Kangaspeska S*, Nicorici D, Hongisto V, Kleivi K, Rye IH, Nyberg S, Wolf M, Borresen-Dale AL, Kallioniemi O. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol. 2011 Jan 19;12(1):R6. [Epub ahead of print]

III. Kleivi Sahlberg K*, Hongisto V*, Edgren H*, Mäkelä R, Hellström K, Due EU, Moen Vollan HK, Sahlberg N, Wolf M, Børresen-Dale AL, Perälä M, Kallioniemi O. The HER2 amplicon includes several genes required for the growth and survival of HER2 positive breast cancer cells. Submitted.

* Equal contribution.

Publication I has been part of another thesis (ISBN 978-82-8072-336-9)

2 ABBREVIATIONS

aCGH	array Comparative Genomic Hybridization
BAC	bacterial artificial chromosome
bp	base pair
cDNA	complementary DNA
CIN	chromosomal instability
cPARP	cleaved poly (ADP-ribose) polymerase protein
CTG	CellTiter-Glo
DCIS	ductal carcinoma in situ
DSB	double stranded break
EC50	half maximal effective concentration
ER	estrogen receptor
FISH	fluorescent *in situ* hybridization
GEO	Gene Expression Omnibus
GINI	gene identification by nonsense inhibition
kb	kilobase, i.e. 1000 base pairs
LMA	lysate microarray
Mb	mega base pair, i.e. one million base pairs
mRNA	messenger RNA
NMD	Nonsense-mediated mRNA decay
pAKT	phosphorylated v-akt murine thymoma viral oncogene homolog protein
PCR	polymerase chain reaction
PR	progesterone receptor
pS6K	phosphorylated p70-S6 kinase
PTC	premature termination codon
RMA	robust multiarray average
RNA-seq	RNA-sequencing
RPKM	reads per kilobase per million aligned sequences
rRNA	ribosomal RNA
TNM	tumor, lymph node, metastasis classification of malignant tumors
TSG	tumor suppressor gene
TSS	transcription start site
UTR	untranslated region
wt	wildtype

3 ABSTRACT

Cancer is a disease characterized by the accumulation of somatic cellular aberrations, whether in the DNA or epigenetic changes, which are inherited from cancer cell generation to generation. During the last decade, many different techniques have been developed to comprehensively characterize these changes in cancer cells, resulting in thousands of publications on various cancer types. Different types of microarrays can now measure the expression of essentially all genes or DNA copy number at up to 1 kilo base pair (kb) resolution in a tumor. Sequencing, whether targeted to specific genes or, increasingly, to all exons (exome sequencing) or whole genome sequencing, has identified genes mutated at various frequencies in many cancer types. In parallel with the development of laboratory techniques, a large variety of bioinformatic methods to analyze data from these have been developed. However, many of these concentrate on the analysis of data from only one laboratory technique, while it is becoming clear that advances in cancer research increasingly depend on integration of multiple different data types for the same tumors. Simultaneously, the recent explosive growth in sequencing data requires the development of new analythical methods. The aim of this thesis was to further characterize the genomic changes in breast cancer, with an emphasis on the development and application of bioinformatic methods to analyze and integrate data from different high throughput analysis techniques.

In the first part of this work, the Gene Identification by Nonsense Inhibition (GINI) method was applied to identify potential tumor suppressor genes (TSGs) in breast cancer. The integration of steady state gene expression, transcript stabilization and array comparative genomic hybridization (aCGH) data for six breast cancer cell lines led to the identification of a nonsense mutation in the *RIC8A* gene located at 11p15, a region deleted in ~15% of breast tumors. Despite being unable to identify further mutations or methylation of *RIC8A* in tumors, low *RIC8A* expression was shown to be associated with estrogen (ER) and progesterone receptor (PR) negative tumors as well as loss of *TP53*. This suggests loss of *RIC8A* expression may be important in a subgroup of aggressive breast cancers.

When study II was started, only a few papers describing fusion gene identification using RNA-sequencing (RNA-seq) data had been published, and all suffered from a high rate of false positive findings, requiring extensive post-sequencing validation. In this study, we developed a bioinformatic method for highly specific fusion gene identification from paired-end RNA-seq data. Application of the bioinformatic pipeline to four breast cancer cell lines led to the identification of 24 novel and three previously published fusion genes, with 95% specificity. In addition to showing that fusion genes are more prevalent in breast cancer than previously thought, several biological characteristics of fusion genes were identified. Most prominently, fusion genes were frequently associated with DNA copy number transitions, particularly high level amplifications, suggesting that most of them are not generated by balanced rearrangements. siRNA knock-down studies furthermore provided evidence for the functional importance of the *VAPB-IKZF3* fusion gene in the BT-474 cell line.

In the final study, we used aCGH to characterize the size distribution of the *ERBB2* amplicon across 71 amplicon carrying tumors and 10 cell lines. The minimal common region of amplification in the tumors was 78.61kb, and included the genes *STARD3*, *TCAP*, *PNMT*, *PERLD1*, *ERBB2* and *MIEN1* (*C17orf37*). To study the possible contribution to cancer of other coamplified genes in the amplicon, 23 genes amplified in 60% of tumors were selected for siRNA screening in two trastuzumab sensitive, two insensitive and one control cell line. In addition to single gene siRNA silencing experiments, *PPP1R1B*, *STARD3*, *PERLD1*, *GRB7* and *PSMD3* were knocked-down together with *ERBB2* to identify synergistic effects. In all *ERBB2* positive cell lines, for instance, simultaneous silencing of *ERBB2* and *PPP1R1B* led to a synergistic inhibition of the Akt pathway, as measured by phosphorylated AKT (pAKT) and phosphorylated S6-kinase (pS6K). Silencing of *PPP1R1B* alone had no effect on pAKT or pS6K in any of the cell lines. Silencing of several other genes, either alone or in combination with *ERBB2*, was also found to have an effect on several endpoints. These results suggest that cancer cells may be dependent on a number of genes in an amplicon besides the primary driver oncogene, a phenomenon termed non-oncogene addiction.

4 INTRODUCTION

Breast cancer is the most common cancer in women in Finland, with 4475 new cases diagnosed in 2009 [1]. It can be classified in a number of ways; based on histology, grade, stage as well as the presence or absence of molecular markers. Histologically breast cancer is divided into two main types, ductal and lobular, depending on the tissue from which the tumor arises. Minor subtypes represent ~5% of breast cancers, the most common being mucinous, tubular and medullary breast cancer [2, 3]. Gene expression microarray data has also been used to subdivide breast cancer into five different subtypes; luminal A and B, basal, ERBB2 positive and normal-like, which differ from each other in terms of prognosis [4-6].

Breast cancer incidence starts rising after the age of 45, with a mean age at diagnosis of 60 [1]. Factors that increase the risk of breast cancer include low age of menarche, late onset of menopause, hormone replacement therapy in conjunction with menopause, obesity, alcohol consumption and smoking [7]. Factors that protect from breast cancer include low age at the time of the birth of the first child as well as the total number of children, breast feeding, exercise and a diet rich in vegetables [7]. Inherited mutations in genes such as *BRCA1*, *BRCA2*, *ATM*, *CHEK2* and *PALB2* also significantly increase the risk of developing breast cancer [8].

Commonly used prognostic factors after a breast cancer diagnosis include TNM status, tumor grade and size, patient age, tumor proliferation as measured by Ki67 expression, expression of molecular markers such as estrogen and progesterone receptors as well as overexpression and possible amplification of the HER2 protein [9]. In recent years, bioinformatic analyses of gene expression microarray data from large sets of breast cancers has lead to the development of multiple expression based classifiers or signatures that are able to predict various features related to the cancer. Currently, the most widely used are the gene expression based classifiers MammaPrint [10] and Oncotype DX [11], the first of which is currently undergoing further validation as part of the MINDACT prospective randomized phase III clinical trial [12]. Both of them primarily have a role in predicting outcome and consequently the benefit of adjuvant chemotherapy in patients with early stage cancer. In the MammaPrint test, patients are divided into good- and poor prognosis groups based on the correlation between the expression levels of 70 genes in their tumor and the average expression of the same genes in a set of previously profiled patients with good prognosis[10]. The Oncotype DX recurrence score, in turn, is based on the weighted sum of the normalized expression levels of 16 cancer associated genes, based on which patients are assigned a low, medium or high risk of recurrence [13]. Other classifiers which predict survival e.g. on the basis of the pattern of DNA amplifications [14] or specific gene expression profiles

[15] have also been developed. However, most of these have not entered into widespread clinical use.

The primary treatment for breast cancer is surgery to remove the tumor and possible lymph node metastases, possibly preceeded by preoperative chemotherapy. After surgery, adjuvant systemic treatments such as endocrine therapy or chemotherapy are prescribed for patients with a high risk of metastasis and radiation to those with risk of local recurrence [16-18]. Adjuvant treatment is estimated to decrease breast cancer mortality by ~30% during 15 years of followup [18]. Inhibitors of the protein product of the *ERBB2* gene, such as trastuzumab, are also used to treat patients whose tumors overexpress HER2. The survival benefit of trastuzumab is well established when it is given concurrently with chemotherapy, whereas the benefit is less clear when given subsequent to chemotherapy [19].

In Europe, breast cancer is both the most common cancer in women, accounting for 26% of all cases, as well as the leading cause of cancer mortality in women, at 17% of all cancer deaths [20]. However, the Europe-wide average age-standardized 5-year relative survival for breast cancer is 79%, higher than for several other common cancers, such as colorectal (53%), ovarian (36%) and lung (12%) cancer [21]. Breast cancer incidence has increased significantly in Finland between 1964 and 2004 [22]. During the same period, the prognosis for breast cancer has also improved considerably [22]. The reason for improved survival is believed to be improved adjuvant treatments, but also earlier diagnosis thanks to screening mammography [23, 24].

Optimal anti-cancer therapies specifically kill cancer cells, while leaving normal cells unaffected. Mutations in oncogenes that the cancer depends on for continued survival are therefore good drug targets. Inhibiting the mutated protein would block cancer cell survival and, by being unique to the cancer cell, hopefully also provide a therapeutic window. The discovery of the mutations and epigenetic changes causing cancer is therefore of great importance. Even when the mutations are not in directly druggable genes, they teach us about the biology and vulnerabilities of the cancer cells and thereby may allow us to identify other nodes in the signaling network that can be targeted. In breast cancer, the primary example of a therapy specifically targeted against a somatic genomic abnormality is trastuzumab, a monoclonal antibody recognizing the HER2 protein, used in the treatment of patients whose tumors have amplified and overexpressed *ERBB2* [25]. However, only some patients with *ERBB2* amplification respond to trastuzumab and even those who do, usually develop resistance over time. A significant fraction of breast cancers, so called triple-negative tumors, also lack both ER and PR expression as well as *ERBB2* overexpression and amplification and no molecularly targeted therapies for them exist. There is therefore a great need to more thoroughly characterize the aquired genetic changes in breast cancer, both in order to identify the mechanisms of resistance against existing drugs, as well as to develop new treatments.

5 REVIEW OF THE LITERATURE

5.1 THE MOLECULAR BIOLOGY OF BREAST CANCER

Breast cancer is a genetic disease in two different ways. On the one hand, inherited high and low penetrance mutations may predispose a person to developing breast cancer. On the other hand, and more central to this thesis, breast cancer is a genetic disease in that the tumor develops as a consequence of the accumulation of mutations in the genomes of the evolving cancer cells. These mutations range from single point mutations and copy number changes affecting only a few nucleotides up to gains, losses or rearrangements of significant portions of whole chromosomes. The common effect of these different aberrations is to activate cancer promoting genes (oncogenes) and inactivate genes that protect cells from malignant transformation (tumor suppressor genes, TSGs). A third category of genes mutated in cancer are the so-called stability or caretaker genes, although these are often classified among TSGs. Mutations in these genes do not, in themselves, lead to cancer. Rather, inactivation of these genes either directly or indirectly increases the rate at which mutations occur, e.g. through defective DNA repair, and thereby increase the likelihood of oncogenic mutations arising [26].

Several external factors are known or suspected to promote carcinogenesis. These include exposure to radiation and chemical agents, either natural or man-made [27]. Infectious agents are also known to promote cancer formation, either as direct (e.g. expression of virus derived oncogene) or indirect (chronic inflammation) carcinogens [28], [29]. In addition, inherited mutations may also predispose to specific cancer types [30], [31]. Immune suppression caused by HIV infection [32] or immunosuppressive drugs required after organ transplantation [33] also increase the likelyhood of developing specific types of cancers. Spontaneous mutations, caused e.g. by proofreading deficient DNA polymerases [34] also occur constantly in human cells, contributing to cancer formation. Most of the time, mutations are either harmless, repaired, or cause the cell to enter apoptosis, but sometimes the mutations are not eliminated and the cell starts on the road towards cancer. The role of randomness and plain bad luck should therefore not be discounted as causative factors for cancer.

Most tumors derive from a single progenitor cell and are therefore clonal. However, this does not preclude heterogeneity within the tumor, which continuously arises when different subclones follow independent developmental trajectories towards increased malignancy [35, 36]. A subclone with the greatest growth rate, i.e. largest fitness in evolutionary terms, may come to dominate the tumor mass. However, unless the physical space that the tumor occupies is restricted, e.g. by the tumor not being able to invade surrounding tissues, the clone is unlikely to outcompete and completely replace less fit subclones and heterogeneity will remain. Metastases in different parts of the body also frequently differ from the primary tumor as well as from each other in terms of mutations [37]. From the tumor's point of view, heterogeneity becomes important as it is the pool from which treatment resistant subclones may arise.

When compared to normal cells, cancer cells have aquired a number of capabilities necessary for cancer formation, the so-called "hallmarks of cancer" [38]. These hallmarks are limitless replicative potential, selfsufficiency in growth promoting signals, the ability to evade growth suppressing signals, resistance to programmed cell death, the ability to induce angiogenes and, most lethally, the capability to invade surrounding tissues and metastasize. Two additional hallmarks have recently been suggested: reprogramming of energy metabolism and the ability to evade destruction by the immune system. In essence, the developing cancer must aquire all these capabilities in order to counter and evade the safety mechanisms the body has evolved as defenses against cancer formation.

5.1.1 ONCOGENES AND TUMOR SUPPRESSOR GENES

Proto-oncogenes are normal genes that, when their function changes, cause cancer [26]. A proto-oncogene can be transformed into an active oncogene in two different ways. A mutation may change the function of the protein, typically making it continuously active, as opposed to the wildtype (wt) form, which is only active under strictly controlled circumstances. An example in breast cancer is the p110α catalytic subunit of the PI3-kinase, *PIK3CA*, which is mutated in ~30% of tumors [39]. Point mutations in *PIK3CA* cluster in the helical and kinase domains of the protein and both lead to increased kinase activity and consequent abnormal activation of the PI3K pathway [39]. The second way is to significantly increase the production of the, frequently wt, proto-oncogene protein product, thereby activating it. Several mechanisms may lead to overexpression of an oncogene. The most commonly amplified and overexpressed oncogene in breast cancer is the receptor tyrosine kinase *ERBB2* (also known as *HER2*) on chromosome 17. Amplification of *ERBB2* is observed in 20-25% of breast cancers [40, 41] and leads to overexpression of it's protein product, the HER2 protein. This in turn leads to the activation of downstream pathways, in particular the PI3K-AKT and MAP-kinase pathways [42], which are central in cancer cell growth and survival. Irrespective of how an oncogene is activated, it functions by contributing to the development one or more of the cancer hallmarks described in the previous section. Mutationally activated *PIK3CA*, for instance, enables both anchorage- and growth factor independent growth as well as protects immortalized mammary epithelial cells from anoikis, a form of programmed cell death induced by loss of contact with extracellular matrix [43, 44].

Tumor suppressor genes (TSGs) function, in the broadest sense, by restraining improper cell growth (table 1). Some, such as *RB1* and the family of cyclin dependent kinase inhibitors function in cell cycle regulation, preventing a cell from passing through specific checkpoints and completing cell division before the preconditions of each checkpoint are met [26].

Table 1. ***Tumor suppressor genes mentioned in the text and their main mechanisms of action.***

Tumor suppressor	Mechanisms of action
RB1	regulation of cell cycle
cyclin dependent kinase inhibitors	regulation of cell cycle
PTEN	regulation of growth factor signaling, genomic stability
TP53	many, including DNA damage response and regulation of apoptosis
BRCA1, *BRCA2*, *MLH1*	DNA repair

Others, such as *PTEN*, work by negatively regulating growth factor signaling. *PTEN* counteracts PI3K signaling by dephosphorylating phosphatidylinositol (3,4,5) trisphosphate, the substrate of the catalytic subunit *PIK3CA* of PI3-kinase. Loss of *PTEN* function, whether through mutation [45], deletion [46], downregulation of expression via methylation [47, 48] or inactivation of the *PTENP1* pseudogene [49] therefore releases PI3K signaling from negative feedback and increases the activity of the pathway. However, it has recently been found that *PTEN* additionally plays a role in e.g. maintaining genomic stability, seemingly independent of the PI3K pathway [50], suggesting yet another way by which *PTEN* loss may cause cancer. In breast cancer, *PTEN* protein expression is lost in 37-48% of tumors [41, 51]. Contrary to the common assumption that an oncogenic pathway typically only needs to be activated by one alteration, *PTEN* loss is not mutually exclusive with *PIK3CA* mutation [41, 52]. Rather, many breast tumors carry both mutated *PIK3CA* in addition to having lost *PTEN*. In addition to *PIK3CA*, *ERBB2* and *PTEN*, alterations in several other genes, such as *AKT1*, *EGFR*, *PDK1* and *KRAS* activate the PI3K pathway in breast cancer [53], pointing to the central importance of this pathway in breast cancer development.

Yet other tumor suppressor genes, such as *TP53*, survey the cell for signs of incorrect behaviour, such as DNA damage, and trigger repair processes or apoptosis if such is found. In breast cancer, *TP53* itself is mutated in 20-40% of tumors [54]. Additional mechanisms for *TP53* inactivation in breast cancer include e.g. amplification of *MDM2*, a negative regulator of *TP53*. A study by Miller *et al.* [55] on 251 consecutively collected breast cancers showed that 58/251 (23%) tumors had *TP53* mutations, while a gene expression microarray-based classifier developed in the study identified an additional 14

(5.6%) tumors that showed an expression profile of *TP53* pathway inactivation.

Some TSGs have their main function in e.g. DNA repair and their inactivation leads to an increased accumulation of mutations, thereby increasing the likelyhood that other oncogenic mutations will occur. In breast cancer, the two most significant predisposing genes *BRCA1* and *BRCA2* both play a role in the repair of double stranded breaks (DSBs) through homologous recombination, and mutations in them lead to genomic instability [56].

The classical model for TSG inactivation is the Knudson two-hit model [57, 58]. In this model, both alleles of a TSG must be inactivated for cancer to develop, with one wt allele being sufficient for normal cellular functions. For some TSGs, however, loss of one allele is sufficient for an altered phenotype, the phenomenon known as haploinsufficiency. Reduction in *PTEN* levels, for instance, correlates in a dose-dependent manner with prostate cancer incidence, latency and progression [59]. High penetrance cancer predisposition genes, such as *BRCA1*, *MLH1* and a large number of other cancer genes [60] are typically categorized as recessive, in that on a cellular level, both copies need to be inactivated, whereas they function dominantly on the whole organism level, i.e. inheriting one mutation is sufficient to cause disease [61]. Rare exceptions to this rule include the *RET* oncogene [62]. One possible explanation for this is that inherited activating mutations in oncogenes would potentially alter and disturb the function of all cell types expressing them and thereby not be compatible with normal development. Cells carrying one mutated TSG allele, however, would function normally and only cells in which both copies are inactivated have altered function [26].

It has recently been suggested that, in addition to oncogenes and TSGs, cancer cells may become dependent on genes that in themselves do not cause cancer, so called "non-oncogene addiction" [63, 64]. This theory proposes that the process of malignant transformation causes the cancer cells to become dependent on genes that protect them from various kinds of stress, creating a synthetic lethal interaction. The source of this stress may be excess DNA damage, protein folding related stress, metabolic stress and the general stress encountered by a cell in a solid tumor (hypoxia, mechanical stress, low nutrient levels) [63, 65]. Examples of such genes include *HSF1* [65] and *PARP1* [64]. Breast cancer cells that have lost both copies of *BRCA2* are deficient in homologous recombination. This makes the cells dependent on *PARP1* for repair of this category of DNA lesions and gives rise to a synthetic lethal interaction between *BRCA2* mutation and *PARP1* [64]. *PARP1* inhibition has been shown to be effective in treating especially *BRCA2* mutant breast cancer [66].

5.1.2 GENOMIC ALTERATIONS IN CANCER

5.1.2.1 Aneuploidy and chromosomal instability

Practically all cancer genomes are altered by different combinations of point mutations, copy number changes and chromosomal rearrangements [67] (table 2). Aneuploidy, defined as the gain or loss of whole chromosomes, is almost ubiquitous in cancer. It is frequently caused by chromosomal instability (CIN), the inability of a cell to correctly divide the chromosomes into daughter cells. In that case, aneuploidy is an evolving process, with every cell division potentially altering the chromosome composition of the daughter cells [68]. However, a cancer may also have a stable though aneuploid karyotype, with no continuing CIN [69]. In solid tumors, CIN in general is associated with poor prognosis [70]. The relationship may not be directly linear, however, as in breast cancer, tumors with the highest level of CIN have a better prognosis than those with a more moderate level of instability [71]. The manner in which aneupoloidy contributes to cancer development is still unclear [72, 73]. Aneuploidy carries a replicative penalty in both normal and cancer cells [74, 75], yet it is seen in nearly all cancers. In mice with widespread aneuploidy due to haploinsufficiency in mitotic checkpoint genes, such as *Rae1* and *Bub3*, aneuploidy has been shown to increase carcinogen induced tumorigenesis [76], even when aneuploidy itself does not increase cancer incidence [77]. Aneuploidy may also be a mechanism by which the cell can get rid of a remaining wt TSG allele or duplicate, and thereby increase the dosage of, a mutated oncogene. In addition, gain or loss of whole chromosomes or chromosome arms may contribute to cancer through dosage effects on a large number of genes, e.g. if they create additional genomic instability or provide a buffer of extra copies of essential genes, such that functional copies of them are more likely to be available even if the mutation rate is high [73]. The importance of the last hypothesis could be tested in cancer types, such as lung cancer, that are known to carry a large number of mutations due to mutagen exposure [78] or individual tumors that have a high mutation rate. If the buffering hypothesis would be true, these cancers should, on average, be more aneuploid.

Table 2. ***Genomic aberrations commonly occurring in cancer.***

Genomic alteration	Description
point mutation	A change of a single nucleotide to another.
insertion / deletion	Addition or loss of one or more consecutive nucleotides.
aneuploidy	An abnormal chromosome number. In the context of cancer, usually somatically acquired.
amplification	An increase in copy number of a genomic region.
deletion	Loss, either of one or both copies of a genomic region.
chromosomal rearrangement	A general term for various chromosomal aberrations, including inversions and translocations.
translocation	Fusion of part of one chromosome to another, non-homologous, chromosome.

Even under normal circumstances, the genome of every cell daily receives and repairs thousands of DNA lesions of various kinds [79, 80]. In most cases, the lesions are either repaired or, if their extent is too large, the cell goes into apoptosis or senescence [81]. Either way, mutations are not transmitted to daughter cells. In cancer, single nucleotide mutations, small insertions and deletions (indels) and e.g. microsatellite instability and larger genomic rearrangements, such as amplifications and translocations occur due to different types of mistakes during DNA repair [82]. Single nucleotide mutations and small indels can arise during DNA replication if the DNA polymerase makes a mistake [83] or at any other point during the cell cycle, mostly as a consequence of normal cellular metabolism. External factors, such as mutagenic chemicals and radiation can also cause DNA damage [79]. The central role of DNA damage repair in cancer formation is exemplified by the large number of tumor suppressor genes that code for proteins involved in DNA damage response.

5.1.2.2 Copy number alterations

Many solid tumors, including breast cancers, contain amplifications or deletions of genomic regions of various sizes [84]. Amplifications are thought to arise primarily through breakage-fusion-bridge cycles, caused e.g. by telomere attrition, and via the formation of double minute chromosomes [85]. High-level amplifications affect cancer development by upregulating the expression of one or more genes in the amplicon [84, 86, 87]. Examples of amplification targets include *ERBB2* and *CCND1* in breast cancer [40, 88] and the Myc family genes *MYCN*, *MYCL1* and *MYC* in a variety of cancers [89-91]. Amplicons containing multiple interacting oncogenes are also known [92, 93], and fusion genes can also be formed within or at the borders of high level amplifications [94]. Additionally, presumably independently of the specific genes amplified, the pattern of complex genomic rearrangements in a breast tumor is prognostic of outcome [95].

Homozygous deletions occur frequently in cancers and are thought to primarily affect cancer development through inactivation of TSGs. As an alternative to TSG inactivation, deletions may also lead to fusion gene formation, as has been shown for *TMPRSS2-ERG* [96]. However, as recurrent homozygous deletions also occur at fragile sites, i.e. parts of the genome that are prone to genomic breaks due to some inherent feature, homozygous loss of a gene as such is not conclusive evidence of cancer relevance. Recent studies indeed suggest that a majority of homozygous deletions occur at fragile sites [97, 98].

5.1.2.3 Translocations

Recurrent chromosomal translocations and the resulting gene fusions are well known mechanisms for oncogene activation and occur frequently in leukemias, lymphomas and sarcomas [99, 100]. Translocations form through double stranded breaks, which can be generated by e.g. immunoglobulin gene processing in B cells [101, 102], DNA damage e.g. caused by genotoxic agents or occuring during DNA replication, double stranded breaks caused by chromosome segregation errors [103], chromotripsis [104, 105] and during amplification formation [94, 106, 107]. Essentially any process that gives rise to two or more double stranded breaks that are then not repaired correctly is capable of generating translocations. In addition, the two sequences that are fused may need to be in close proximity in the nucleus [108]. The best studied example is *BCR-ABL* in chronic myelogenous leukemia [109, 110], which is formed by a translocation between chromosomes 9 and 22. The discovery of translocations involving Ets-family members in prostate cancer [111], *EML4-ALK* in lung cancer [112] and *CD44-SLC1A2* in gastric cancer [113] now suggests that fusion genes may play a more prominent role in the development of epithelial cancers than previously anticipated. In breast cancer, both primary tumors and cell lines have been found to contain fusion genes [94, 114-116], but recurrent fusions have only been known in rare subtypes, such as *ETV6-NTRK3* in secretory breast carcinoma [117] and *MYB-NFIB* in adenoid cystic carcinoma of the breast [118]. Recently, rare but recurrent rearrangements of NOTCH and MAST family genes as well as the recurrent *RPS6KB1-VMP1* fusion have been reported in breast cancer [116, 119]. Individual examples of both fusion categories have also been reported previously, e.g. *NOTCH1-NUP214* and *RPS6KB1-VMP1* (previously known as *RPS6KB1-TMEM49*) [94] and *ARID1A-MAST2* [114]. However, in regard to *RPS6KB1-VMP1*, Inaki *et al.* [116] suggest that it may rather be a marker of genomic instability or amplification of the 17q23 locus in which both genes are located, than an oncogenic fusion transcript.

5.1.2.4 Driver and passenger mutations

Although there is significant intertumoral variation in the number of mutations they carry, most tumors contain a large number of mutations, especially point mutations [120]. Mutations that attain a significant frequency in a tumor are unlikely to have a negative impact on cancer cell growth, as cells carrying these would have been removed by negative selection. Mutations can therefore be divided into two categories, drivers and passengers, based on whether they increase the net growth rate of a developing cancer cell, or whether they are selectively neutral [120]. The average selective growth advantage of an individual driver mutation has been estimated to be only ~0.4% [121], suggesting a developing tumor must accumulate a surprisingly large number of driver mutations, before becoming life threathening. The percentage of all mutations in a tumor that are drivers is poorly known, but results from glioblastoma multiforme suggest 8% of missense mutations may be drivers [122].

5.1.2.5 Epigenetic alterations

In addition to the DNA changing alterations described above, epigenetic changes are also common in cancers [123], and cancer genomes as a whole are frequently hypomethylated [123]. However, hypermethylation of CpG islands close to the promoters of genes leads to their silencing, and this is a common mechanism for TSG inactivation in cancer. Genome wide, several cancer types show alterations in CpG island methylation boundaries and significantly increased between tumor heterogeneity in the methylation status of a large number of specific genomic regions, compared to their tissues of origin [124]. This indicates a general loss of epigenetic stability in cancer and results in both inter- and intratumoral heterogeneity through its effects on gene expression levels [124]. Methylated cytosines in CpG dinucleotides are also more prone to mutation, either spontaneously or when exposed to ultraviolet light or tobacco carcinogenes [123]. The importance of altered methylation in cancer development is also supported by the recent discovery of frequent mutations in e.g. the *DNMT3A* DNA methyltransferase in acute myeloid leukemia, myelodysplastic syndrome and T-cell lymphoma [125-127]

5.2 METHODS FOR DETECTING CANCER ALTERATIONS

5.2.1 ARRAY COMPARATIVE GENOMIC HYBRIDIZATION

Array comparative genomic hybridization is based on the concept of competitive hybridization of DNA from two samples to the probes on a microarray [128-131]. Both cDNAs, bacterial artificial chromosomes (BACs) and synthetic oligonucleotides have been used as probes, typically printed or synthesized onto glass microscope slides [130, 132]. If some part of the genome is not present in equal number of copies in both samples, this will be visible as either a gain or loss of fluorescent signal from probes measuring that region, indicating the presence of an amplification or deletion. aCGH is always comparative, in other words, gains and losses are defined in relation to a reference sample. This applies also to Affymetrix SNP microarray derived copy number data, even if the hybridizations themselves are done with one sample per microarray and therefore are not competitive. Current aCGH microarrays can contain up to 1 million probes (Agilent SurePrint G3 Human High-Resolution Discovery 1M arrays), providing an average resolution of 3kb across the genome. aCGH does not detect balanced genomic rearrangements, such as translocations, in which no genetic material is gained or lost. In practice, however, it seems that many, if not most, translocations are accompanied by small copy number changes (either deletions or gains), which may be visible using aCGH [94, 133]. aCGH has been used most widely in cancer research [131]. Compared to G-band karyotyping, aCGH is able to identify much smaller copy number variants, and is therefore increasingly used in the diagnosis of e.g. idiopathic mental retardation and developmental malformations [134] as well as in prenatal diagnosis [135]. Beyond medical applications, aCGH has been used to study population wide copy number variation in several species, including humans [136, 137], various great apes [138] and dogs [139].

As aCGH data is comparative, results are nearly universally reported as ratios of sample divided by reference, frequently log-transformed to make them symmetric around zero. Visualization of the ratios in the context of their genomic positions then allows the determination of copy number profiles for all examined chromosomes. Simultaneous analysis of copy number profiles from multiple samples can be used to identify minimal common regions of amplification and deletion, the locations of potential oncogenes and TSGs [140, 141]. Minimal common region identification rests on dividing the genome into non-overlapping regions of differing copy number by segmentation [142-144]. Segmentation provides smoothed DNA copy number estimates for genomic regions by using the ratios from multiple adjacently located probes to derive an average copy number value for the region. Gene level copy number values, for integration with e.g. gene expression data, can be derived directly from the values of the segment in which the gene is located. Alternatively, gene level copy number data can be calculated, on a gene by gene level, from probes located in a specified window surrounding the gene's location [145]. Recently, next generation whole genome- or exome sequencing data has also been used for estimation of copy number [35, 146, 147], and this may come to replace aCGH in areas of research in which sequencing becomes common.

5.2.2 GENE EXPRESSION ARRAYS & GINI

Gene expression microarrays are miniaturized assays that enable measuring the expression of nearly all protein coding genes in the human genome in a single experiment. Expression arrays can be divided into two main types. One is based on the competitive hybridization of two samples on the same array, as done with aCGH [148]. In the other type, only one sample is hybridized onto the microarray, and the quantified signal is therefore the absolute fluorescent intensity measured, not a ratio of signal from two samples [149, 150]. As with aCGH, the probes may be either cDNAs or synthesized oligos, the latter being used almost exclusively these days. Oligo-based expression arrays range from relatively simple designs using long 60 base pair (bp) oligos (e.g. Agilent) to the more complex short oligo-based design of Affymetrix. Expression arrays have been used extensively in cancer research, contributing to identifying, in breast cancer alone, new subtypes [4, 5, 151], expression profiles predictive of disease outcome [152] and the impact of DNA copy number changes on expression levels [86, 87]. Outside of cancer research, they have been used in anything from researching the effects of parabolic flight on plant gene expression [153] to stydying gene expression changes in the brain caused by the domestication of dogs from wolves [154].

Although the bioinformatic methods used to analyze microarray data are almost as varied as the hypotheses being studied, all analyses start with preprocessing the microarray data [155]. The first step in data aquisition is the segmentation of scanned microarray images to obtain signal intensities for all probes [156]. In the literature, the post signal aquisition steps in microarray data preprocessing are frequently simply called "data normalization", although formally normalization is only one of the steps in preprocessing. Microarray preprocessing methods vary depending on the type of microarray, but all aim at correcting for technical noise and variation in the data [157]. For single-color microarrays, such as Affymetrix, one of the most commonly used preprocessing methods is the Robust Multiarray Average (RMA) [158]. RMA consists of three processing steps. The first step is background adjustment, in which an estimate of background signal intensity is subtracted from probe signals, under the assumption that background signal represents nonspecific hybridization. After background adjustment, probe intensities are normalized using quantile normalization.

Finally, data is summarized at the level of probe sets or other probe groupings, such as Ensembl gene definitions [159, 160].

Clustering and classification methods, also termed unsupervised and supervised classification, have frequently been used in microarray data analysis, the former especially in exploratory data analysis. Examples include the previously mentioned identification of breast cancer subtypes using hierarchical clustering [5] and definition of new subtypes of diffuse large B-cell lymphoma [161]. One major aim of microarray data analysis is the identification of genes that are differentially expressed between two or more groups of samples, e.g. samples subjected to a treatment compared to an untreated control group. Methods range from simple log fold change calculation [155] to more complex methods, such as Gene Set Enrichment Analysis [162, 163], that do not rely on defining a ratio cutoff for differential gene expression, but rather identify simultaneous changes in groups of genes that share a biological function. During the last decade, the data from tens of thousands of microarray hybridizations has been made public through repositories such as Gene Expression Omnibus (GEO) [164] and ArrayExpress [165]. This has prompted the development of meta-analysis methods to integrate data across multiple studies to be able to answer questions that no single study is powered to answer. Examples include GeneSapiens [166] and Oncomine [167], both of which concentrate on integrating data from cancer microarray studies. GeneSapiens normalizes Affymetrix gene expression data for altogether 9783 healthy, cancer and other disease samples onto the same scale, enabling e.g. studying the expression profile of all kinases across ~5600 different healthy and malignant tissue samples [168] as well as determining the origin of cancers of unknown primary origin [169].

Nonsense-mediated messenger RNA (mRNA) decay (NMD) is an eukaryotic quality control mechanism that triggers the decay of mRNAs that contain premature termination codons (PTCs) [170]. In addition, NMD also regulates the expression of a set of target transcripts under normal physiological conditions [171]. A PTC mutation is an effective way for a cancer cell to inactivate one copy of a TSG. Methods to identify such mutations based on the stabilization of PTC carrying mRNAs after either chemical (emetine with or without actinomycin D) [145, 172, 173] or siRNA-based [174] inhibition of NMD have therefore been developed (figure 1). One of the strengths of the gene identification by nonsense inhibition (GINI) method is that no *a priori* information about candidate genes or location in the genome is necessary, although if available, this information can be integrated with the GINI data [172]. Mutations have been found using an NMD-based approach in e.g. *EPHB2* in prostate cancer [173], *RIC8A* and *ARID1A* in breast cancer cell lines [145, 175], as well as several genes in colon [176] and prostate cancer [177], mantle-cell lymphoma [178] and melanoma [179].

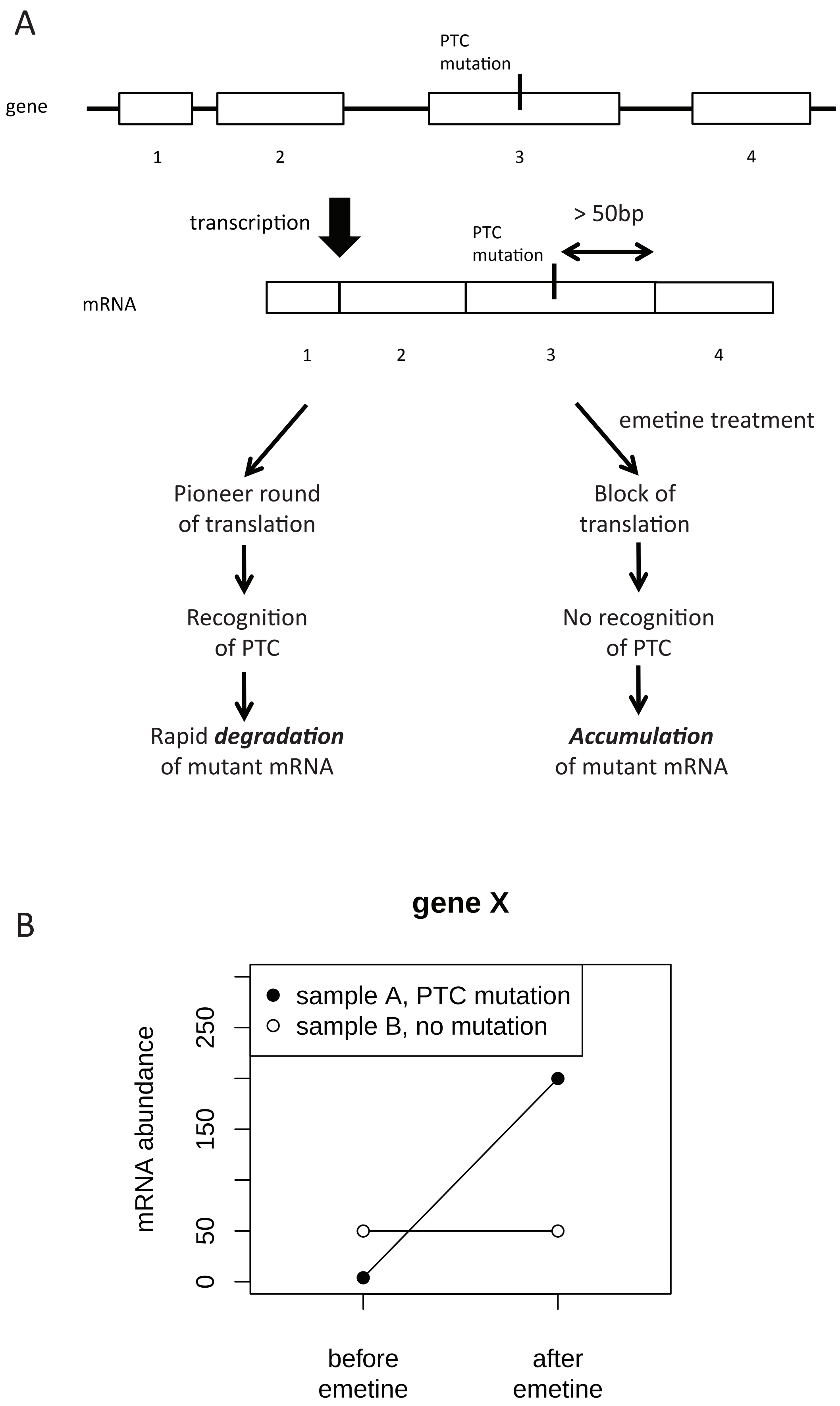

A
PTC
mutation
gene
1
2
3
4
transcription
> 50bp
PTC
mutation
mRNA
1
2
3
4
emetine treatment
Pioneer round of translation
Recognition of PTC
Rapid *degradation* of mutant mRNA
Block of translation
No recognition of PTC
Accumulation of mutant mRNA
B
gene X
● sample A, PTC mutation
○ sample B, no mutation
mRNA abundance
0
50
150
250
before emetine
after emetine

Figure 1 **Principle of emetine mediated NMD block. A)** If a PTC mutation occurs at least 50-54 bp before the last exon-exon junction, the mutation will be recognized during the pioneer round of translation and the mRNA molecule is degraded. Emetine is a general inhibitor of the translation process. Emetine treatment therefore also blocks the pioneer round of translation and prevents the NMD machinery from recognizing and degrading the mutated transcript, leading to accumulation of mutation carrying transcripts. **B)** An idealized example of the effects of emetine treatment on the abundance of mRNA transcripts of gene *X*. In sample A that carries a PTC mutation in *X*, emetine treatment leads to an increase in mRNA from gene *X*. Conversely, in sample B that has no mutation, transcript levels of *X* are not affected by emetine treatment. Note also that, compared to sample B, continued degradation of mutated transcripts from *X* in sample A leads to lower expression of the gene in the untreated state.

NMD microarray data analysis is in principle simple: a matter of identifying the mRNA transcripts that increase in amount following inhibition of NMD. In practice, however, a large number of transcripts are induced by NMD inhibition, whether chemical or siRNA-based [145, 172]. The main task of data analysis is therefore to prioritize a short list of the most likely mutation carrying genes. Several of the above mentioned studies have arrived at similar filtering algorithms. One of the two main filtering criteria follows; increased transcript level in only one out of several cell lines studied, this rests on the assumption that only one of the cell lines is likely to have inactivated a gene through a PTC, and transcripts upregulated in multiple cell lines are therefore likely to be physiological NMD targets. The second main criteria is that in untreated cells, expression of the transcript should be low compared to other samples, as would be expected based on a PTC containing transcript being degraded when NMD is intact [145, 178]. A futher criterion used in several publications is that the candidate gene should be located in a region of heterozygous deletion or loss of heterozygosity [145, 173, 175].

5.2.3 NEXT GENERATION SEQUENCING

Next generation sequencing is a collective term used to describe several different new sequencing technologies that utilize massive parallelization to achieve large increases in sequencing throughput in comparison to traditional capillary sequencing (Sanger sequencing) using e.g. ABI Prism 3730 DNA Sequencer instruments (Applied Biosystems). Currently, the main technologies in use are provided by Illumina (HiSeq, MiSeq, GA-family of instruments), Applied Biosystems (SOLiD), Roche (454), Life Technologies (Ion Torrent) and the technology of Complete Genomics [180, 181]. Massively parallel sequencing of RNA allows the comprehensive characterization of the features of a transcriptome, including gene expression levels, alternative splicing, identification of new transcripts as well as chimeric RNA molecules [182-185]. Chimeric RNAs, such as fusion transcripts, can be detected using paired-end sequencing of mRNA or ribosomal RNA (rRNA) depleted total RNA, in which 35-150 bp are sequenced from both ends of DNA molecules in the sequencing library (typically 200-500bp long). Whole-genome sequencing is also able to identify chromosomal rearrangements that potentially can create fusion genes (figure 2). However, RNA-seq can directly identify the expressed fusion genes, out of a potentially large set of rearrangements, and is therefore a more cost effective and straightforward method for detecting potentially oncogenic gene fusions.

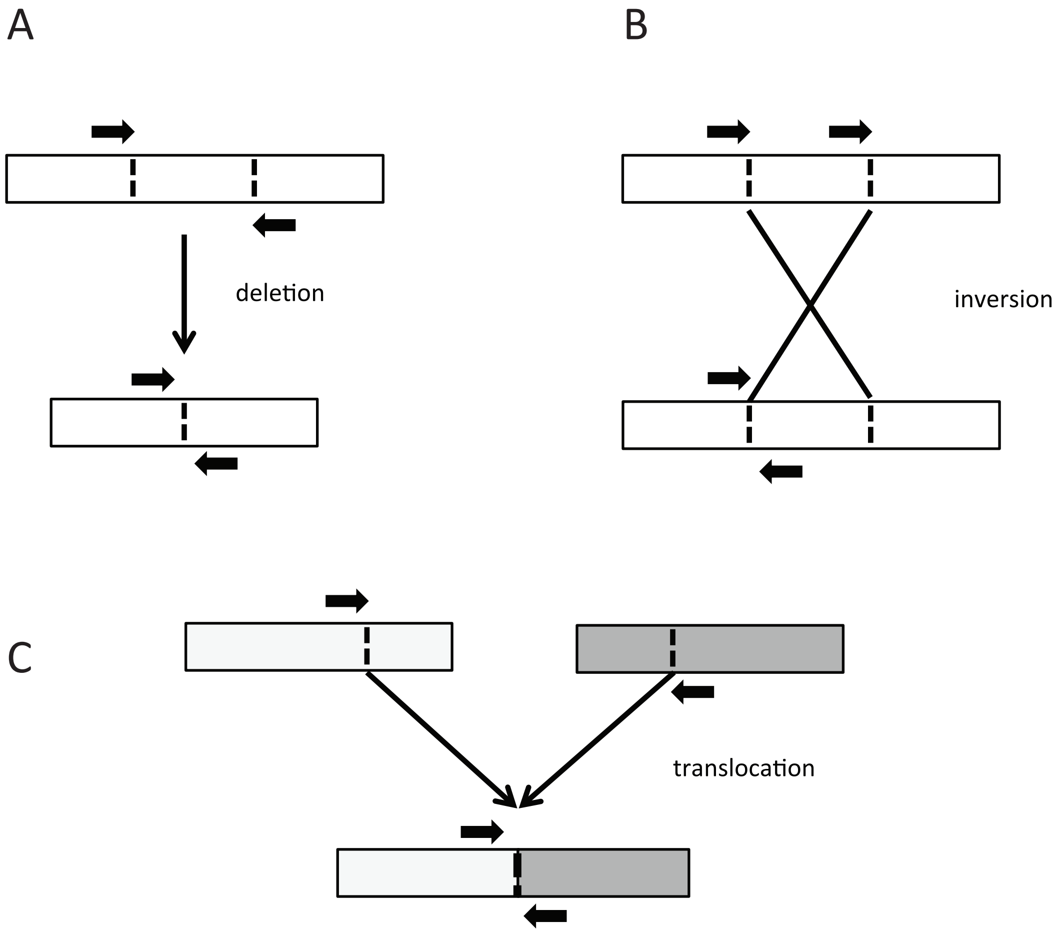


Figure 2 **Identification of deletions, inversions and translocations using paired-end DNA sequencing data.** Vertical arrows indicate sequence read pairs, and arrow directions show the strand they align on (arrow pointing to the right: forward strand). Vertical dashed lines indicate chromosomal breakpoints. In each subgraph A-C, the lower part shows the chromosome after the rearrangement, i.e. the state assayed by sequencing. The upper parts of each subgraph show how the reads in the readpair align to a normal reference genome. Rearrangements are identified as follows. **A)** When sequencing across a deletion point, the reads align further away from each other on the reference genome than would be expected. If the insert size is e.g. on average 300bp, reads that align 10kb from each other on the reference genome are likely to flank a roughly 9-10kb deletion. **B)** When sequencing across an inversion point, both reads will align on the forward strand when aligning them to a normal reference genome. In addition, depending on the size of the inversion, the reads may align further from each other than expected. **C)** When sequencing across a translocation point, both reads will align to different chromosomes in a normal reference genome.

Several pipelines have been published for fusion gene identification [94, 114, 186-189], but most methods that achieve a high specificity converge on very similar solutions. In all approaches, paired-end reads are first aligned and filtered to identify those pairs, in which the reads align to two different genes. This, however, does not distinguish between true fusion genes and readthrough transcription between genes that lie next to each other in the genome. Various solutions for filtering out transcriptional readthrough have been proposed, such as excluding all gene-gene pairs that lie closer to each other than some specified bp distance [114] or only considering gene-gene pairs that are separated by at least one other gene that lies between them [94]. The exon-exon junction at which the fusion occurs is then identified by searching non-aligned single-end reads for ones that align partially to exons from both genes. This search is typically done by bioinformatically constructing a library of all possible exon-exon junctions, i.e. potential fusion junctions, between a candidate gene-gene pair, against which alignments are performed. Fusion gene validation is then typically performed by polymerase chain reaction (PCR) and Sanger sequencing across the predicted fusion junction(s). Additional filtering criteria employed by some pipelines include filtering out gene-gene pairs with high sequence similarity, on the assumption that they are false positives derived from misaligned sequence reads [94, 114]. Additionally, the locations of alignment start positions for fusion junction spanning reads have proven to be a good criterion for excluding false positive fusion candidates [94]. One of the main points at which pipelines differ is whether they can identify fusions that do not occur at known exon-exon junctions. Here, the TopHat-Fusion [187] algorithm seems to provide the most robust detection of fusion junctions, in which one or both fusion breakpoints reside within exons.

6 AIMS OF THE STUDY

In broad terms, the aim of my PhD thesis work was to identify and characterize genomic mutations in human breast cancer, as well as to study their impact on breast carcinogenesis. In particular developing and then applying bioinformatic methods to help answer these questions.

The specific aims of the study were:

- To identify new tumor suppressor genes in breast cancer using the NMD microarray methodology.
- To develop a bioinformatic method for fusion gene identification using RNA-sequencing data.
- To study whether fusion genes exist in breast cancer, and if they do, characterize them and their potential impact on breast carcinogenesis.
- To characterize the *ERBB2* amplicon in breast cancer, its extent and the biological impact of both *ERBB2* and the other genes in the core *ERBB2* amplicon.

7 MATERIALS AND METHODS

7.1 BREAST CANCER CELL LINES AND CLINICAL SAMPLES USED

Table 3. ***Cell lines used in studies I-III.*** *Cell lines used in more than one study were separately analyzed by aCGH in each. Normal breast total RNA is listed under cell lines, even if it was derived directly from in vivo tissue and not cultured.*

Cell line	Study	Type
MDA-MB-468	I	breast cancer
MDA-MB-231	I	breast cancer
ZR-75-1	I	breast cancer
MCF-7	I, II, III	breast cancer
BT-474	I, II, III	breast cancer, ERBB2 amplified
T-47D	I	breast cancer
HMEC	I	normal human mammary epithelial
IMR90	I	normal human lung fibroblasts
WS1	I	normal human skin fibroblasts
SK-BR-3	II, III	breast cancer, ERBB2 amplified
KPL-4	II, III	breast cancer, ERBB2 amplified
normal breast total RNA	II	normal breast total RNA
HCC202	III	breast cancer, ERBB2 amplified
UACC812	III	breast cancer, ERBB2 amplified
HCC1954	III	breast cancer, ERBB2 amplified
HCC1569	III	breast cancer, ERBB2 amplified
JIMT-1	III	breast cancer, ERBB2 amplified
SUM190	III	breast cancer, ERBB2 amplified
SUM225	III	breast cancer, ERBB2 amplified

Table 4. ***Clinical microarray datasets used in studies I-III.*** *The samples used for sequencing or methylation analyses in study I were first described in the references given in the table. IKZF3 expression data was accessed via the GeneSapiens database described in Kilpinen et al., 2008. aCGH data used in study III is available from GEO with accession numbers GSE17907, GSE32291 and GSE20394.*

N of samples	Study	Purpose	Reference
127	I	*RIC8A* sequencing	Naume *et al.* 2007, Wiedswang *et al.* 2003
115	I	*RIC8A* expression	Naume *et al.* 2007, Wiedswang *et al.* 2003
86	I	*RIC8A* methylation	Warnberg *et al.* 2001
75	I	*RIC8A* methylation	Geisler *et al.* 2001
251	I	*RIC8A* expression	Miller *et al.* 2005
761	II	*IKZF3* expression	Kilpinen *et al.* 2008
54	III	*ERBB2* amplicon copy number	Sircoulomb *et al.* 2010
17	III	*ERBB2* amplicon copy number	Langeröd *et al.* 2007

7.2 MICROARRAY EXPERIMENTS (I-III)

7.2.1 NMD MICROARRAYS (I)

All cell lines were grown in replicate cultures. For inhibition of NMD, half of the cultures were treated with 100 µg ml^{-1} of emetine dihydrochloride hydrate (Sigma-Aldrich, St Louis, MO, USA) and incubated for 10h at 37°C, while the other half were retained as untreated controls. After incubation, total RNA was extracted, subjected to quality control and hybridized onto Affymetrix Human Genome U133 plus 2.0 GeneChip oligonucleotide microarrays (Affymetrix, Santa Clara, CA, USA) according to the manufacturer's instructions. Microarrays were scanned using a GeneChip Scanner 3000 (Affymetrix) and images processed using GeneChip Operating Software 1.1 (Affymetrix). Gene expression microarray data have been deposited in Gene Expression Omnibus (GEO) and are accessible through GEO accession number GSE15477.

Microarray data were normalized using the dChip method [190]. All further data analysis was carried out using the R statistical programming language [191]. NMD candidates for further validation were selected in each breast cancer cell line based on the following criteria:

I. *Emetine treatment induced the expression of the gene by at least 1.5 fold (log2 scale) only in the cell line being analyzed.* This filter was implemented to exclude the majority of physiological NMD substrates, as well as noise caused by emetine treatment.

II. *Expression in the non-emetine treated sample was lower than in the untreated sample of any other cell line.* This filter was based on the asumption that PTC containing transcripts are normally degraded by NMD and expression is therefore lower.

III. *The gene was located in a heterozygously deleted region.* This filter restricted the search to TSGs for which the other allele has been lost through deletion.

IV. *Expression after emetine treatment was higher than 50 units (non-log scale).* This filter was used to remove genes expressed at levels not measured reliably by the Affymetrix microarrays.

7.2.2 ARRAY-CGH (I-III)

During the years the project was ongoing, several versions of Agilent aCGH microarrays with increasing probe numbers were used. These were 44k Agilent Human Genome CGH microarrays (study I), 244k Agilent Human Genome CGH microarrays (study III) and SurePrint G3 Human 1M oligo CGH microarrays (study II) (Agilent, Palo Alto, CA, USA). As the labeling, hybridization and scanning protocols have not changed significantly, only one account of these is given below.

Overall, aCGH experiments were performed as described previously [96] following protocols for respective microarray types (Agilent). Briefly, genomic DNA from untreated cancer cells was extracted and labeled using Cy5-dUTP. Commercially available female genomic reference DNA (i.e. not matching germline DNA) was labeled with Cy3-dUTP and used as reference in all experiments. After hybridization and washing, microarrays were scanned using a laser scanner (Agilent) and images processed for signal acquisition using Feature Extraction Software (Agilent). CGH Analytics or Genomic Workbench Lite software (Agilent) was used for data visualization.

In study I, gene expression and aCGH data were combined by calculating, for each Affymetrix probe set, the median log2 ratio of all aCGH oligos located between the start and stop base pair positions of the gene the probe set mapped to. Mappings between probe sets and genes were retrieved from the NetAffx database (april 2005) [192] and all base pair positions are based on human genome build hg17 (NCBI 35, May 2004). In study II, the association between genes taking part in gene fusions and copy number changes was assesed visually using Agilent Genomic Workbench Lite (Agilent). In study III, aCGH data for cell lines and tumors from both our own cohort, as well as Sircoulomb *et al.* 2010 [193], were segmented using the Piecewise Constant Fit algorithm with settings K-min = 5 and Gamma = 15 [194]. For a list of cell line and tumor samples, see tables 3 and 4. Heatmaps to visualize segmented data were drawn using R [191].

For studies I and II, aCGH data have been deposited in GEO and are accessible through GEO Series records GSE15477 and GSE23949. aCGH data for the cell lines used in study III have been submitted to GEO, accession number GSE34236, but remains private pending publication of the paper.

7.3 RNA-SEQ AND FUSION GENE IDENTIFICATION (II)

In brief, RNA-seq libraries were created as follows. Total RNA from the four breast cancer cell lines BT-474, SK-BR-3, KPL-4 and MCF-7 was isolated, followed by messenger RNA (mRNA) extraction using oligo-dT Dynabeads (Invitrogen Inc., Carlsbad, CA, USA). Extracted mRNA was randomly fragmented to an average size of 200bp and converted into cDNA using random hexamers. Fragment length and cDNA concentration was measured using Bioanalyzer DNA 1000 kit (Agilent). The median insert sizes of the final sequencing libraries were 100bp for MCF-7 and KPL-4, whereas libraries of both 100 and 200bp were created for BT-474 and SK-BR-3. For the normal breast sample, median insert size was 200bp. 2*56bp paired-end sequencing was carried out using a 1G Illumina Genome Analyzer IIx (Illumina). Raw sequencing data have been deposited in the NCBI Sequence Read Archive [SRA:SRP003186].

A workflow describing the fusion gene identification pipeline is shown in figure 3. Short reads were trimmed from 56 to 50bp for all analyses. Sequence alignment was done using the Bowtie software [195], allowing a maximum of 3 mismatches. Ensembl version 55 was used as reference for all analyses concerning BT-474, MCF-7, KPL-4 and normal breast, whereas version 56 was used for SK-BR-3. Both are based on human genome build NCBI37. Short reads were first filtered by aligning against ribosomal RNA (18S, 28S, 5S and 5.8S) and complete repeating unit ribosomal DNA, excluding any aligning reads from further analyses. In addition, short reads matching adapter sequences, mitochondrial DNA or containing long homopolymeric stretches were removed. The filtered short reads were next aligned against the human genome and a library of splice site junctions based on the transcript structures of each gene. Short reads were divided into three categories:

I. Reads that do not align to the genome.
II. Reads that align uniquely to the genome.
III. Reads that align to multiple locations in the genome.

A short read was considered uniquely aligning if there was a single best alignment, defined as having the smallest number of mismatches.

To identify fusion genes, non-aligned and uniquely aligned short reads were aligned to the Ensembl transcript definitions and reads were assigned to genes based on the transcript they aligned to. Short read pairs in which the two reads align to different genes were selected for further analysis. A first set of fusion gene candidates was identified by selecting all the gene-gene pairs that were supported by at least two (MCF-7, KPL-4, normal breast) or three (BT-474, SK-BR-3) short read pairs. As sequencing depth was higher for BT-474 and SK-BR-3, a higher threshold was chosen for them in an effort to keep the likelyhood of false positives the same for all samples (table 5). These lists were further filtered by excluding all fusion gene candidate pairs in which the two genes are either known paralogs or adjacent to each other in the genome. Both genes in a candidate fusion were also required to be protein coding. In addition, fusion gene candidates involving genes taking part in more than a few potential fusions were excluded. Paralog and gene biotype data were retrieved from Ensembl. Genes were defined as non-adjacent if there was a third gene, the start and stop positions of which lie between the two other genes.

The exon-exon fusion junctions were identified as follows. A library of artificial fusion junctions was created by generating all the potential exon-exon combinations between the two genes in each candidate fusion gene pair. Short reads not aligning to either the genome or the transcriptome were aligned against the library of fusion junctions. Short reads were required to overlap the exon-exon junction by at least 10 bp. For each candidate fusion gene, the exon-exon junction supported by the greatest number of short reads was nominated as the most likely fusion junction. This also defined which gene is the 5' partner in the fusion. At least two fusion junction spanning reads were required. The final list of 28 candidate fusion genes was selected for laboratory validation primarily based on the number of unique short read alignment start positions across the fusion junction. A secondary criterion stating that predicted fusion junctions should lie close to a copy number transition was used for fusion junctions with a low number of aligning reads.

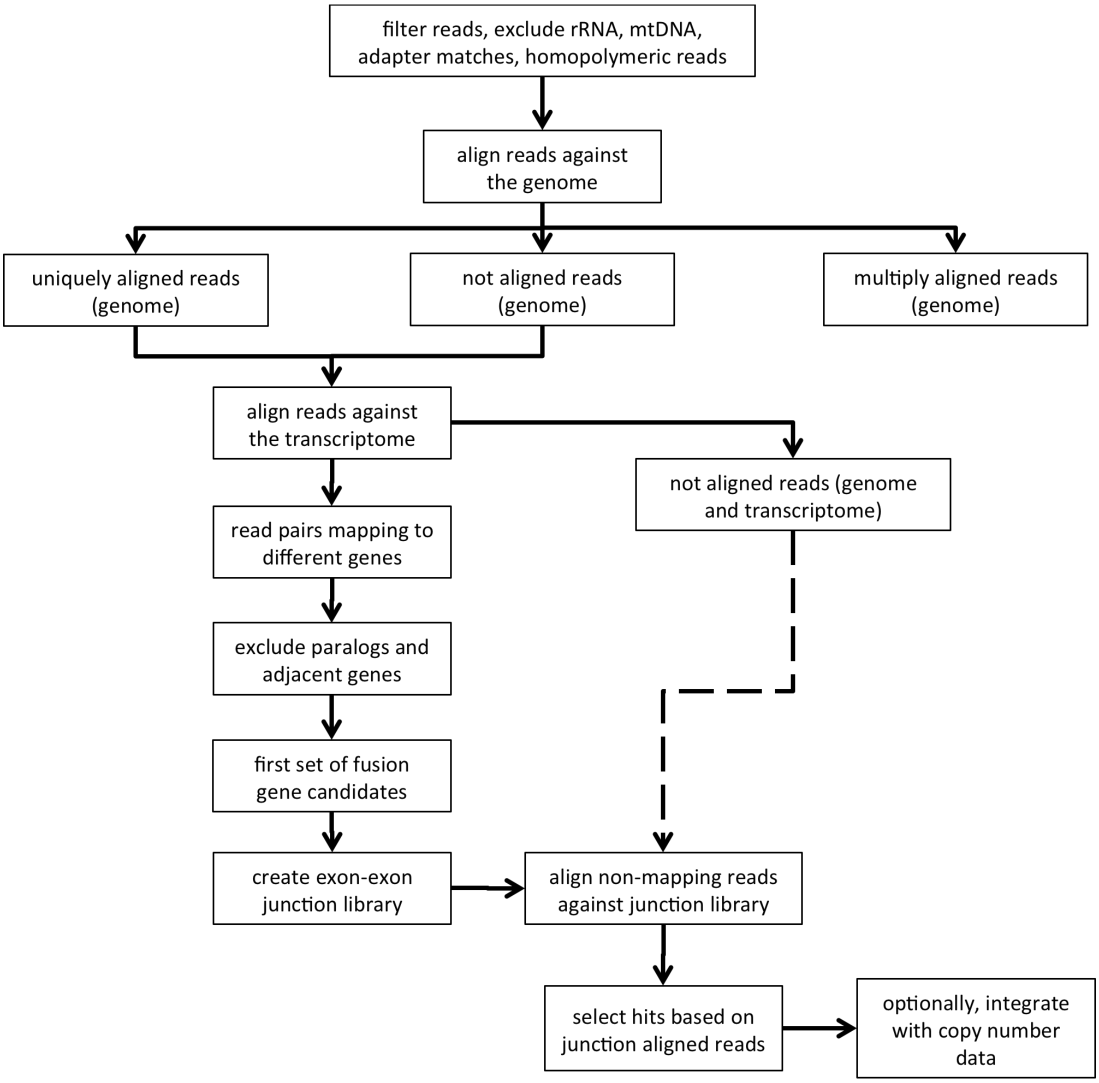

filter reads, exclude rRNA, mtDNA, adapter matches, homopolymeric reads
align reads against the genome
uniquely aligned reads (genome)
not aligned reads (genome)
multiply aligned reads (genome)
align reads against the transcriptome
not aligned reads (genome and transcriptome)
read pairs mapping to different genes
exclude paralogs and adjacent genes
first set of fusion gene candidates
create exon-exon junction library
align non-mapping reads against junction library
select hits based on junction aligned reads
optionally, integrate with copy number data

Figure 3 Overview of the fusion gene identification pipeline.

To determine if fusion gene candidates were in frame, all possible fusions between those Ensembl transcripts of both genes that contain the fused exons were created. If any of these transcript-transcript fusions retained the same reading frame across the junction, the fusion gene was predicted to be in frame. Expression of fusion genes and their wt partners was calculated as reads per kilobase per million uniquely aligned sequences (RPKM) [184]. Reads aligning to the fusion junction were used to calculate fusion gene specific RPKM values.

Graphs visualizing fusion genes together with aCGH data were created using R and Circos [196]. The graph illustrating fusion genes in the chromosome 17q amplicons was drawn using R and a modified version of the Bioconductor [197] package GenomeGraphs [198]. Graphs integrating aCGH, sequencing data and gene structures were drawn using GenomeGraphs. Unless otherwise noted, all steps in the fusion gene identification and prioritization pipeline are based on custom in-house tools written in Python and R.

7.4 SIRNA EXPERIMENTS (II, III)

In study III, siRNA screening was performed on five cell lines; BT-474, SK-BR-3, JIMT-1, KPL-4 and MCF-7. Out of 27 genes in the core *ERBB2* amplicon, 23 were silenced using 2-4 unique siRNAs per gene (Qiagen, Hilden, Germany), with siRNAs against *PLK1*, *KIF11* as well as AllStars Negative Control and AllStars Cell Death Control (Qiagen) used as controls. Screening was performed with one individual siRNA per well on 384 well plates. CellTiter-Glo (CTG, Promega) was used to measure cell viability 72 hours after transfection. Screening data was normalized for plate and row/column effects and replicate screens were merged. Co-transfections of *ERBB2* with other genes were done using two siRNAs for both genes in the same well.

A siRNA screen against genes taking part in the identified fusion genes was performed in triplicate in KPL-4, SK-BR-3 and BT-474 cells largely as described above. Separate validation was carried out for *IKZF3* in BT-474 using two siRNAs, Hs_IKZF3_3 and HS_ZNFN1A3_5 (Qiagen) in 96 well plates using CellTiter-Glo Cell Viability Assay (Promega, Madison, WI, USA) as an endpoint after 168 hour incubation. Quantitative RT-PCR using LightCycler 480 (Roche Applied Science, Penzberg, Germany) and *GAPDH* as an internal control was used to validate *VAPB-IKZF3* fusion knock down.

7.5 PROTEIN LYSATE MICROARRAYS (III)

Protein lysate microarray analysis (LMA) was carried out as in Leivonen *et al.* [199]. Briefly, protein lysates were created from siRNA transfected cells in 384-well plates 72 hours after transfection by lysing the cells. Cell lysates were printed on nitrocellulose-coated microarray slides and stained with antibodies for cleaved PARP (cPARP), Ki67, HER2, phospho-Akt, phospho-p70-S6K and p27. Scanned signal intensities were normalized using Z-score normalization and data analyzed using Array-Pro Analyzer microarray software (Median Cybernetics Inc., Bethesda, MD, USA). Hit genes were selected using three cutoff levels, in ascending order of reliability: Z-score > 1 but < 2 for two siRNAs, siRNAs with one Z-score > 1 and another > 2 and the third category with two siRNAs with Z-score > 2.

7.6 VALIDATION OF MUTATIONS AND GENE FUSIONS (I, II)

In study I, primers were designed to amplify all 10 exons as well as 50bp of flanking intronic sequences of *RIC8A*, *PGPEP1* and *COL12A1* and the genes sequenced in all six breast cancer cell lines using Sanger sequencing. *RIC8A* was further sequenced in two cohorts of 127 and 86 tumors, respectively (table 4) using the same primers as described above. The association between low *RIC8A* expression and *TP53* mutation was validated using qRT-PCR in a subset of 38 tumors from the Naume *et al.* and Wiedswang *et al.* [200, 201] cohort, using TaqMan Gene Expression Assays (Applied Biosystems, Carlsbad, CA, USA) on an ABI Prism 7900 HT sequence detection system (Applied Biosystems). Methylation analysis of 98 CpGs in the region from -943 to +1338 around the *RIC8A* transcription start site was performed in two cohorts of 86 and 75 tumors (table 4) using pyrosequencing of bisulphite treated DNA [202].

In study II, several approaches were used to validate fusion gene candidates identified using RNA-seq. Predicted fusion genes were first validated using RT-PCR across the predicted fusion junctions, followed by Sanger sequencing of amplification products. Obtained sequences were aligned to human genome build hg19 (february 2009) using the BLAST-Like Alignment Tool alignment program [203] to ensure that they uniquely matched the exon boundaries of the predicted fusion junctions. DNA level rearrangements were validated using long-range genomic PCR and fluorescent *in situ* hybridization (FISH). Primers for genomic PCR were placed based on the positions of the fused exons such that the PCR product covered the fusion junction. When a copy number transition was evident close to the fused exon(s), this information was used to place PCR primers closer to the likely genomic fusion point. Interphase FISH was performed using BAC probes located as close as possible on each side of the breakpoint. Fusions were detected as fused signals from the FISH probes.

7.7 ANALYSIS OF PUBLICLY AVAILABLE DATA (I, II)

RIC8A expression was analyzed in two published breast cancer microarray studies [55, 200]. The Miller *et al.* dataset (table 4, GEO accession number GSE3494), consisting of 251 breast tumors profiled on both Affymetrix U133A and U133B microarrays, was normalized in R using RMA [158] and probes summarized on the level of Ensembl gene ids, using the alternative CDF file definitions of Dai et al [159]. Data from both array types were combined. When genes appeared on both array types, data were combined by calculating their median expression values across each sample from both arrays. Normalized data together with sample annotations were transformed into a Bioconductor ExpressionSet [197] and further into an R data package.

PCNA, Ki67 and ERBB2 status was estimated from the expression data and tumor subtype classification done based on [6]. An R package, "phenoplots", was written to enable automated plotting of gene expression data in relation to clinical parameters as either "phenoplots" (figure 5) or annotated heatmaps. Functionality to draw correlation plots between two genes as well as Kaplan-Meier survival plots based on the expression of a gene was also implemented. The functions in the phenoplots package require that data sets are formated into Bioconductor ExpressionSets with sample annotations stored in the phenoData slot, and transformed into R data packages. Statistical association between *RIC8A* expression and clinical parameters was tested using the Mann-Whitney test. *IKZF3* expression in breast cancer was analyzed using the GeneSapiens database [166].

8 RESULTS

8.1 IDENTIFICATION OF *RIC8A* (PUBLICATION I)

8.1.1 *RIC8A* IDENTIFICATION VIA NMD MICROARRAYS

The NMD microarray method identified 51 candidate genes in the six breast cancer cell lines MDA-MB-468, MDA-MB-231, ZR-75-1, MCF-7, BT-474 and T-47D. Out of these, three genes were selected for further sequencing-based validation. The selected genes were *PGPEP1* in BT-474 and *COL12A1* and *RIC8A* in ZR-75-1. The NMD induction ratios of the genes were 1.89, 5.16 and 1.51, respectively. Sequencing of the three genes from the genomic DNA of all six cell lines identified a CAG -> TAG nonsense mutation in the last codon of the third exon of *RIC8A* in ZR-75-1. No nonsense mutations were identified in other cell lines for any of the genes. As can be seen in figure 4, *RIC8A* is heterozygously deleted in ZR-75-1, suggesting that both alleles of the gene are lost. Furthermore, based on Affymetrix data, expression of *RIC8A* was significantly lower in the untreated ZR-75-1 cells compared to all other cell lines, further pointing to complete loss of the gene. This differential expression in untreated cells was also confirmed using qRT-PCR.

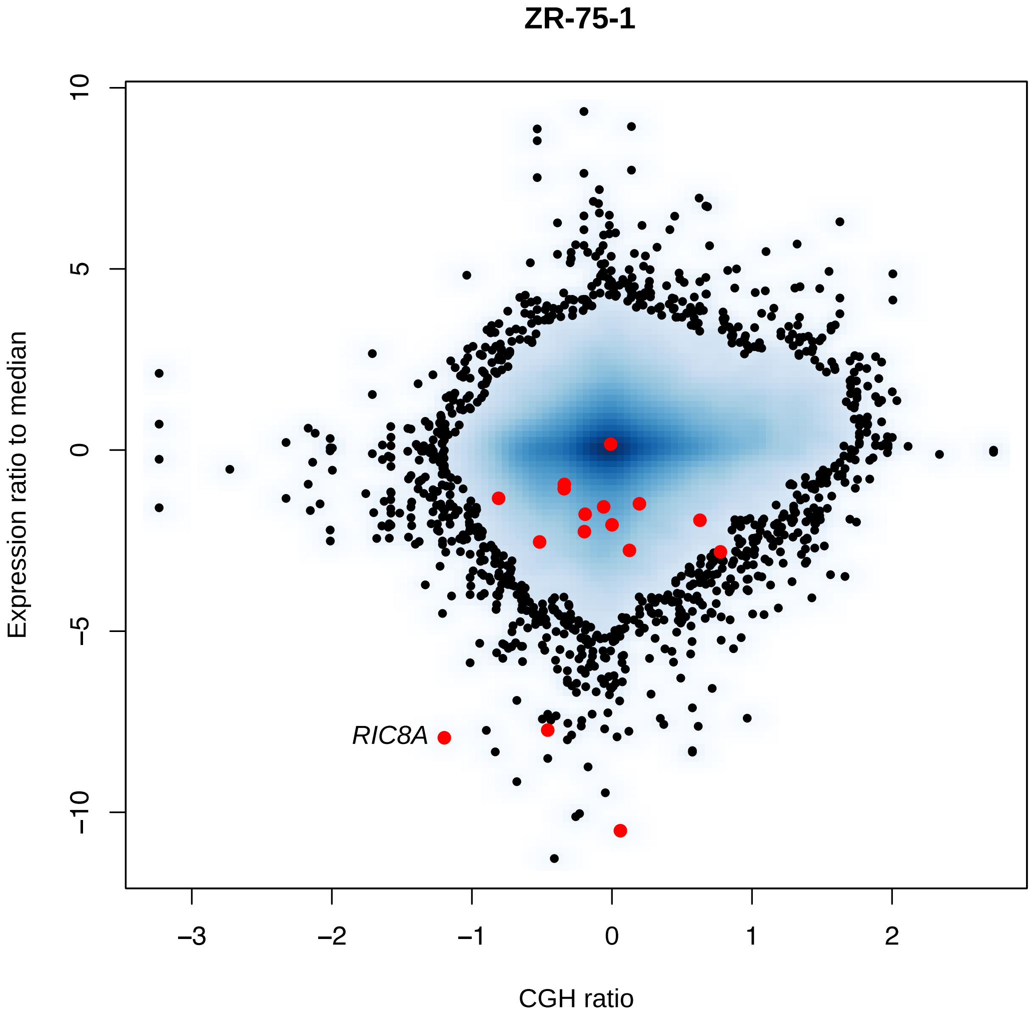


Figure 4 **Heterozygous deletion and down regulation of expression of *RIC8A* in ZR-75-1.** The x-axis shows the gene copy number for each probe set, while the y-axis shows, for each probe set, the log2 expression ratio of the probe set's expression in ZR-75-1 divided by its median expression across all cell lines used in the study. Candidate NMD target genes are shown in red. The intensity of the blue color indicates the number of probe sets located in the region and only the most outlying observations are drawn as individual dots. *RIC8A* stands out as being both heterozygously deleted and underexpressed in untreated ZR-75-1.

8.1.2 CLINICAL RELEVANCE OF *RIC8A*

To study whether *RIC8A* is mutated in clinical breast tumors, all ten exons of the gene were sequenced in 127 early-stage breast cancers. No nonsense or missense mutations were identified. We additionally performed pyrosequencing on 98 CpGs within 1kb upstream of the *RIC8A* transcription start site (TSS), as well as CpGs in the 3' end of the large CpG island downstream of the TSS in a total of 161 breast tumors of varying severity (27 ductal carcinoma *in situ* (DCIS), 32 invasive tumors with DCIS components, 27 early invasive tumors and 75 locally advanced tumors). No methylation was observed in the CpGs upstream of the TSS. Downstream of the TSS, in the 3' region of the CpG island, methylation levels reached 40-80%, but this did not differ from the methylation pattern observed in normal breast tissue samples. Taken together, the DNA- and pyrosequencing data suggest *RIC8A* is not a frequent target of mutation or epigenetic silencing by DNA methylation. The chromosome band 11p15 is frequently deleted in breast cancer [204], suggesting an alternative mechanism for *RIC8A* loss. To examine whether expression of *RIC8A* is lost in a subset of breast cancers through deletion or other mechanisms, we studied its expression in published microarray data sets. In the Miller *et al.* data set of 251 consequtively collected breast cancers [55], low *RIC8A* expression was seen in 15% of tumors and low expression was statistically significantly associated with ER-negativity (P < 0.001), PR-negativity (P < 0.003) and *TP53* mutation (P < 0.0001) (figure 5). Validation in a second data set consisting of 115 early-stage breast tumors [200] showed a borderline statistically significant association between low expression and PR-negativity (P = 0.054) and *TP53* mutation (P = 0.071). Taqman qRT-PCR validation of the association between low *RIC8A* expression and *TP53* mutation in 38 tumors from the second cohort was statistically significant (Mann-Whitney test: P = 0.006).

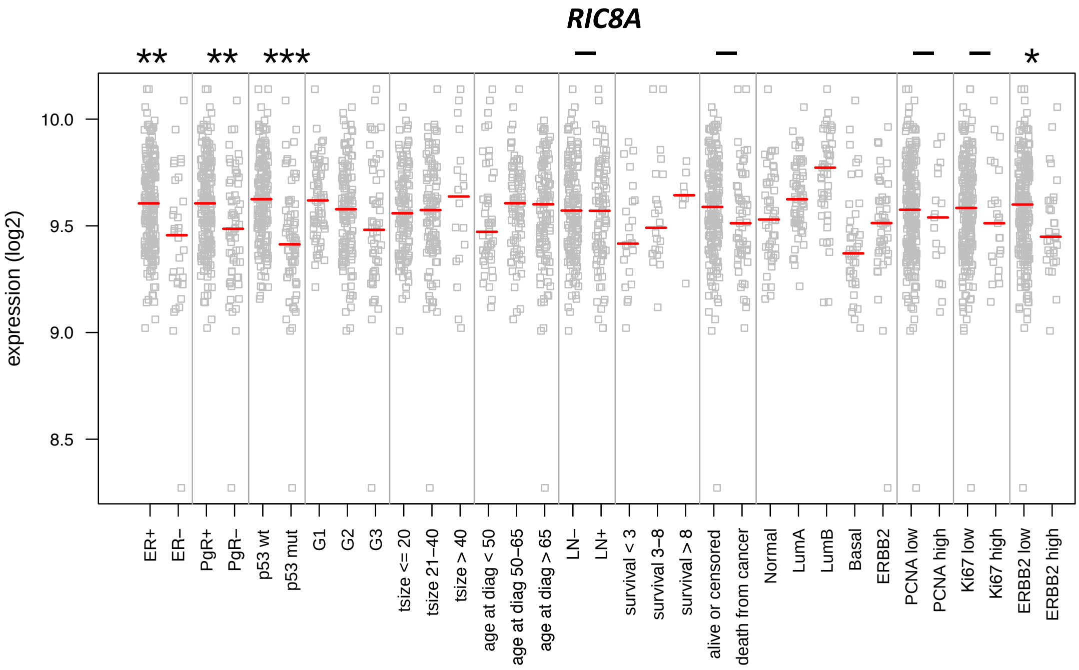


Figure 5 **Association of *RIC8A* expression with clinical parameters.** *RIC8A* expression across 251 consequtively collected breast tumors [55], stratified by clinical parameters. Grey squares represent expression values. The y-axis is drawn on log2 scale. Grey vertical lines separate clinical parameters. Red horizontal bars indicate the median expression value for each group. For each clinical parameter, data are drawn for all tumors for which the parameter is given. Expression data in relation to survival is only plotted for tumors from patients that eventually died of breast cancer. Stars indicate a statistically significant difference in expression between the two groups * p < 0.05, ** p < 0.01, *** p < 0.001, - p > 0.05. Statistical testing was only performed for clinical variables with two categories. p53mut = sample is classified as p53 null based on the gene expression classifier created in [55]. G1 = grade 1, tsize = tumor size, LN- = negative lymph nodes, Normal, LumA, LumB, Basal, ERBB2 = breast cancer subtypes. *PCNA* low = tumors with low *PCNA* expression, Ki67 low = tumors with low *MKI67* expression, *ERBB2* low = tumors with low *ERBB2* expression.

8.2 FUSION GENE IDENTIFICATION (PUBLICATION II)

8.2.1 RAW SEQUENCING OUTPUT

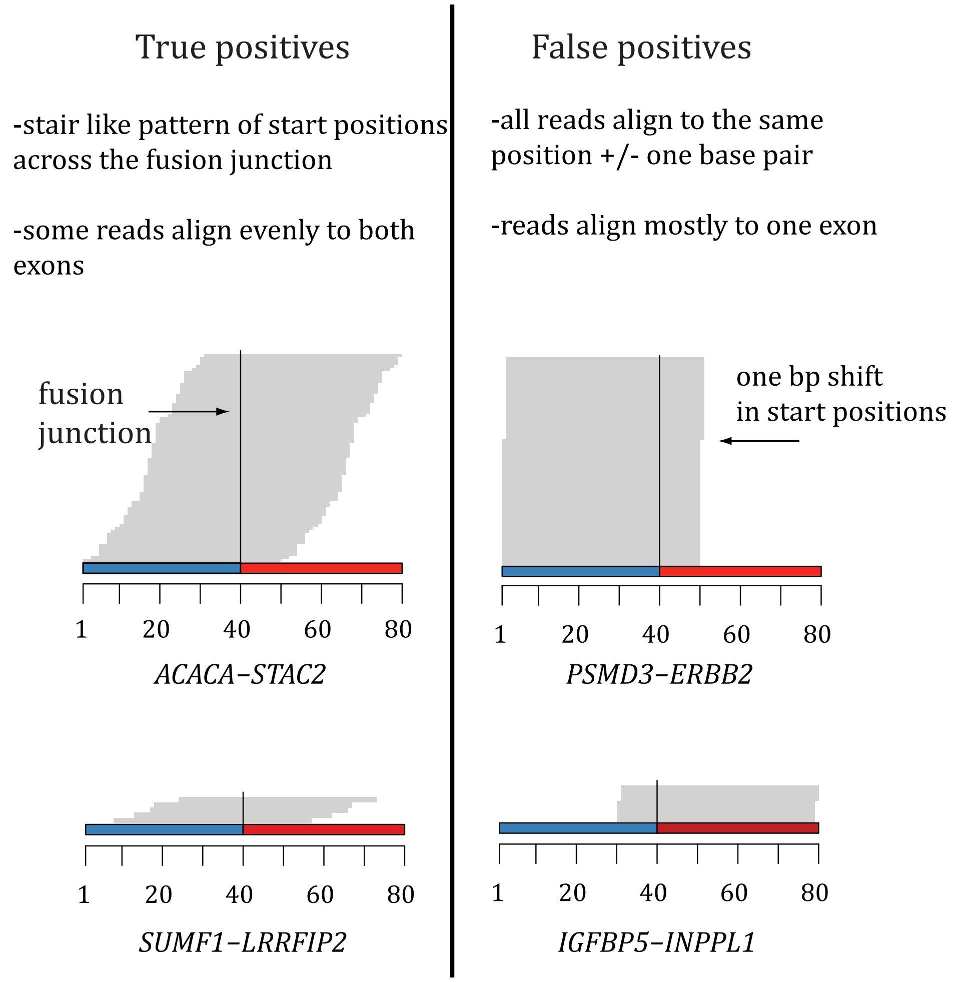
Each cell line was sequenced on one to three lanes on the 1G Illumina Genome Analyzer IIx sequencer (table 5). Optimization of the sequencing protocol, especially the amount of sequencing library loaded onto the flow cell, allowed a significant increase in reads obtained per lane in later instrument runs (BT-474, SK-BR-3) compared to the samples sequenced first (MCF-7, KPL-4, normal breast).

Table 5. ***Summary statistics for alignment results.*** *The number of lanes sequenced for each sample is given on row one. Alignment statistics and total read numbers are counted after reads aligning to rRNA, mitochondrial DNA and adapter sequences were removed.*

	BT-474	SK-BR-3	KPL-4	MCF-7	Normal breast
Number of lanes	2	2	1	3	1
Uniquely aligning reads	22.729.557	32.131.811	7.979.513	9.391.444	8.554.829
Multiple aligning reads	3.053.108	4.832.155	1.031.888	1.551.671	1.261.876
Non-aligning reads	3.947.238	5.373.572	1.188.192	1.862.559	1.317.916
Total	29.729.903	42.337.538	10.199.593	12.805.674	11.134.621

8.2.2 A METHOD FOR DETECTING FUSION GENES

In our preliminary analysis, we identified between 303 and 349 fusion gene candidates in each of the four breast cancer cell lines, plus an additional 152 in the normal breast sample. A set of 83 candidate fusion genes was selected for RT-PCR validation, but only seven of these were validated. This suggested that a majority of the 152-349 fusion gene candidates were likely to be false positives. In an effort to increase the specificity of our method, we reasoned that a true positive fusion gene would be expected to create a staggered pattern of alignment start positions across the fusion junction, whereas potential PCR artifacts or incorrectly aligned reads would be likely to align all in the same position (figure 6). Phrased differently, the number of unique alignment start positions across a fusion junction should be large for true positive fusion genes, but low for false positives. Comparing the pattern of alignment start sites between the seven validated and 76 non-validated fusion gene candidates indeed confirmed our hypothesis. Additionally, it showed that for false positive fusions, most of the length of the sequencing read aligned to one of the exons, suggesting these represent incorrect alignments, not PCR artifacts. This was further supported by the fact that the paired-end reads of these short reads did not align within close proximity of each other. Reanalyzing the data using this additional criterion resulted in the identification of 28 fusion gene candidates in the four breast cancer cell lines and none in the normal breast sample (table 6).



True positives

-stair like pattern of start positions across the fusion junction

-some reads align evenly to both exons

False positives

-all reads align to the same position +/- one base pair

-reads align mostly to one exon

Figure 6 **Pattern of alignment start positions across the fusion junction.** Examples of alignment start position patterns for fusion genes expressed at high and low levels. True positive fusions are shown on the left, false positive on the right. Grey horizontal lines represent short reads. The exon-exon junction is marked by a vertical black line. True positive fusions have a staggered pattern of alignment start positions and some reads align equally much to both exons. False positive fusions are characterized by piles of reads in the same position or offset by one bp. These reads also align to one of the exons for most of their lengths.

Table 6. ***28 fusion gene candidates and their validation.*** *Fusion genes are listed with the 5' partner gene first. N. paired-end reads and N. junction reads are the number of read pairs and single reads supporting the gene fusion. NA indicates that in frame status and validation of a DNA level rearrangement were not applicable to RBM41-MAN1A2, as it was not validated with RT-PCR and was therefore excluded from all further analyses.*

Cell line	5' gene	5' chr	3' gene	3' chr	N. paired-end reads	N. junction reads	In frame?	RT-PCR validated?	DNA rearrangement validated?
BT-474	*ACACA*	chr17	*STAC2*	chr17	57	72	yes	yes	yes
BT-474	*RPS6KB1*	chr17	*SNF8*	chr17	43	68	yes	yes	yes
BT-474	*VAPB*	chr20	*IKZF3*	chr17	41	26	yes	yes	yes
BT-474	*ZMYND8*	chr20	*CEP250*	chr20	35	14	no	yes	yes
BT-474	*RAB22A*	chr20	*MYO9B*	chr19	9	12	no	yes	yes
BT-474	*SKA2*	chr17	*MYO19*	chr17	8	7	yes	yes	yes
BT-474	*STARD3*	chr17	*DOK5*	chr20	4	6	yes	yes	yes
BT-474	*LAMP1*	chr13	*MCF2L*	chr13	5	3	no	yes	yes
BT-474	*GLB1*	chr3	*CMTM7*	chr3	6	2	yes	yes	yes
BT-474	*CPNE1*	chr20	*PI3*	chr20	4	2	no	yes	yes
BT-474	*DIDO1*	chr20	*KIAA0406*	chr20	8	1	yes	yes	no
SK-BR-3	*TATDN1*	chr8	*GSDMB*	chr17	28	447	yes	yes	yes
SK-BR-3	*CSE1L*	chr20	*ENSG00000236127*	chr20	10	20	yes	yes	no
SK-BR-3	*RARA*	chr17	*PKIA*	chr8	13	10	yes	yes	yes
SK-BR-3	*ANKHD1*	chr5	*PCDH1*	chr5	12	6	yes	yes	yes
SK-BR-3	*CCDC85C*	chr14	*SETD3*	chr14	6	6	yes	yes	yes
SK-BR-3	*SUMF1*	chr3	*LRRFIP2*	chr3	14	5	yes	yes	no
SK-BR-3	*WDR67*	chr8	*ZNF704*	chr8	3	3	yes	yes	yes
SK-BR-3	*CYTH1*	chr17	*EIF3H*	chr8	38	2	yes	yes	yes
SK-BR-3	*DHX35*	chr20	*ITCH*	chr20	3	2	yes	yes	yes
SK-BR-3	*NFS1*	chr20	*PREX1*	chr20	5	9	yes	yes	no
KPL-4	*BSG*	chr19	*NFIX*	chr19	22	14	yes	yes	yes
KPL-4	*PPP1R12A*	chr12	*SEPT10*	chr2	2	6	yes	yes	yes
KPL-4	*NOTCH1*	chr9	*NUP214*	chr9	4	6	yes	yes	yes
KPL-4	*RBM41*	chrX	*MAN1A2*	chr1	2	2	NA	no	NA
MCF-7	*BCAS4*	chr20	*BCAS3*	chr17	133	142	yes	yes	reported previously
MCF-7	*ARFGEF2*	chr20	*SULF2*	chr20	17	25	yes	yes	reported previously
MCF-7	*RPS6KB1*	chr17	*TMEM49*	chr17	2	7	yes	yes	reported previously

8.2.3 FUSION GENE VALIDATION

From the list of 28 fusion gene candidates in the four breast cancer cell lines, we were able to validate 27 using RT-PCR across the fusion junction followed by Sanger sequencing. Only the candidate fusion *RBM41-MAN1A2* in KPL-4 was not validated as it did not produce a band in RT-PCR despite multiple attempts. When multiple bands were observed for a fusion gene, all were picked for sequencing and subsequent validation. One possible mechanism for generating false positive fusions that would pass our criteria would be for a gene to contain an unannotated exon or a retained intronic sequence that is highly homologous to another gene, creating an apparent gene fusion. To exclude this possibility, the fusion junction sequences obtained from Sanger sequencing were aligned to the human genome to ensure they align uniquely only to the expected exons of respective fusion partner genes. All Sanger sequences from the 27 validated fusion genes aligned uniquely as expected.

The three fusion genes identified in MCF-7 (*BCAS4-BCAS3, ARFGEF2-SULF2, RPS6KB1-TMEM49*) were previously known [205-207], but the remaining 24 were novel. Validation of *NFS1-PREX1* is tentative, as only a short stretch of *NFS1* is included in the fusion, complicating PCR primer design and alignment of the short sequence uniquely to the genome.

mRNA *trans*-splicing is a mechanism in which exons or other sequences from two pre-mRNA molecules, transcribed from different genes, are spliced together to form a chimeric mRNA [208]. On the mRNA level, such a chimeric mRNA is indistinguishable from a fusion gene formed through a genomic rearrangement [209]. To exclude the possibility that fusion genes were formed by mRNA *trans*-splicing we tried to validate an underlying genomic rearrangement for all the 24 novel fusion genes by either genomic DNA PCR or interphase FISH. We were able to confirm a genomic rearrangement with either method for 20 of 24 novel fusion genes. In the remaining cases, the intron between the fused exons is likely to be so large that a PCR product could not be amplified, although we cannot formally exclude the possibility that the fusion transcripts arose through mRNA *trans*-splicing.

8.2.4 THE BIOLOGICAL FEATURES OF THE FUSION GENES

Integration of RNA-seq with aCGH data showed that in 23 of 27 fusion genes, one or both partner genes were associated with a copy number transition close to exons taking part in the fusions. This suggested that most of the fusion genes are not balanced translocations in the traditional sense of no change in DNA copy number. Additionally, in 17 fusion genes, one or both genes were located in high level amplifications on chromosomes 8, 17 and 20. Some of the fusion genes, such as *TATDN1-GSDMB* and *VAPB-IKZF3* bridge separate amplification regions on either the same or different chromosomes, suggesting that these amplicons reside on the same derivative chromosome (figure 7). FISH analysis showed that the fusion genes were on average only seen in two to five copies per cell, indicating that they are not amplified to the same extent as the amplicons they are associated with.

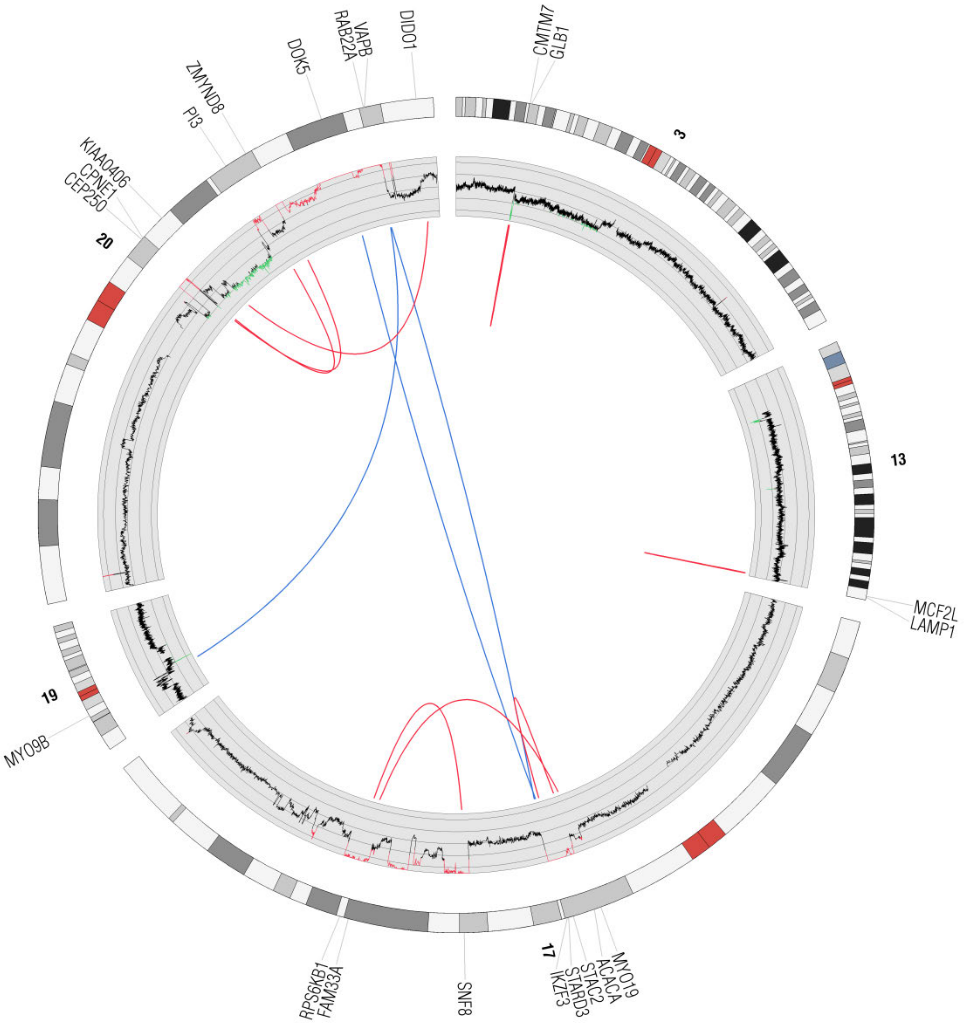


Figure 7 **Association of fusion genes with genomic rearrangements in BT-474.** Circos plot of chromosomes taking part in gene fusions in BT-474. Chromosomes are drawn to scale along the outer rim of the plot, except for chromosomes 17 and 20 which are drawn at five times magnification. The next track inwards shows aCGH data, with amplifications in red and deletions in green. Fusion genes are represented by red (intra-) and blue (interchromosomal) arcs.

Some fusion gene producing rearrangements are quite complex. In the *ERBB2*-amplicon, for instance, six genes take part in gene fusions, some being fused to genes outside the amplicon (figure 8). Other genomic regions contain genes that take part in gene fusions in several different cell lines. The region between 45.8 and 47.8 million bp (Mb) on chromosome 20 contains five genes (*ZMYND8*, *SULF2*, *PREX1*, *ARFGEF2*, *CSE1L*) that take part in gene fusions in BT-474, SK-BR-3 and MCF-7. Despite this clustering, since no gene takes part in a gene fusion in more than one cell line, it is unlikely that the gene fusions are functionally important themselves. A more likely explanation is that the genomic region is a fragile site [98]. Parts or all of the 45.8-47.8 Mb region are also amplified in the three cell lines and the clustering of fused genes may also be driven by the amplification of a gene in the vicinity.

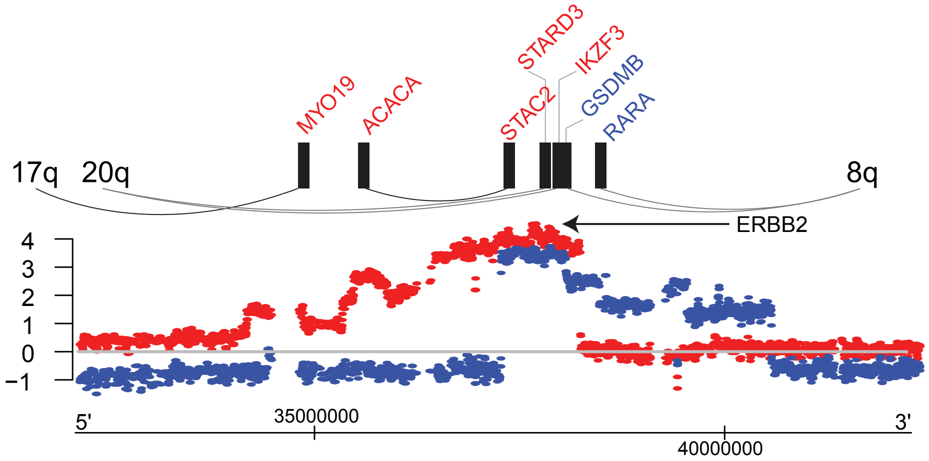


Figure 8 **Genes taking part in fusion events in the ERBB2 amplicon.** aCGH data for BT-474 is shown in red and for SK-BR-3 in blue. The axis below the plot denotes bp positions. The y-axis is on log2 scale. Several genes in the ERBB2 amplicon in both cell lines are fused to either other genes in the same amplicon or other regions of amplification on 8q, 17q and 20q. Genes fused in BT-474 are in red, those fused in SK-BR-3 in blue.

Eight out of 27 fusion genes (*BSG-NFIX, CCDC85C-SETD3, DHX35-ITCH, CMTM7-GLB1, LAMP1-MCF2L, NOTCH1-NUP214, PPP1R12A-SEPT10, SUMF1-LRRFIP2*) were not associated with high-level amplifications. However, in all fusion genes except *PPP1R12A-SEPT10*, one or both partners were associated with low level gains or deletions. This pattern of low level copy number changes is similar to that observed for leukemic translocations [133] and e.g. *TMPRSS2-ERG* in prostate cancer [96].

In addition to the association between fusion genes and copy number changes described above, several other patterns could be discerned. The most common feature is that 23 out of 27 fusion genes were predicted to be in frame, although as fusion genes were detected from mRNA, this was not entirely unexpected. Out of frame fusion genes would likely contain a premature stop codon, leading to degradation of the transcripts by the nonsense-mediated mRNA decay pathway. Given that a few of the out of frame fusion genes, such as *ZMYND8-CEP250*, were highly expressed, it is possible that these are in frame as a result of alternative splicing or secondary mutations that restored correct frame. Second, intrachromosomal translocations were found to be twice as common (19) as interchromosomal ones (8). The same pattern has been reported based on DNA sequencing of breast cancers [106]. In 9 out of 27 fusion genes, the partners are located on opposite strands and must therefore have been fused through inversions (figure 9). Third, some genes were exclusively expressed as parts of fusion genes, since no wt expression could be detected (*IKZF3*, figure 9). Fourth, genes taking part in the fusions contributed both promoters (5' untranslated region (UTR); e.g. *TATDN1-GSDMB*), coding sequences (e.g. *ACACA-STAC2*) as well as 3' UTRs (e.g. *CSE1L-ENSG00000236127*), suggesting they can give rise to both fusion proteins as well as alter protein production by promoter replacement or altered miRNA-based regulation. Fifth, a small number of fusion genes, including *SKA2-MYO19* and *CPNE1-PI3* displayed alternative splicing at the fusion junction.

8.2.5 FURTHER STUDIES ON *VAPB-IKZF3*

To study whether the newly discovered fusion genes were important for cancer cell growth, we performed a siRNA screen using siRNAs targeted against the exons of the 3' partner genes that were included in the fusion transcripts. Based on the screen, *VAPB-IKZF3* in BT-474 was selected for further validation. In addition, no wt *IKZF3* was expressed, simplifying validation experiments since it was not necessary to isolate the effect of fusion gene knock-down from that caused by knock-down of a wt transcript.

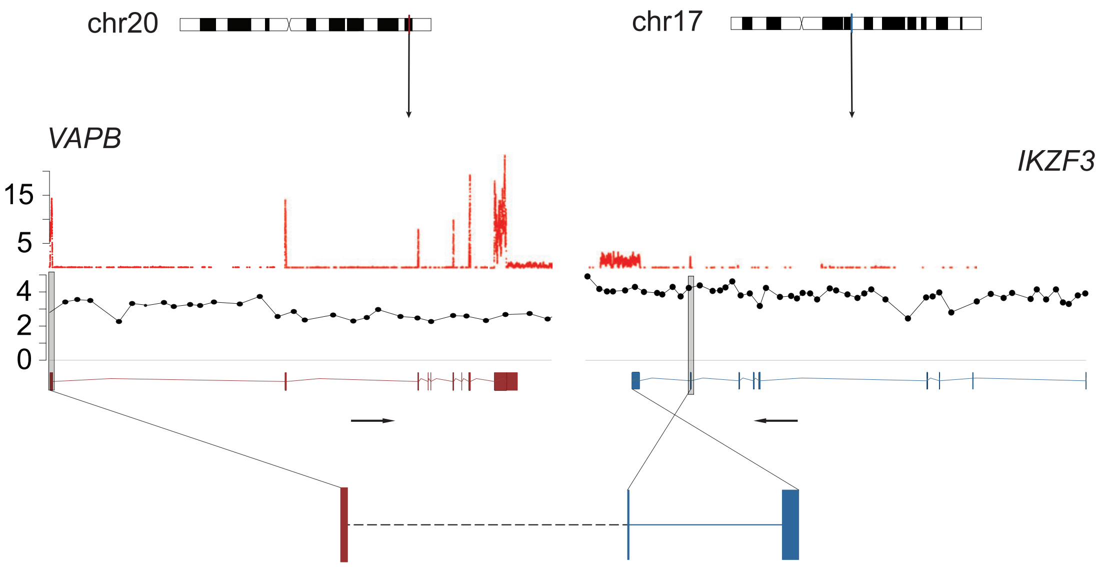


Figure 9 **Fusion of *VAPB* and *IKZF3*.** Gene positions are shown on the ideograms at the top of the figure. The structure of the fusion gene is shown at bottom, with gene structures above it. Black dots represent aCGH data and red dots RNA-sequencing coverage. For *IKZF3*, only the two most 3' exons taking part in the fusion are expressed. As the genes are on opposite strands, the fusion is required to occur via an inversion.

VAPB-IKZF3 arises through a t(17:20)(q12;q13) translocation that fuses the *VAPB* promoter to the 3' part of *IKZF3*, and encodes a protein containing two Zn-finger domains from *IKZF3*. Knock-down of *VAPB-IKZF3* using two different *IKZF3* targeting siRNAs in BT-474 lead to an 80% decrease in fusion gene expression. It also led to a statistically significant inhibition of cancer cell growth (P < 0.001). This suggests that BT-474 growth is dependent on the expression of *VAPB-IKZF3*.

8.3 ERBB2 AMPLICON PROJECT (PUBLICATION III)

8.3.1 ERBB2 AMPLICON SIZE

The primary aim of our project was to study the cancer relevance of other genes in the *ERBB2* amplicon than *ERBB2* itself. We therefore started by studying the size of the amplicon using 244K aCGH arrays in 71 tumors and our panel of cell lines. The average size of the amplicon in tumors was 1.74 mb, with a range of 0.31-13.6 mb.

The minimal common region of amplification was 78.61 kb and included the genes *STARD3*, *TCAP*, *PNMT*, *PERLD1*, *ERBB2* and *MIEN1* (previously known as *C17orf37*) (figure 10). However, most of the tumors (64/71) shared a larger common region of amplification of 255.74 kb. This contained, in addition to the genes above, *NEUROD2*, *PPP1R1B, GRB7* and *IKZF3* (previously known as *ZNFN1A3*). Sixty percent (43/71) of tumors shared a common region of amplification containing 27 genes (928.93 kb), delimited by *RPL19* on the centromeric and *NR1D1* on the telomeric side (figure 10).

Cell lines displayed a very similar pattern of *ERBB2* amplicon sizes, despite a more narrow size range (figure 10). The average size of the amplicon was 1.47 mb, with a range of 0.37-3.29 mb. Trastuzumab-sensitive and -resistant cell lines did not differ in amplicon size. The mean amplification height was somewhat higher in resistant cell lines (mean log2 ratio 4.59 in resistant, 3.54 in sensitive). However, as our study only identified two sensitive cell lines, this result should be taken with caution.

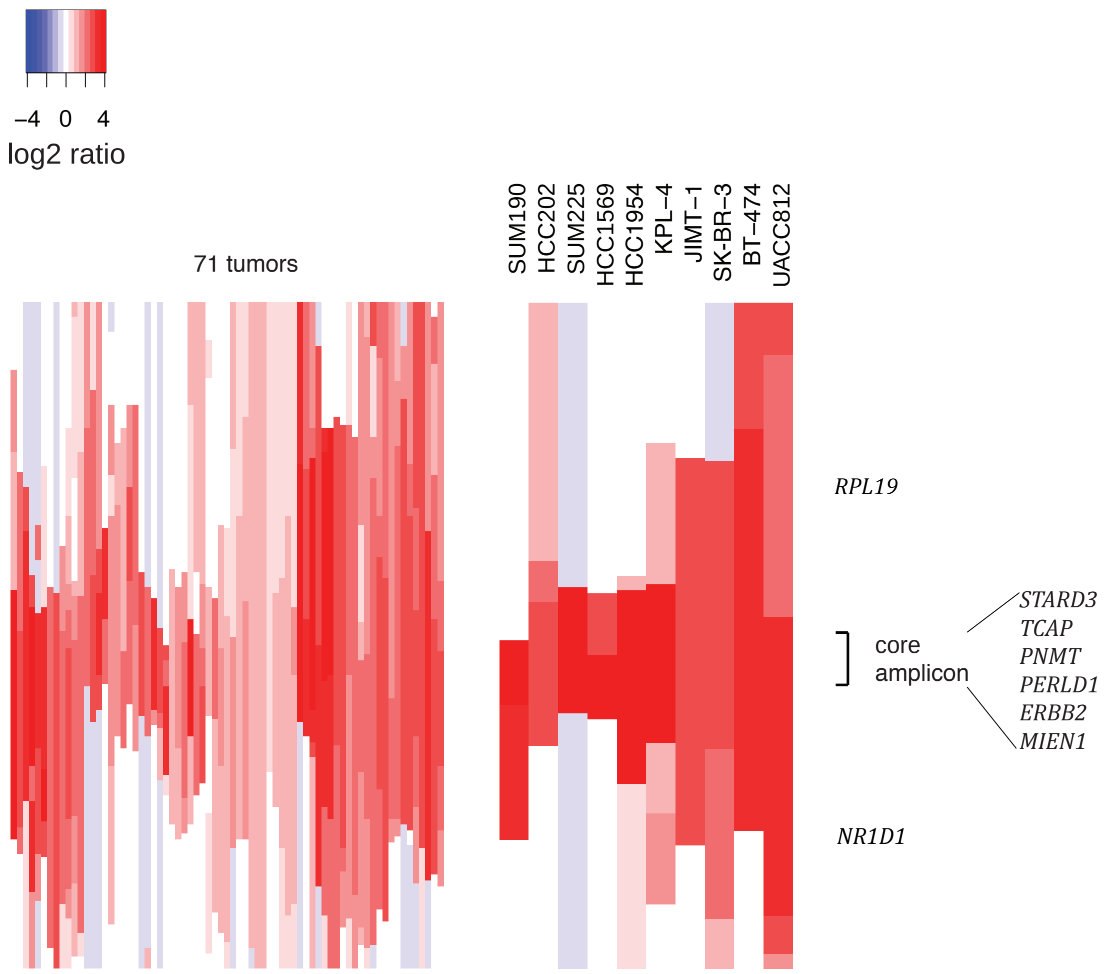


Figure 10 ***ERBB2* amplicon size in clinical tumors and cell lines.** Tumor aCGH data is shown on the left, cell lines on the right. The region from which genes were selected for siRNA screening is delimited by *RPL19* and *NR1D1*. The minimal common region of amplification is shown on the right.

8.3.2 TRASTUZUMAB RESPONSE AND PIK3CA MUTATION ANALYSIS

Two out of nine breast cancer cell lines, BT-474 and SK-BR-3, were sensitive to trastuzumab with half maximal effective concentrations (EC_{50}) of 1.02 nM +/- 0.71 nM and 0.98 nM +/- 0.08 nM, respectively. The corresponding decreases in cell viability were 26.7% and 34.3%. KPL-4, JIMT-1, HCC1954, HCC1569, HCC202 and SUM225 did not respond to trastuzumab (2 - 18% decrease in cell viability). SUM190 had a 22.3% decrease in cell growth in response to trastuzumab, but as it had an EC_{50} value of 0.67 mM +/- 0.6 mM, it was considered resistant.

Mutations in the gene *PIK3CA* have been linked to trastuzumab resistance [210]. To study a possible link in our cell lines, we sequenced exons 9 and 20, which contain the majority of *PIK3CA* mutations, in 11 cell lines. Both trastuzumab-responsive cell lines, BT-474 and SK-BR-3, as well as non-responsive cell lines HCC1569, JIMT-1, UACC812 and SUM225 had wt sequence. The non-responsive cell lines HCC202 (exon 9), HCC1954, SUM190, KPL-4 (all exon 20) and the control cell line MCF-7 (exon 9) carried mutations. JIMT-1 has previously been reported to have a mutation in *PIK3CA* exon 7 [52].

8.3.3 SILENCING THE AMPLICON GENES AND EFFECTS ON CELLULAR SIGNALING.

For both biological and practical reasons, we decided to concentrate our siRNA screen on those 27 genes that were amplified in 60% of tumors. siRNAs were available for 23/27 genes, and 2-4 siRNAs were utilized for each gene, depending on availability. Two trastuzumab sensitive, BT-474 and SK-BR-3, two resistant JIMT-1 and KPL-4 as well as one control cell line, MCF-7, were subjected to siRNA screening with cell viability as the endpoint. Target gene silencing was separately validated for *ERBB2*, *GRB7*, *STARD3*, *PERLD1*, *PSMD3* and *PPP1R1B* using Taqman in SK-BR-3 and KPL-4, with an average knock-down of 80% for the six genes.

As expected, *ERBB2* silencing decreased cell viability in the trastuzumab-sensitive cell lines. No significant effect was seen in the resistant cell lines. This was separately confirmed by fluorescence-activated cell sorting, which showed that *ERBB2* silencing caused a significantly larger G1 cell cycle phase arrest in the sensitive cell lines. This suggests that KPL-4 and JIMT-1 are not only resistant to the specific effects of the drug trastuzumab, but are overall less dependent on the *ERBB2* pathway for their viability.

The LMA technique was used to study the impact of silencing the *ERBB2* amplicon genes on other cell signaling endpoints than cell viability. Antibodies against HER2, cPARP, Ki67, pAKT, pS6K and p27 Kip1 were used in the LMA experiments to quantify either the amount or phosphorylation status of key signaling proteins. As expected, after *ERBB2* silencing, the Ki67 proliferation marker was downregulated more in trastuzumab-sensitive cell lines when compared to the resistant. Trastuzumab sensitive cell lines also had lower phosphorylation of AKT and S6K following *ERBB2* silencing, which fits with a previous study which showed that trastuzumab treatment decreased pAKT in BT-474 but not JIMT-1 [211]. Silencing of *ERBB2* was equally effective at decreasing HER2 protein in both resistant and sensitive cell lines, confirming the functionality of siRNAs in the cell lines used. Among the other genes in the *ERBB2* amplicon, silencing of *STARD3* led to an increase in cPARP in three out of four cell lines (SK-BR-3, JIMT-1, KPL-4), suggesting an increase in apoptosis. *PPARBP* silencing altered cell cycle regulation (p27) in the resistant cell lines. Silencing of both *GSDM1* and *PSMD3*, had an effect on several of the signaling markers in all cell lines. This suggests an important role in cancer cell signaling irrespective of whether the cells are trastuzumab sensitive or not. Alternatively, as the non-*ERBB2* amplified cell line MCF-7 also showed similar results, the genes may be generally important for the measured signaling pathways also in *ERBB2* negative breast cancer cells.

8.3.4 CO-SILENCING EXPERIMENTS AND THEIR RESULTS

PPP1R1B, *STARD3*, *PERLD1*, *GRB7* and *PSMD3* were selected for co-silencing experiments together with *ERBB2*. Our hypothesis was that coamplification of these genes is either directly beneficial for the cancer cell or, alternatively, the cell becomes dependent on their functions later during its evolution, as a form of non-oncogene addiction (see section 5.1.1). Generally, the trastuzumab-sensitive cell lines were more likely to show synergistic effects in the double silencing experiments, which fits their higher dependence on *ERBB2* overall. Synergistic effects were, however, detected for some combinations in all cell lines, whether trastuzumab-resistant or -sensitive. As expected, the cell line MCF-7, which does not contain an *ERBB2* amplification, showed barely any synergistic effects in the double silencing experiments.

Silencing of *PPP1R1B* alone, for instance, had no effect on phosphorylation of AKT, but co-silencing together with *ERBB2* caused a synergistic inhibition of the AKT pathway, as measured by both AKT and S6K phosphorylation, as well as decreased cell viability. This effect was seen in all four *ERBB2* amplified cell lines. Silencing of *ERBB2* together with either *STARD3* or *PERLD1* also synergistically inhibited both AKT signaling (downregulation of pAKT and pS6K) and proliferation (downregulation of Ki67) in several of the cell lines. Co-silencing of *ERBB2* and *GRB7* affected both cell viability and pAKT in a synergistic manner in the sensitive cell lines. This fits previous work, which has shown that silencing of *GRB7* leads to decreased cell viability, decreased pAKT and enhances the response to the HER2 inhibitor lapatinib [212].

9 DISCUSSION

9.1 DEVELOPMENT OF GENOMIC TECHNOLOGIES

The projects that form this thesis have been undertaken over a six year period, starting with study I in late 2005. The overarching theme has been the development of bioinformatic methods and their application to breast cancer research, especially the identification and characterization of somatically mutated or amplified genes. The new bioinformatic methods and biological results presented here expand our knowledge about the biology of breast cancer, especially with regards to fusion genes and how they form. During the same period of time, genomic technologies, such as microarrays, have developed significantly and completely new techniques, such as high throughput sequencing, have sprung into existence. For instance, aCGH microarrays have developed from cDNA or BAC based low resolution arrays to *in situ* synthesized arrays that interrogate up to 1.000.000 different genomic positions. This has necessitated the development of, among others, more computationally efficient segmentation algorithms to deal with the increased number of data points [213]. The first next generation sequencing publications were published in the fall of 2005 [214, 215], and the first full cancer genome sequence was reported in late 2008 [216]. As the sequencing output of these methods is orders of magnitude larger than that achieved with Sanger sequencing, this has required the development of new and significantly faster approaches to sequence alignment [195, 217]. These in turn have enabled the development of new bioinformatic approaches to specific biological questions, exemplified e.g. by the fusion gene identification method described in study II. Technology development has therefore also significantly driven the need to develop and refine new bioinformatic methods.

9.2 *RIC8A* AND THE GINI METHODOLOGY (PUBLICATION I)

Study I lead to the identification of biallelic inactivation of *RIC8A* in the breast cancer cell line ZR-75-1, as well as found a link between low *RIC8A* expression and aggressive breast cancer. We were unable to identify additional mutations in *RIC8A* when sequencing it in 127 early stage breast cancers. Curation results for *RIC8A* mutations in 515 cancers of various types (primarily glioblastoma, medulloblastoma, pancreatic, breast and colorectal cancers) reported in version 57 of the COSMIC database [218] identified one somatic missense mutation in a squamous cell carcinoma of the skin [219]. It therefore seems that *RIC8A* is mutated at low frequency or the mutation may be private to the individual from which the ZR-75-1 cell line was derived. Alternatively, it may be mutated primarily in late stage breast cancers not included in our validation set. The rarity of *RIC8A* mutations also raises the possibility that *RIC8A* may be a passenger mutation with no cancer relevance. However, our results showing biallelic inactivation of *RIC8A* and low gene expression being associated with more aggressive breast cancer argues against this. Recent publications have shown that *RIC8A* plays a role in both orienting the mitotic metaphase spindle [220] and is required for coupling receptor tyrosine kinases to actin skeleton remodeling [221]. This indicates a role for it in the central cancer processes of cell division and migration, though in the latter case, *RIC8A* inhibition puzzlingly inhibits cell migration in response to PDGF and EGF [221]. Examining mutations on the level of pathways has shown that signaling pathways can be altered at many different nodes, with a large variation in how frequently individual nodes are mutated [222]. In these kinds of approaches, private and rare mutations such as *RIC8A* can contribute to the identification of the altered pathways and therefore be more important than their low frequency alone suggests.

Despite the successes of the GINI method in finding mutated genes [145, 172, 173], Buffart *et al.* have critisized it for producing a high number of false positive hits that do not validate in sequencing [174]. Our experience has also been that the NMD block leads to stabilization of a large number of transcripts, many of which are likely normal physiological targets of NMD. This observation led to us to include the data analysis criteria that a gene should only be upregulated by NMD inhibition in one cancer cell line and none of the non-transformed cell lines, essentially excluding most physiological NMD targets or transcripts otherwise upregulated by emetine treatment. This greatly decreased the number of false positive candidates. In principle, the specificity of the GINI method should increase as more samples, either cancers or "normal" cell lines, are added to the analysis. The major limitation of this approach is that the same criteria, if strictly applied, limits one to the identification of genes inactivated by PTC mutations only in one of the cancer cell lines in a study. However, the same gene can be inactivated by other types of mutations in other samples. Several recent publications have also shown that many genes are mutated at low frequencies or are unique to the tumor in which they are found [223-225], with few genes reaching even a 10% mutation frequency. This criterion is therefore unlikely to prevent the identification of most mutated TSGs, while significantly increasing the likelyhood of finding true positive PTC mutations.

These days, next generation sequencing is videly available and allows the validation of many more NMD candidate genes in the same experiment, circumventing many of the problems associated with validating a large number of candidate mutated genes. However, the cost of exome sequencing is rapidly approaching the point at which it will be cheaper to simply sequence all exons in the samples, finding both nonsense as well as other types of mutations. This suggests that the NMD microarray method may not remain in use for the purpose of mutation discovery for very much longer.

9.3 FUSION GENES IN BREAST CANCER (PUBLICATION II)

The same arguments regarding *RIC8A* cancer relevance based on mutation frequency may also be true for many of the fusion genes identified in publication II, although this is conjecture as little or no validation data in tumors is available for them. The largest study on fusion genes in breast cancer to date [119] analyzed 41 breast cancer cell lines and 38 tumors using paired-end RNA-seq and Illumina sequencing. Even though they reported no highly recurrent fusion genes, they found repeated fusions involving NOTCH and MAST family members in several breast tumors and cell lines, suggesting that they together may be present in 5-7% of breast cancers. In the supplementary material, Robinson et al [119] reported 384 fusion gene candidates from the 79 cancer samples. Besides *NOTCH1,* another 17 of the fusion partner genes identified in study II are listed in the appendix to Robinson *et al.*: *STARD3, EIF3H, CYTH1, SNF8, TMEM49, GSDMB, MCF2L, ACACA, IKZF3, RAB22A, DIDO1, NFIX, RARA, ARFGEF2, ITCH, BSG* and *ZMYND8*, suggesting they may also be recurrently fused. However, Robinson *et al.* did not validate most of the 384 fusion gene candidates. Since the paper does not discuss the specificity of their fusion gene identification pipeline, it remains unclear how many of the fusion gene candidates they report are true and therefore whether the 17 genes listed above are genuinely recurrent. As a major criterion for defining driver oncogenes and mutations is observing them repeatedly, the question remains open whether these are driver mutations or not. Of the genes found fused in study II, *STARD3, SNF8, MCF2L, ACACA* and *NFIX* were each validated once by Robinson *et al.*, suggesting they may be recurrently fused.

In addition to the five genes listed above, Robinson *et al.* promisingly reported two additional validated instances of *IKZF3* fusions; the fusion *CDC6-IKZF3* in UACC812 and *MED1-IKZF3* in UACC893. This brings the total of observed *IKZF3* gene fusions to three, which together with the functional data reported in study II increasingly suggest *IKZF3* fusion genes play a role in a subset of breast cancer. One possible caveat is that all three cell lines with *IKZF3* fusions contain the *ERBB2* amplification. As *IKZF3* is located ~29kb from *ERBB2*, the two genes are frequently coamplified, and the *IKZF3* locus has been reported to be the most common telomeric breakpoint of the *ERBB2* amplicon [193]. *IKZF3* gene fusions may therefore in principle be byproducts of *ERBB2* amplification, e.g. because the locus is a fragile site at which DSBs preferentially occur during amplicon formation. Nevertheless, the observation that *ERBB2* amplicons frequently end at *IKZF3* can also be interpreted to support the importance of *IKZF3*;

breakpoint clustering would indicate the location of an important gene. Further study of *IKZF3* fusion genes in breast cancer therefore seems a promising project, both in terms of *IKZF3* fusions themselves, as well as the possibility that it could point to a more general importance for fusion genes in high level amplifications.

9.4 FUSION GENE IDENTIFICATION (PUBLICATION II)

In study II, our fusion gene detection pipeline achieved near perfect specificity, and the pipeline has since been used to analyze several clinical RNA-seq samples, mainly leukemias, achieving comparable specificity (unpublished). The main step that provides high specificity in our pipeline is the pattern of short read alignments across the fusion junction, which has subsequently also been noted by others [187]. False positive fusion gene candidates that did not pass this filtering step were almost universally supported by short reads that aligned the minimum number of base pairs across the fusion junction (10bp in study II), as well as contained several mismatches in the 10bp that aligned to the other exon. This suggests that false positive rate could be further controlled by both increasing the required length of junction overlap, as well as by limiting the number of allowed mismatches. The downside would, however, be that some genuine fusion genes might be missed. Either due to mutations or single nucleotide polymorphisms close to the exon-intron junction creating mismatches indistinguishable from mismatches in misaligned reads. Alternatively, some true junction spanning short reads could be excluded because they do not sufficiently overlap the fusion junction. In either case, a fusion gene expressed at a low level might not be identified.

Besides specificity, the sensitivity of an analysis method is of equal importance. The minimum number of fusion mRNA:s that must be sequenced to generate the required number of paired-end reads (2 for MCF-7 and KPL-4, 3 for BT-474 and SK-BR-3) places a theoretical lower bound on the abundance a fusion mRNA must have to be detected in each of the breast cancer cell lines. This abundance is a direct function of sequencing depth and the number of read pairs required to support a fusion. KPL-4 has roughly 3.989.756 uniquely aligning paired-end read pairs and, since two paired-end reads was the limit for detection, a fusion mRNA must be present at at least one fusion mRNA per 1.994.878 paired-end reads. This is ~1/3 of the theorethical sensitivity for SK-BR-3; 1 fusion mRNA per 5.355.302 paired-end reads. In practice, however, the sensitivity will clearly be lower, but the numbers give an indication of the differences in sensitivity between the cell lines. As all the fusion genes are not known in any of our samples, a precise sensitivity can not be given, though others have found largely the same fusion genes when reanalyzing either our data [187] or conducting independent experiments on the same cell lines [114, 116], as well having found new fusion genes. Our reanalysis of the RNA-seq data from study II in the spring of 2011 also identified several new fusion genes, many of which have also been found by e.g. Kim *et al.* [187]. Direct comparison of the sensitivity of different analysis pipelines is complicated if they use different sets of transcript definitions, such as those provided by Ensembl and RefSeq. Transcript structure definitions may also change between versions of the same database. The fusion gene *CSE1L-ENSG00000236217*, for instance, was not identified by Kim *et al.* [187], as *ENSG00000236217* had been removed from the Ensembl database version they used. Out of the new fusion genes found in our reanalysis, some were not found in study II because it used older versions of the Ensembl database. The other main reason was that, in our reanalysis, we relaxed the criteria for how many gene fusions the same gene could take part in in the same sample, which previously caused us to miss e.g. several *MED1* fusions in BT-474. More generally, most fusion gene detection pipelines are based on the current knowledge of gene and transcript structures and will not find fusion genes involving nonannotated exons or genes. Periodic reanalysis of RNA-seq data sets for fusion gene identification is therefore advisable as long as the identification of new transcript variants causes gene models to evolve. TopHat-fusion [187] is the main exception, in that it can detect fusion genes involving unknown exons or genes, as long as one of the fusion gene partners is a known gene from RefSeq.

9.5 THE BIOLOGICAL CHARACTERISTICS OF FUSION GENES (PUBLICATION II)

One of the central findings in publication II was the strong link between fusion genes and high level amplifications on one hand and lower level copy number alterations on the other. Traditionally, recurrent translocations observed in leukemias were considered balanced, in that no DNA was lost in the process. This now seems more a product of the limited resolution of techniques such as G-banding, as a large fraction of leukemic translocations have also been shown to involve low level copy number changes, typically gains of the fused parts of both genes as well as losses of the parts not taking part in the fusion [133]. More broadly, it seems likely that a large number of amplicon associated fusion genes will be identified in cancer types that typically contain high level amplifications. If that turns out to be true, some of the genes will, if the sample set is large enough, be recurrently fused simply as a function of the large number of DSBs occuring in the vicinity of the driving oncogene during amplicon formation. Therefore, distinguishing these random recurrent events from recurrent oncogenic fusion genes associated with the amplicon may require functional experiments.

In study II, we identified two gene fusions, *CSE1L-ENSG00000236217* and *ANKHD1-PCDH1*, in which the 3' partner gene primarily contributed its 3' untranslated region (UTR) to the fusion gene. For fusion genes such as these, the primary result of the rearrangement may not be a gene with altered protein function. Instead, the result could be altered miRNA mediated regulation of the 5' gene due to 3' UTR swapping, analogous to e.g. the t(8;14) translocation that place the *MYC* oncogene next to the highly active immunoglobulin heavy chain gene promoter [226]. Gene fusion may also be a mechanism for TSG inactivation, either by disrupting the allele taking part in the fusion gene or by creating a dominant negative version of the TSG. This has e.g. been suggested for *SULF2*, a potential breast cancer TSG [227], in the *ARFGEF2-SULF2* gene fusion in MCF-7 [206]. Although as *SULF2* has also been reported to be an oncogene in hepatocellular carcinoma [228], the question is open whether this is a true TSG inactivating event.

Just as the pattern of single nucleotide or other mutations seen in a tumor may reveal the type of mutagens that initiated the tumor or what kinds of DNA repair mechanisms have been active [229], so too can the type of copy number changes associated with a gene fusion point to the mechanism that led to its formation. Gene fusions associated with high level amplifications are likely to have arisen together with the amplicon, either as the selectively advantageous target of the amplicon or as a byproduct of double strand breaks occurring during repeated rounds of e.g. breakage-fusion-bridge cycles. This model would imply a continuous accumulation of new fusion genes as the magnitude of an amplicon increases, since the DSBs generated during each round of amplification provide new opportunities for fusion formation. On the other hand, fusion genes associated with low level copy number changes may arise through incorrect repair of DSBs occurring as a consequence of mutagen exposure or during normal cellular functions. Especially if they lie in heavily rearranged genomic regions, the DSBs may represent chromothripsis [104]. The causes of chromothripsis, i.e. many chromosomal rearrangements ocurring in one or a few narrow regions of the genome, are not known, but speculated to be caused by e.g. ionizing radiation exposure during mitosis when chromosomes are condensed.

9.6 THE *ERBB2* AMPLICON (PUBLICATION III)

Sircoulomb *et al.* [193] identified a minimal common region of amplification consisting of only 3 genes: *ERBB2, MIEN1/C17orf37* and *GRB7*, while our reanalysis of their data together with 17 additional tumors in study III identified a minimal common region containing six genes (*STARD3, TCAP, PNMT, PERLD1, ERBB2* and MIEN1/C17orf37), which is in accordance with previous studies [230]. This is a discrepancy, as adding more samples should, at most, further narrow the minimal common region, not enlarge it. The most likely explanation is the use of different segmentation methods, Genomic Identification of Significant Targets in Cancer [231] in Sircoulomb *et al.* and Piecewise Constant Fit in study III [194]. If data is noisy or amplicon boundaries not well defined in some samples, breakpoint definition can be difficult and small differences in algorithm behaviour may lead to differences in results. Especially when defining a minimal common region, differences in the result for a single sample may be sufficient, if it happens to be the one defining either border of the common region.

Few, if any, tumors however amplify only the minimal common region. Our results in study III showed that 10 genes were amplified in 90% of tumors and 60% of tumors amplified a region containing 27 genes around *ERBB2*. The concept of non-oncogene addiction suggests that, in individual tumors, several genes in the amplicon may contribute to the cancer phenotype, even if they are not located in the minimal common region. If that is true, then their inhibition may be therapeutically beneficial. Our results in study III showed that several genes in the *ERBB2* amplicon are needed for survival of amplicon containing cells, lending support to the notion that several genes in an amplicon may be important for cancer development. The results showing synergistic effects on AKT phosphorylation in practically all *ERBB2* amplified cell lines when silencing *ERBB2* together with *PPP1R1B, STARD3* or *PERLD1*, also supports this theory, as well as points to the importance of the PI3K/AKT signaling pathway as a downstream target of several genes in the amplicon.

One limitation of any siRNA screen is the number of screen endpoints that can be detected. An siRNA screen will typically have only one or a few endpoints, and may require replicate experiments if the endpoints can not be measured from the same wells. The LMA technique used in study III allowed us to measure several additional signaling endpoints from the same individual screening plates, significantly expanding the scope of our screen beyond the primary cell viability endpoint.

One of the findings in study III is that while the trastuzumab resistant cell lines JIMT-1 and KPL-4 showed little or no effect when *ERBB2* alone was silenced, we were still able to see synergistic effects on AKT signaling in all cell lines when it was silenced together with e.g. *PPP1R1B*. Köninki *et al.* [52] have previously linked trastuzumab resistance in JIMT-1 to activation of the PI3K-AKT pathway by both a mutation in *PIK3CA* exon 7 as well as low PTEN expression. This suggests activation of *ERBB2* signaling downstream of the HER2 protein, thereby possibly making its presence in the cells unimportant. Even if the synergistic effect was somewhat lower in the resistant cell lines, it is still a puzzling observation, as it implies that *ERBB2* still has a function in the cells that have become resistant to its inhibition. One possible model is that in e.g. JIMT-1, the alterations in *ERBB2, PIK3CA* and PTEN all feed into the same downstream signaling network in an additive manner. As a result, loss of the oncogenic signal from any one of the three aberrant proteins would be largely buffered by the two others, thereby resulting in resistance to inhibition of *ERBB2* alone. In this scenario, the cells would be resistant to *ERBB2* inhibition while simultaneously, when additional components of the downstream signaling network are inhibited, still dependent on it.

10 CONCLUSIONS AND FUTURE PROSPECTS

My reasons for studying the genomic alterations found in breast cancer have been twofold. Pragmatically, oncogenic mutations make for good drug targets, in that the vulnerability is not present in normal cells. This hopefully provides a therapeutic window, enabling interference with the functions of the cancer cell without significantly affecting normal cells. Even if a mutation is not directly druggable, the more we know about the biology of breast cancer, the greater our chances are of better tailoring existing treatments for individual patients, as well as inventing completely new therapies. The other reason, certainly less altruistic, is scientific curiosity; a drive to figure out things that are unknown. Much like climbing a mountain because it is there.

One of the primary lessons I have learned during this project is the need to document bioinformatic analyses in an understandable and reproducible fashion. As has been noted by Ioannidis *et al.* [232], many published articles do not contain enough details to allow independent bioinformatic reanalysis of the data and exact reproduction of the published results. Even though we have tried to document bioinformatic analyses in detail and have made all data publicly available, the publications that make up this thesis also fall short of the full disclosure model proposed by e.g. Baggerly and Coombes [233]. Irrespective of how one publishes methods, this question is equally acute when it comes to how you document your own analyses, such that you can reproduce them years later, or at least understand what was done in detail. Since 2005 when work on publication I was started, I have repeatedly needed to go back to examine minute details of how something was done, often years before, to confirm that I had not made any errors. If I had not saved the code of practically every analysis, included comments on what I was doing and organized this in a version control system, I would frequently have been lost.

As more and more laboratory methods come to depend on bioinformatic analysis of their results, there is a danger that significant fractions of molecular biological research will become essentially nonreproducible, unless steps are taken towards more transparent publication of the data analysis methods used in a paper. In practice, this would mean requiring publication of the code used to analyze the data. Frameworks for doing this exist, either using literate programming tools such as Sweave [234] or as workflows in e.g. GenePattern [235] or Galaxy [236]. My own goal is to move towards publishing the code in future projects, and I hope that, just as many journals and funding bodies now require that raw data is made public, they will also start requiring that data analysis methods are published in a reproducible and transparent fashion.

It has for some time been clear that most cancers can in practice be divided into several different subtypes, with varying prognoses and treatment responses. In parallel, the recent explosive growth in sequencing capacity and drop in costs has made it both technically and financially feasible to sequence the whole exomes, transcriptomes or even genomes of cancer patients. This has lead to the establishment of projects aiming to personalize cancer treatment through large scale sequencing of cancer and corresponding germline genomes in many countries. The low hanging fruit in these personalized oncology projects will be the comprehensive identification of essentially all currently clinically actionable mutations in a tumor at a reasonable cost. Already today, exome sequencing of a tumor and the corresponding normal DNA is less expensive than 4-6 single gene tests, suggesting these projects will provide costs savings both in diagnostics as well as by better identifying those patients that would benefit from, often expensive, targeted drugs. In the slightly longer term, these projects will identify practically all recurrent cancer mutations, at least in more common cancer types. This information, together with detailed information on treatment histories should uncover many links between specific sets of mutations and treatment success, thereby improving cancer treatments. These projects will also generate very large sets of data on different cancer types, expanding our knowledge about the molecular biology of cancer and hopefully leading to the development of completely new therapies.

This undertaking will, however, not be quite as straightforward as described above. In any sufficiently large collection of data, many spurious correlations will show up simply due to chance. Rigorous validation of links between specific sets of mutations and treatments through clinical trials should therefore be a priority. More generally, the problem, often sidestepped in "large data" projects, is that data alone does not equal knowledge. To quote the author Henning Mankell, "*Many people make the mistake of confusing information with knowledge. They are not the same thing. Knowledge involves the interpretation of information.*" [237]. It is therefore unreasonable to expect or present these sequencing projects as revolutionizing cancer treatment on their own. They will likely generate important advances in cancer treatment but, equally importantly, they will provide raw material for new hypotheses and targeted studies for years to come.

11 ACKNOWLEDGEMENTS

This project has been carried out over a period of ten years and multiple different institutions. For the early years, I would like to thank Biomedicum Biochip Center, as it was known back then, as well as the Medical Biotechnology department of VTT and the University of Turku. For the last several years, this project has been carried out in the Genome Scale Biology research program in Biomedicum and at the Institute for Molecular Medicine Finland (FIMM). I want to thank all my colleagues at FIMM for making it an inspiring place to work in, as well as the administrative staff, in particular Reetta and Susanna, for making working there easy. I also want to thank my *alma mater*, the University of Helsinki, a central part of my life for 16 years. I am also grateful for the financial support received from the Academy of Finland, Sigrid Juselius Foundation, Cancer Organizations of Finland and Arvid och Greta Olins fond.

My heartfelt thanks to all the people that have contributed to this study, especially:

My supervisor Olli Kallioniemi. It has been a great privilege to work with you for all these years. I am still amazed at your ability to, as if from thin air, conjure up visionary ideas on nearly any topic. You have served as a role model in always being willing to share your ideas and knowledge with others, as well as fostering an atmosphere in which others follow your example. You have also gone far beyond the call of duty in providing excellent working facilities and financial support for this project. For all these things, and many more, my warmest thanks.

The official reviewers, Minna Tanner and Garry Wong for their valuable and insightful comments on my thesis.

My many coauthors: Anne-Lise Børresen-Dale, Emelyne Dejeux, Eldri U Due, Kirsi Hellström, Vesa Hongisto, Sara Kangaspeska, Kristine Kleivi Sahlberg, Aslaug Muggerud, Astrid Murumägi, Rami Mäkelä, Daniel Nicorici, Sandra Nyberg, Merja Perälä, Inga H Rye, Niko Sahlberg, Therese Sørlie, Jörg Tost, Hans Kristian Moen Vollan and Maija Wolf. This project would have come to nothing without your contributions. Aslaug Muggerud for your hard work and dedication, including travelling abroad to learn new techniques, in pushing the NMD project forward. Sara Kangaspeska and Astrid Murumägi for the rare kind of collaboration in which the whole becomes much bigger than the sum of its parts. Daniel Nicorici for your work on creating the fusion gene identification pipeline. Vesa Hongisto and Kristine Kleivi for your enthusiasm and laboratory wizardry in all things

breast cancer related. Anna Järvinen for sharing the early steps of PhD work with me and remaining a friend ever since. Outi Monni for providing facilities and support when this project was starting, continued valuable and friendly advice throughout the work as well as giving me the opportunity to teach. Sampsa Hautaniemi for many interesting and entertaining discussions on bioinformatics and everything else. Carl Blomqvist for helping with my thesis introduction. All members of the extended Kallioniemi group over the years, many thanks!

Sami Kilpinen and Kalle Ojala for being colleagues and close friends for many many years. Without your company, lunch breaks, retreats and parties would have been much less fun and, without your sharp minds, my scientific ideas riddled with many more holes.

Maija Wolf, for being invaluable in so many ways, whether as a mentor, colleague, scientific collaborator or the one who always knew how to get the practical things done. I would not be where I am now without you.

My friends from our studying days, for all the fun we have had when not studying. Jacke and Petter for our many hikes all over arctic Scandinavia. Nick, Viska, Make and Niko, there is no sound like that of rolling dice.

My extended family on both sides, no occasion is too insignificant not to be used as an excuse to get together, celebrate and enjoy each other's company. Mummu Flora for taking care of me so many times when I was young and for continued support to this day. My brother Johannes for countless fun and exciting escapades over the years.

Dear mother and father. I would need another 75 pages to list all the things I am grateful for. Thank you for your loving and unwavering support during this project in particular and my whole life in general.

Last, but certainly not least, my beloved wife Elli and our "lilla gumman" Freja. Thank you Elli for all your support during this project and for making my life so much better in innumerable ways. Thank you Freja for bringing a smile to my face every day, and for reminding me that there are more important things than work.

Helsinki, June 2012
Henrik Edgren

12 REFERENCES

1. Registry, F.C., *Cancer in Finland 2008 and 2009*. Cancer Society of Finland Publication. Vol. 84. 2011, Helsinki: Cancer Society of Finland.
2. Li, C.I., D.J. Uribe, and J.R. Daling, *Clinical characteristics of different histologic types of breast cancer.* Br J Cancer, 2005. 93(9): p. 1046-52.
3. Li, C.I., *Risk of mortality by histologic type of breast cancer in the United States.* Horm Cancer, 2010. 1(3): p. 156-65.
4. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. 98(19): p. 10869-74.
5. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. 406(6797): p. 747-52.
6. Parker, J.S., et al., *Supervised risk predictor of breast cancer based on intrinsic subtypes.* J Clin Oncol, 2009. 27(8): p. 1160-7.
7. Key, T.J., P.K. Verkasalo, and E. Banks, *Epidemiology of breast cancer.* Lancet Oncol, 2001. 2(3): p. 133-40.
8. Shuen, A.Y. and W.D. Foulkes, *Inherited mutations in breast cancer genes--risk and response.* J Mammary Gland Biol Neoplasia, 2011. 16(1): p. 3-15.
9. Fitzgibbons, P.L., et al., *Prognostic factors in breast cancer. College of American Pathologists Consensus Statement 1999.* Arch Pathol Lab Med, 2000. 124(7): p. 966-78.
10. Cardoso, F., et al., *Clinical application of the 70-gene profile: the MINDACT trial.* J Clin Oncol, 2008. 26(5): p. 729-35.
11. Albain, K.S., 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): p. 55-65.
12. MINDACT, E.T.B.-.-. Available from: http://www.eortc.be/services/unit/mindact/MINDACT_websiteii.asp .
13. Paik, S., et al., *A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.* N Engl J Med, 2004. 351(27): p. 2817-26.
14. Hicks, J., et al., *Novel patterns of genome rearrangement and their association with survival in breast cancer.* Genome Res, 2006. 16(12): p. 1465-79.
15. Chang, H.Y., et al., *Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival.* Proc Natl Acad Sci U S A, 2005. 102(10): p. 3738-43.

16. (EBCTCG), E.B.C.T.C.G., *Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100 000 women in 123 randomised trials.* Lancet, 2011(Ahead of Print).
17. Darby, S., et al., *Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10,801 women in 17 randomised trials.* Lancet, 2011. 378(9804): p. 1707-16.
18. Davies, C., et al., *Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials.* Lancet, 2011. 378(9793): p. 771-84.
19. Perez, E.A., et al., *Sequential versus concurrent trastuzumab in adjuvant chemotherapy for breast cancer.* J Clin Oncol, 2011. 29(34): p. 4491-7.
20. Bray, F., et al., *Estimates of cancer incidence and mortality in Europe in 1995.* Eur J Cancer, 2002. 38(1): p. 99-166.
21. Berrino, F., et al., *Survival for eight major cancers and all cancers combined for European adults diagnosed in 1995-99: results of the EUROCARE-4 study.* Lancet Oncol, 2007. 8(9): p. 773-83.
22. Tryggvadottir, L., et al., *Trends in the survival of patients diagnosed with breast cancer in the Nordic countries 1964-2003 followed up to the end of 2006.* Acta Oncol, 2010. 49(5): p. 624-31.
23. Autier, P., et al., *Advanced breast cancer and breast cancer mortality in randomized controlled trials on mammography screening.* J Clin Oncol, 2009. 27(35): p. 5919-23.
24. Tabar, L., et al., *Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare.* Lancet, 1985. 1(8433): p. 829-32.
25. Slamon, D.J., et al., *Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2.* N Engl J Med, 2001. 344(11): p. 783-92.
26. Weinberg, R.A., *The Biology of Cancer*2007, New York: Garland Science, Taylor and Francis Group. 796.
27. Fontham, E.T., et al., *American Cancer Society perspectives on environmental factors and cancer.* CA Cancer J Clin, 2009. 59(6): p. 343-51.
28. Moore, P.S. and Y. Chang, *Why do viruses cause cancer? Highlights of the first century of human tumour virology.* Nat Rev Cancer. 10(12): p. 878-89.
29. Polk, D.B. and R.M. Peek, Jr., *Helicobacter pylori: gastric cancer and beyond.* Nat Rev Cancer. 10(6): p. 403-14.
30. Varghese, J.S. and D.F. Easton, *Genome-wide association studies in common cancers--what have we learnt?* Curr Opin Genet Dev. 20(3): p. 201-9.
31. Fletcher, O. and R.S. Houlston, *Architecture of inherited susceptibility to common cancer.* Nat Rev Cancer. 10(5): p. 353-61.

32. Grulich, A.E., et al., *Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis.* Lancet, 2007. 370(9581): p. 59-67.
33. Vajdic, C.M. and M.T. van Leeuwen, *Cancer incidence and risk factors after solid organ transplantation.* Int J Cancer, 2009. 125(8): p. 1747-54.
34. Matsuda, T., et al., *Low fidelity DNA synthesis by human DNA polymerase-eta.* Nature, 2000. 404(6781): p. 1011-3.
35. Tao, Y., et al., *Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data.* Proc Natl Acad Sci U S A. 108(29): p. 12042-7.
36. Navin, N., et al., *Tumour evolution inferred by single-cell sequencing.* Nature, 2011. 472(7341): p. 90-4.
37. Campbell, P.J., et al., *The patterns and dynamics of genomic instability in metastatic pancreatic cancer.* Nature. 467(7319): p. 1109-13.
38. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. 144(5): p. 646-74.
39. Zhao, L. and P.K. Vogt, *Helical domain and kinase domain mutations in p110alpha of phosphatidylinositol 3-kinase induce gain of function by different mechanisms.* Proc Natl Acad Sci U S A, 2008. 105(7): p. 2652-7.
40. Wolff, A.C., et al., *American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer.* Arch Pathol Lab Med, 2007. 131(1): p. 18-43.
41. Perez-Tenorio, G., et al., *PIK3CA mutations and PTEN loss correlate with similar prognostic factors and are not mutually exclusive in breast cancer.* Clin Cancer Res, 2007. 13(12): p. 3577-84.
42. Hynes, N.E. and H.A. Lane, *ERBB receptors and cancer: the complexity of targeted inhibitors.* Nat Rev Cancer, 2005. 5(5): p. 341-54.
43. Isakoff, S.J., et al., *Breast cancer-associated PIK3CA mutations are oncogenic in mammary epithelial cells.* Cancer Res, 2005. 65(23): p. 10992-1000.
44. Zhang, H., et al., *Comprehensive analysis of oncogenic effects of PIK3CA mutations in human mammary epithelial cells.* Breast Cancer Res Treat, 2008. 112(2): p. 217-27.
45. Ali, I.U., L.M. Schriml, and M. Dean, *Mutational spectra of PTEN/MMAC1 gene: a tumor suppressor with lipid phosphatase activity.* J Natl Cancer Inst, 1999. 91(22): p. 1922-32.
46. Li, J., et al., *PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer.* Science, 1997. 275(5308): p. 1943-7.
47. Kang, Y.H., H.S. Lee, and W.H. Kim, *Promoter methylation and silencing of PTEN in gastric carcinoma.* Lab Invest, 2002. 82(3): p. 285-91.
48. Garcia, J.M., et al., *Promoter methylation of the PTEN gene is a common molecular change in breast cancer.* Genes Chromosomes Cancer, 2004. 41(2): p. 117-24.

49. Poliseno, L., et al., *A coding-independent function of gene and pseudogene mRNAs regulates tumour biology.* Nature. 465(7301): p. 1033-8.

50. Zhang, S. and D. Yu, *PI(3)king apart PTEN's role in cancer.* Clin Cancer Res. 16(17): p. 4325-30.

51. Saal, L.H., 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): p. 7564-9.

52. Koninki, K., et al., *Multiple molecular mechanisms underlying trastuzumab and lapatinib resistance in JIMT-1 breast cancer cells.* Cancer Lett, 2010. 294(2): p. 211-9.

53. Miller, T.W., et al., *Mutations in the phosphatidylinositol 3-kinase pathway: role in tumor progression and therapeutic implications in breast cancer.* Breast Cancer Res, 2011. 13(6): p. 224.

54. Borresen-Dale, A.L., *TP53 and breast cancer.* Hum Mutat, 2003. 21(3): p. 292-300.

55. Miller, L.D., et al., *An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival.* Proc Natl Acad Sci U S A, 2005. 102(38): p. 13550-5.

56. O'Donovan, P.J. and D.M. Livingston, *BRCA1 and BRCA2: breast/ovarian cancer susceptibility gene products and participants in DNA double-strand break repair.* Carcinogenesis, 2010. 31(6): p. 961-7.

57. Knudson, A.G., Jr., *Mutation and cancer: statistical study of retinoblastoma.* Proc Natl Acad Sci U S A, 1971. 68(4): p. 820-3.

58. Knudson, A.G., *Two genetic hits (more or less) to cancer.* Nat Rev Cancer, 2001. 1(2): p. 157-62.

59. Trotman, L.C., et al., *Pten dose dictates cancer progression in the prostate.* PLoS Biol, 2003. 1(3): p. E59.

60. MacDonald, D.J., *Germline mutations in cancer susceptibility genes: an overview for nurses.* Semin Oncol Nurs, 2011. 27(1): p. 21-33.

61. Sherr, C.J., *Principles of tumor suppression.* Cell, 2004. 116(2): p. 235-46.

62. Mulligan, L.M., et al., *Germ-line mutations of the RET proto-oncogene in multiple endocrine neoplasia type 2A.* Nature, 1993. 363(6428): p. 458-60.

63. Luo, J., N.L. Solimini, and S.J. Elledge, *Principles of cancer therapy: oncogene and non-oncogene addiction.* Cell, 2009. 136(5): p. 823-37.

64. Bryant, H.E., et al., *Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase.* Nature, 2005. 434(7035): p. 913-7.

65. Dai, C., et al., *Heat shock factor 1 is a powerful multifaceted modifier of carcinogenesis.* Cell, 2007. 130(6): p. 1005-18.

66. O'Shaughnessy, J., et al., *Iniparib plus chemotherapy in metastatic triple-negative breast cancer.* N Engl J Med. 364(3): p. 205-14.

67. Albertson, D.G., et al., *Chromosome aberrations in solid tumors.* Nat Genet, 2003. 34(4): p. 369-76.

68. Lengauer, C., K.W. Kinzler, and B. Vogelstein, *Genetic instability in colorectal cancers.* Nature, 1997. 386(6625): p. 623-7.

69. Lingle, W.L., et al., *Centrosome amplification drives chromosomal instability in breast tumor development.* Proc Natl Acad Sci U S A, 2002. 99(4): p. 1978-83.
70. Carter, S.L., et al., *A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers.* Nat Genet, 2006. 38(9): p. 1043-8.
71. Birkbak, N.J., et al., *Paradoxical relationship between chromosomal instability and survival outcome in cancer.* Cancer Res, 2011. 71(10): p. 3447-52.
72. Schvartzman, J.M., R. Sotillo, and R. Benezra, *Mitotic chromosomal instability and cancer: mouse modelling of the human disease.* Nat Rev Cancer. 10(2): p. 102-15.
73. Holland, A.J. and D.W. Cleveland, *Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis.* Nat Rev Mol Cell Biol, 2009. 10(7): p. 478-87.
74. Torres, E.M., et al., *Identification of aneuploidy-tolerating mutations.* Cell. 143(1): p. 71-83.
75. Lee, A.J., et al., *Chromosomal instability confers intrinsic multidrug resistance.* Cancer Res, 2011. 71(5): p. 1858-70.
76. Babu, J.R., et al., *Rae1 is an essential mitotic checkpoint regulator that cooperates with Bub3 to prevent chromosome missegregation.* J Cell Biol, 2003. 160(3): p. 341-53.
77. Kalitsis, P., et al., *Increased chromosome instability but not cancer predisposition in haploinsufficient Bub3 mice.* Genes Chromosomes Cancer, 2005. 44(1): p. 29-36.
78. Lee, W., et al., *The mutation spectrum revealed by paired genome sequences from a lung cancer patient.* Nature, 2010. 465(7297): p. 473-7.
79. Jackson, S.P. and J. Bartek, *The DNA-damage response in human biology and disease.* Nature, 2009. 461(7267): p. 1071-8.
80. De Bont, R. and N. van Larebeke, *Endogenous DNA damage in humans: a review of quantitative data.* Mutagenesis, 2004. 19(3): p. 169-85.
81. Gorgoulis, V.G., et al., *Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions.* Nature, 2005. 434(7035): p. 907-13.
82. Evert, B.A., et al., *Spontaneous DNA damage in Saccharomyces cerevisiae elicits phenotypic properties similar to cancer cells.* J Biol Chem, 2004. 279(21): p. 22585-94.
83. Arana, M.E. and T.A. Kunkel, *Mutator phenotypes due to DNA replication infidelity.* Semin Cancer Biol, 2010. 20(5): p. 304-11.
84. Santarius, T., et al., *A census of amplified and overexpressed human cancer genes.* Nat Rev Cancer. 10(1): p. 59-64.
85. Bignell, G.R., et al., *Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution.* Genome Res, 2007. 17(9): p. 1296-303.
86. Pollack, J.R., et al., *Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors.* Proc Natl Acad Sci U S A, 2002. 99(20): p. 12963-8.

87. Hyman, E., et al., *Impact of DNA amplification on gene expression patterns in breast cancer.* Cancer Res, 2002. 62(21): p. 6240-5.

88. Ormandy, C.J., et al., *Cyclin D1, EMS1 and 11q13 amplification in breast cancer.* Breast Cancer Res Treat, 2003. 78(3): p. 323-35.

89. Vainio, P., et al., *Integrative genomic, transcriptomic, and RNAi analysis indicates a potential oncogenic role for FAM110B in castration-resistant prostate cancer.* Prostate, 2011.

90. Murphy, D.M., et al., *Dissection of the oncogenic MYCN transcriptional network reveals a large set of clinically relevant cell cycle genes as drivers of neuroblastoma tumorigenesis.* Mol Carcinog, 2011. 50(6): p. 403-11.

91. Iwakawa, R., et al., *MYC amplification as a prognostic marker of early-stage lung adenocarcinoma identified by whole genome copy number analysis.* Clin Cancer Res, 2011. 17(6): p. 1481-9.

92. Yang, Z.Q., et al., *Multiple interacting oncogenes on the 8p11-p12 amplicon in human breast cancer.* Cancer Res, 2006. 66(24): p. 11632-43.

93. Zender, L., et al., *Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach.* Cell, 2006. 125(7): p. 1253-67.

94. Edgren, H., et al., *Identification of fusion genes in breast cancer by paired-end RNA-sequencing.* Genome Biol, 2011. 12(1): p. R6.

95. Russnes, H.G., et al., *Genomic architecture characterizes tumor progression paths and fate in breast cancer patients.* Sci Transl Med, 2010. 2(38): p. 38ra47.

96. Iljin, K., et al., *TMPRSS2 fusions with oncogenic ETS factors in prostate cancer involve unbalanced genomic rearrangements and are associated with HDAC1 and epigenetic reprogramming.* Cancer Res, 2006. 66(21): p. 10242-6.

97. Cox, C., et al., *A survey of homozygous deletions in human cancer genomes.* Proc Natl Acad Sci U S A, 2005. 102(12): p. 4542-7.

98. Bignell, G.R., et al., *Signatures of mutation and selection in the cancer genome.* Nature, 2010. 463(7283): p. 893-8.

99. Mitelman, F., B. Johansson, and F. Mertens, *The impact of translocations and gene fusions on cancer causation.* Nat Rev Cancer, 2007. 7(4): p. 233-45.

100. Mitelman, F. *Mitelman Database of Chromosome Aberrations in Cancer*. Available from: http://cgap.nci.nih.gov/Chromosomes/Mitelman.

101. Klein, I.A., et al., *Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes.* Cell, 2011. 147(1): p. 95-106.

102. Chiarle, R., et al., *Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells.* Cell, 2011. 147(1): p. 107-19.

103. Janssen, A., et al., *Chromosome segregation errors as a cause of DNA damage and structural chromosome aberrations.* Science, 2011. 333(6051): p. 1895-8.

104. Stephens, P.J., et al., *Massive genomic rearrangement acquired in a single catastrophic event during cancer development.* Cell, 2011. 144(1): p. 27-40.
105. Kloosterman, W.P., et al., *Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer.* Genome Biol, 2011. 12(10): p. R103.
106. Stephens, P.J., et al., *Complex landscapes of somatic rearrangement in human breast cancer genomes.* Nature, 2009. 462(7276): p. 1005-10.
107. Makela, T.P., et al., *A fusion protein formed by L-myc and a novel gene in SCLC.* EMBO J, 1991. 10(6): p. 1331-5.
108. Mani, R.S., et al., *Induced chromosomal proximity and gene fusions in prostate cancer.* Science, 2009. 326(5957): p. 1230.
109. Rowley, J.D., *Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining.* Nature, 1973. 243(5405): p. 290-3.
110. de Klein, A., et al., *A cellular oncogene is translocated to the Philadelphia chromosome in chronic myelocytic leukaemia.* Nature, 1982. 300(5894): p. 765-7.
111. Tomlins, S.A., et al., *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.* Science, 2005. 310(5748): p. 644-8.
112. Soda, M., et al., *Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer.* Nature, 2007. 448(7153): p. 561-6.
113. Tao, J., et al., *CD44-SLC1A2 gene fusions in gastric cancer.* Sci Transl Med, 2011. 3(77): p. 77ra30.
114. Asmann, Y.W., et al., *A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines.* Nucleic Acids Res, 2011. 39(15): p. e100.
115. Ha, K., et al., *Identification of gene fusion transcripts by transcriptome sequencing in BRCA1-mutated breast cancers and cell lines.* BMC Med Genomics, 2011. 4: p. 75.
116. Inaki, K., et al., *Transcriptional consequences of genomic structural aberrations in breast cancer.* Genome Res, 2011. 21(5): p. 676-87.
117. Lae, M., et al., *Secretory breast carcinomas with ETV6-NTRK3 fusion gene belong to the basal-like carcinoma spectrum.* Mod Pathol, 2009. 22(2): p. 291-8.
118. Persson, M., et al., *Recurrent fusion of MYB and NFIB transcription factor genes in carcinomas of the breast and head and neck.* Proc Natl Acad Sci U S A, 2009. 106(44): p. 18740-4.
119. Robinson, D.R., et al., *Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer.* Nat Med, 2011. 17(12): p. 1646-51.
120. Greenman, C., et al., *Patterns of somatic mutation in human cancer genomes.* Nature, 2007. 446(7132): p. 153-8.
121. Bozic, I., et al., *Accumulation of driver and passenger mutations during tumor progression.* Proc Natl Acad Sci U S A, 2010. 107(43): p. 18545-50.

122. Carter, H., et al., *Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations.* Cancer Res, 2009. 69(16): p. 6660-7.

123. Jones, P.A. and S.B. Baylin, *The fundamental role of epigenetic events in cancer.* Nat Rev Genet, 2002. 3(6): p. 415-28.

124. Hansen, K.D., et al., *Increased methylation variation in epigenetic domains across cancer types.* Nat Genet, 2011. 43(8): p. 768-75.

125. Couronne, L., C. Bastard, and O.A. Bernard, *TET2 and DNMT3A mutations in human T-cell lymphoma.* N Engl J Med, 2012. 366(1): p. 95-6.

126. Ley, T.J., et al., *DNMT3A mutations in acute myeloid leukemia.* N Engl J Med, 2010. 363(25): p. 2424-33.

127. Walter, M.J., et al., *Recurrent DNMT3A mutations in patients with myelodysplastic syndromes.* Leukemia, 2011. 25(7): p. 1153-8.

128. Kallioniemi, A., et al., *Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.* Science, 1992. 258(5083): p. 818-21.

129. Pinkel, D., et al., *High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays.* Nat Genet, 1998. 20(2): p. 207-11.

130. Pollack, J.R., et al., *Genome-wide analysis of DNA copy-number changes using cDNA microarrays.* Nat Genet, 1999. 23(1): p. 41-6.

131. Kallioniemi, A., *CGH microarrays and cancer.* Curr Opin Biotechnol, 2008. 19(1): p. 36-40.

132. Wicker, N., et al., *A new look towards BAC-based array CGH through a comprehensive comparison with oligo-based array CGH.* BMC Genomics, 2007. 8: p. 84.

133. Wang, X.S., et al., *An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer.* Nat Biotechnol, 2009. 27(11): p. 1005-11.

134. Siggberg, L., et al., *Array CGH in molecular diagnosis of mental retardation - A study of 150 Finnish patients.* Am J Med Genet A, 2010. 152A(6): p. 1398-410.

135. Fiorentino, F., et al., *Introducing array comparative genomic hybridization into routine prenatal diagnosis practice: a prospective study on over 1000 consecutive clinical cases.* Prenat Diagn, 2011. 31(13): p. 1270-82.

136. Sebat, J., et al., *Large-scale copy number polymorphism in the human genome.* Science, 2004. 305(5683): p. 525-8.

137. Redon, R., et al., *Global variation in copy number in the human genome.* Nature, 2006. 444(7118): p. 444-54.

138. Gazave, E., et al., *Copy number variation analysis in the great apes reveals species-specific patterns of structural variation.* Genome Res, 2011. 21(10): p. 1626-39.

139. Nicholas, T.J., et al., *A high-resolution integrated map of copy number polymorphisms within and between breeds of the modern domesticated dog.* BMC Genomics, 2011. 12: p. 414.

140. Mermel, C.H., et al., *GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.* Genome Biol, 2011. 12(4): p. R41.

141. Shah, S.P., *Computational methods for identification of recurrent copy number alteration patterns by array CGH.* Cytogenet Genome Res, 2008. 123(1-4): p. 343-51.
142. Autio, R., et al., *CGH-Plotter: MATLAB toolbox for CGH-data analysis.* Bioinformatics, 2003. 19(13): p. 1714-5.
143. Olshen, A.B., et al., *Circular binary segmentation for the analysis of array-based DNA copy number data.* Biostatistics, 2004. 5(4): p. 557-72.
144. Baross, A., et al., *Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data.* BMC Bioinformatics, 2007. 8: p. 368.
145. Muggerud, A.A., et al., *Data integration from two microarray platforms identifies bi-allelic genetic inactivation of RIC8A in a breast cancer cell line.* BMC Med Genomics, 2009. 2: p. 26.
146. Sulonen, A., et al., *work in progress.*
147. Lonigro, R.J., et al., *Detection of somatic copy number alterations in cancer using targeted exome capture sequencing.* Neoplasia, 2011. 13(11): p. 1019-25.
148. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. 270(5235): p. 467-70.
149. Fodor, S.P., et al., *Light-directed, spatially addressable parallel chemical synthesis.* Science, 1991. 251(4995): p. 767-73.
150. Pease, A.C., et al., *Light-generated oligonucleotide arrays for rapid DNA sequence analysis.* Proc Natl Acad Sci U S A, 1994. 91(11): p. 5022-6.
151. Krijgsman, O., et al., *A diagnostic gene profile for molecular subtyping of breast cancer associated with treatment response.* Breast Cancer Res Treat, 2011.
152. van de Vijver, M.J., et al., *A gene-expression signature as a predictor of survival in breast cancer.* N Engl J Med, 2002. 347(25): p. 1999-2009.
153. Paul, A.L., et al., *Parabolic Flight Induces Changes in Gene Expression Patterns in Arabidopsis thaliana.* Astrobiology, 2011.
154. Saetre, P., et al., *From wild wolf to domestic dog: gene expression changes in the brain.* Brain Res Mol Brain Res, 2004. 126(2): p. 198-206.
155. Allison, D.B., et al., *Microarray data analysis: from disarray to consolidation and consensus.* Nat Rev Genet, 2006. 7(1): p. 55-65.
156. Lehmussola, A., P. Ruusuvuori, and O. Yli-Harja, *Evaluating the performance of microarray segmentation algorithms.* Bioinformatics, 2006. 22(23): p. 2910-7.
157. Quackenbush, J., *Microarray data normalization and transformation.* Nat Genet, 2002. 32 Suppl: p. 496-501.
158. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Res, 2003. 31(4): p. e15.
159. Dai, M., et al., *Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.* Nucleic Acids Res, 2005. 33(20): p. e175.

160. Flicek, P., et al., *Ensembl 2011.* Nucleic Acids Res, 2011. 39(Database issue): p. D800-6.
161. Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, 2000. 403(6769): p. 503-11.
162. Mootha, V.K., et al., *PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.* Nat Genet, 2003. 34(3): p. 267-73.
163. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.
164. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets--10 years on.* Nucleic Acids Res, 2011. 39(Database issue): p. D1005-10.
165. Parkinson, H., et al., *ArrayExpress--a public database of microarray experiments and gene expression profiles.* Nucleic Acids Res, 2007. 35(Database issue): p. D747-50.
166. Kilpinen, S., et al., *Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues.* Genome Biol, 2008. 9(9): p. R139.
167. Rhodes, D.R., et al., *ONCOMINE: a cancer microarray database and integrated data-mining platform.* Neoplasia, 2004. 6(1): p. 1-6.
168. Kilpinen, S., K. Ojala, and O. Kallioniemi, *Analysis of kinase gene expression patterns across 5681 human tissue samples reveals functional genomic taxonomy of the kinome.* PLoS One, 2010. 5(12): p. e15068.
169. Ojala, K.A., S.K. Kilpinen, and O.P. Kallioniemi, *Classification of unknown primary tumors with a data-driven method based on a large microarray reference database.* Genome Med, 2011. 3(9): p. 63.
170. Silva, A.L. and L. Romao, *The mammalian nonsense-mediated mRNA decay pathway: to decay or not to decay! Which players make the decision?* FEBS Lett, 2009. 583(3): p. 499-505.
171. Wittmann, J., E.M. Hol, and H.M. Jack, *hUPF2 silencing identifies physiologic substrates of mammalian nonsense-mediated mRNA decay.* Mol Cell Biol, 2006. 26(4): p. 1272-87.
172. Noensie, E.N. and H.C. Dietz, *A strategy for disease gene identification through nonsense-mediated mRNA decay inhibition.* Nat Biotechnol, 2001. 19(5): p. 434-9.
173. Huusko, P., et al., *Nonsense-mediated decay microarray analysis identifies mutations of EPHB2 in human prostate cancer.* Nat Genet, 2004. 36(9): p. 979-83.
174. Buffart, T.E., et al., *NMD inhibition fails to identify tumour suppressor genes in microsatellite stable gastric cancer cell lines.* BMC Med Genomics, 2009. 2: p. 39.
175. Mamo, A., et al., *An integrated genomic approach identifies ARID1A as a candidate tumor-suppressor gene in breast cancer.* Oncogene, 2011.
176. Ivanov, I., et al., *Identifying candidate colon cancer tumor suppressor genes using inhibition of nonsense-mediated mRNA decay in colon cancer cells.* Oncogene, 2007. 26(20): p. 2873-84.

177. Rossi, M.R., et al., *Identification of inactivating mutations in the JAK1, SYNJ2, and CLPTM1 genes in prostate cancer cells using inhibition of nonsense-mediated decay and microarray analysis.* Cancer Genet Cytogenet, 2005. 161(2): p. 97-103.
178. Pinyol, M., et al., *Inactivation of RB1 in mantle-cell lymphoma detected by nonsense-mediated mRNA decay pathway inhibition and microarray analysis.* Blood, 2007. 109(12): p. 5422-9.
179. Bloethner, S., et al., *Identification of ARHGEF17, DENND2D, FGFR3, and RB1 mutations in melanoma by inhibition of nonsense-mediated mRNA decay.* Genes Chromosomes Cancer, 2008. 47(12): p. 1076-85.
180. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. 11(1): p. 31-46.
181. Rothberg, J.M., et al., *An integrated semiconductor device enabling non-optical genome sequencing.* Nature, 2011. 475(7356): p. 348-52.
182. Sultan, M., et al., *A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.* Science, 2008. 321(5891): p. 956-60.
183. Nagalakshmi, U., et al., *The transcriptional landscape of the yeast genome defined by RNA sequencing.* Science, 2008. 320(5881): p. 1344-9.
184. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. 5(7): p. 621-8.
185. Lynch, V.J., et al., *Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals.* Nat Genet, 2011.
186. Iyer, M.K., A.M. Chinnaiyan, and C.A. Maher, *ChimeraScan: a tool for identifying chimeric transcription in sequencing data.* Bioinformatics, 2011. 27(20): p. 2903-4.
187. Kim, D. and S.L. Salzberg, *TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.* Genome Biol, 2011. 12(8): p. R72.
188. McPherson, A., et al., *deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data.* PLoS Comput Biol, 2011. 7(5): p. e1001138.
189. Sboner, A., et al., *FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data.* Genome Biol, 2010. 11(10): p. R104.
190. Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.* Proc Natl Acad Sci U S A, 2001. 98(1): p. 31-6.
191. Team, R.D.C., *R: A Language and Environment for Statistical Computing.* 2011.
192. Liu, G., et al., *NetAffx: Affymetrix probesets and annotations.* Nucleic Acids Res, 2003. 31(1): p. 82-6.
193. Sircoulomb, F., et al., *Genome profiling of ERBB2-amplified breast cancers.* BMC Cancer, 2010. 10: p. 539.
194. Baumbusch, L.O., et al., *Comparison of the Agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors.* BMC Genomics, 2008. 9: p. 379.

195. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. 10(3): p. R25.
196. Krzywinski, M., et al., *Circos: an information aesthetic for comparative genomics.* Genome Res, 2009. 19(9): p. 1639-45.
197. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biol, 2004. 5(10): p. R80.
198. Durinck, S., et al., *GenomeGraphs: integrated genomic data visualization with R.* BMC Bioinformatics, 2009. 10: p. 2.
199. Leivonen, S.K., et al., *Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines.* Oncogene, 2009. 28(44): p. 3926-36.
200. Naume, B., et al., *Presence of bone marrow micrometastasis is associated with different recurrence risk within molecular subtypes of breast cancer.* Mol Oncol, 2007. 1(2): p. 160-71.
201. Wiedswang, G., et al., *Detection of isolated tumor cells in bone marrow is an independent prognostic factor in breast cancer.* J Clin Oncol, 2003. 21(18): p. 3469-78.
202. Tost, J. and I.G. Gut, *DNA methylation analysis by pyrosequencing.* Nat Protoc, 2007. 2(9): p. 2265-75.
203. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. 12(4): p. 656-64.
204. Chin, S.F., et al., *High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer.* Genome Biol, 2007. 8(10): p. R215.
205. Barlund, M., et al., *Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer.* Genes Chromosomes Cancer, 2002. 35(4): p. 311-7.
206. Hampton, O.A., et al., *A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome.* Genome Res, 2009. 19(2): p. 167-77.
207. Ruan, Y., et al., *Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs).* Genome Res, 2007. 17(6): p. 828-38.
208. Murphy, W.J., K.P. Watkins, and N. Agabian, *Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: evidence for trans splicing.* Cell, 1986. 47(4): p. 517-25.
209. Li, H., et al., *A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells.* Science, 2008. 321(5894): p. 1357-61.
210. Berns, K., et al., *A functional genetic approach identifies the PI3K pathway as a major determinant of trastuzumab resistance in breast cancer.* Cancer Cell, 2007. 12(4): p. 395-402.
211. Tanner, M., et al., *Characterization of a novel cell line established from a patient with Herceptin-resistant breast cancer.* Mol Cancer Ther, 2004. 3(12): p. 1585-92.
212. Nencioni, A., et al., *Grb7 upregulation is a molecular adaptation to HER2 signaling inhibition due to removal of Akt-mediated gene repression.* PLoS One, 2010. 5(2): p. e9024.

213. Venkatraman, E.S. and A.B. Olshen, *A faster circular binary segmentation algorithm for the analysis of array CGH data.* Bioinformatics, 2007. 23(6): p. 657-63.
214. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. 437(7057): p. 376-80.
215. Shendure, J., et al., *Accurate multiplex polony sequencing of an evolved bacterial genome.* Science, 2005. 309(5741): p. 1728-32.
216. Ley, T.J., et al., *DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.* Nature, 2008. 456(7218): p. 66-72.
217. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. 25(14): p. 1754-60.
218. Forbes, S.A., et al., *COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.* Nucleic Acids Res, 2011. 39(Database issue): p. D945-50.
219. Durinck, S., et al., *Temporal dissection of tumorigenesis in primary cancers.* Cancer Discov, 2011. 1(2): p. 137-143.
220. Woodard, G.E., et al., *Ric-8A and Gi alpha recruit LGN, NuMA, and dynein to the cell cortex to help orient the mitotic spindle.* Mol Cell Biol, 2010. 30(14): p. 3519-30.
221. Wang, L., et al., *Resistance to inhibitors of cholinesterase-8A (Ric-8A) is critical for growth factor receptor-induced actin cytoskeletal reorganization.* J Biol Chem, 2011. 286(35): p. 31055-61.
222. Jones, S., et al., *Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.* Science, 2008. 321(5897): p. 1801-6.
223. Parsons, D.W., et al., *An integrated genomic analysis of human glioblastoma multiforme.* Science, 2008. 321(5897): p. 1807-12.
224. Dalgliesh, G.L., et al., *Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes.* Nature, 2010. 463(7279): p. 360-3.
225. Kumar, A., et al., *Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers.* Proc Natl Acad Sci U S A, 2011. 108(41): p. 17087-92.
226. Taub, R., et al., *Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human Burkitt lymphoma and murine plasmacytoma cells.* Proc Natl Acad Sci U S A, 1982. 79(24): p. 7837-41.
227. Peterson, S.M., et al., *Human Sulfatase 2 inhibits in vivo tumor growth of MDA-MB-231 human breast cancer xenografts.* BMC Cancer, 2010. 10: p. 427.
228. Lai, J.P., et al., *Sulfatase 2 protects hepatocellular carcinoma cells against apoptosis induced by the PI3K inhibitor LY294002 and ERK and JNK kinase inhibitors.* Liver Int, 2010. 30(10): p. 1522-8.
229. Pleasance, E.D., et al., *A small-cell lung cancer genome with complex signatures of tobacco exposure.* Nature, 2010. 463(7278): p. 184-90.
230. Staaf, J., et al., *High-resolution genomic and expression analyses of copy number alterations in HER2-amplified breast cancer.* Breast Cancer Res, 2010. 12(3): p. R25.

231. Beroukhim, R., et al., *Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma.* Proc Natl Acad Sci U S A, 2007. 104(50): p. 20007-12.

232. Ioannidis, J.P., et al., *Repeatability of published microarray gene expression analyses.* Nat Genet, 2009. 41(2): p. 149-55.

233. Baggerly, K.A. and K.R. Coombes, *What Information Should Be Required to Support Clinical "Omics" Publications?* Clin Chem, 2011.

234. Leisch, F., *Sweave: Dynamic generation of statistical reports using literate data analysis.* Compstat 2002: Proceedings in Computational Statistics, 2002: p. 575-580.

235. Mesirov, J.P., *Computer science. Accessible reproducible research.* Science, 2010. 327(5964): p. 415-6.

236. Goecks, J., A. Nekrutenko, and J. Taylor, *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.* Genome Biol, 2010. 11(8): p. R86.

237. Mankell, H., *In Africa, the Art of Listening*, in *New York Times*2011.