Search for fusion genes in breast cancers which have amplified regions

Hjörleifur Einarsson

Faculty of life and environmental sciences
University of Iceland
2014

Search for fusion genes in breast cancers which have amplified regions

Hjörleifur Einarsson

14 ECTS thesis submitted in partial fulfillment of
Baccalaureus Scientiarum degree in molecular biology

Supervisor: Inga Reynisdóttir
Supervisory teacher: Zophanías Oddur Jónsson

Faculty of life end environmental sciences
School of engineering and natural sciences
University of Iceland
Reykjavík, May 2014

Search for fusion genes in breast cancers which have amplified regions
14 ECTS thesis submitted in partial fulfillment of Baccalaureus *Scientiarum* degree in molecular biology

Faculty of life and environmental sciences
School of engineering and natural sciences
University of Iceland
Sturlugata 7
107 Reykjavík

Sími: 525 4000

Registration information:
Hjörleifur Einarsson, 2014, *Search for fusion genes in breast cancers which have amplified regions*, Faculty of life and environmental sciences, University of Iceland, 47p.

Printing: Háskólaprent
Reykjavík, May 2014

Útdráttur

Brjóstakrabbamein er algengasta krabbamein í konum og eru erfðabrenglanir algengar í æxlunum. Magnanir á litningasvæðum og litningayfirfærslur eru dæmi um slíkar brenglanir sem geta leitt til æxlismyndunar vegna yfirtjáningar gena eða myndunar samrunagena. Markmið þessa verkefnis er að leita að samrunagenum í brjóstakrabbameinum með áherslu á þekkt mögnunarsvæði. Sérstök áhersla er á 8p11-12 svæðið en það er magnað i 10-15% brjóstaæxla og aðeins eitt áhrifagen æxlismyndunar hefur verið skilgreint á þessu svæði. Til að finna möguleg samrunagen keyrðum við háhraðaraðgreiningargögn fyrir valin brjóstakrabbameinsæxli og brjóstakrabbameins frumulínur í gegnum SOAPfuse forritið sem finnur (leitar eftir) samrunagenum. Af þessum mögulegu samrunagenum voru nokkur valin til frekari greiningar með Sanger raðgreiningu til að staðfesta að þau væru til staðar. Með þessum aðferðum tókst okkur að finna og staðfesta fyrirfram þekkt samrunagen í MCF7 brjóstakrabbameinsfrumulínunni. Könnun á greiningu úr SOAPfuse leiddi í ljós samrunagen sem hafa orðið til vegna litningayfirfærslu, viðsnúninga litningabúta og samruna litningasvæða á sama litningi. Sum samrunagenin koma fyrir í mörgum sýnum, meðan önnur eru einstök fyrir hvert sýni. NOTCH2NL samruni var fundinn af SOAPfuse í T-47D brjóstakrabbameinsfrumulínunni og í 5/8 ER+ brjóstakrabbameinum. NOTCH2NL samruninn var staðfestur í T-47D með Sanger raðgreiningu. NOTCH genafjölskyldan hefur áður verið fundin í samrunagenum sem hafa áhrif á æxlismyndun í brjóstakrabbameinum. Unnið er að því að staðfesta fleiri áhugaverð samrunagen og fylgt verður eftir völdum staðfestum samrunagenum í íslenskum brjóstaæxlum.

Abstract

Breast cancer is the most common cancer in women and genomic aberrations are common in the tumors. Amplifications of chromosomes and chromosomal translocations are examples of such aberrations which can affect tumor development by upregulating gene expression and through the formation of fusion genes. The aim of this assignment is to search for fusion genes in breast cancer tumors with emphasis on known amplified regions. Special focus is on the 8p11-12 region which is amplified in 10-15% of all breast cancers and only one gene in the region has been linked to tumor development. To predict possible fusion genes we ran paired-end RNA sequencing data for chosen breast cancer cell lines and breast cancer tumors through SOAPfuse, an algorithm that detects fusion genes. Chosen fusion gene predictions were further analyzed using Sanger sequencing for validation. With these methods we were able to find and verify fusion genes that were previously known in the MCF7 breast cancer cell line. SOAPfuse analysis for the samples revealed fusion genes which were formed due to translocations, chromosomal inversions and deletions. Some of these fusions were recurrent while others were special for each sample. A NOTCH2NL fusion was detected by SOAPfuse in the T-47D breast cancer cell line and in 5/8 ER+ breast cancer tumors. We verified the fusion in T-47D using Sanger sequencing. NOTCH family genes have previously been found in fusion genes which affect tumor development in breast cancers. Other interesting fusion genes await verification and chosen verified fusion genes will be further studied in Icelandic breast cancer tumors.

Table of contents

Figures

Tables

Acknowledgments

I would like to express my gratitude to my supervisor Inga Reynisdóttir for giving me the opportunity to work on this project, always being ready to assist me and for answering all my questions.

I would like to thank Eydís Þórunn Guðmundsdóttir, Edda Sigríður Freysteinsdóttir and Guðrún Jóhannesdóttir at the Department of Pathology, Landspitali University Hospital for instructing me and helping me with the lab work,

I also would like to thank Bylgja Hilmarsdóttir in the laboratory of Þórarinn Guðjónsson in Lækngarður, University of Iceland for providing the cell lines and Daníel Óskarsson in the laboratory of Arnar Pálsson in Askja, University of Iceland for helping me with running SOAPfuse.

1 Introduction

1.1 Breast cancer epidemiology

Breast cancer is the most common type of cancer diagnosed in females worldwide. They are 23% of total diagnosed cancers and the cause of 14% of cancer related deaths in 2008. Incidence rates are higher in the developed countries where North America and Europe alone account for almost 60% of the total cases worldwide (Jemal et al. 2011). While the incidence rate is higher in the developed countries the mortality rate is higher in the economically developing countries. The reason for increased incidence rate but lower mortality rate in the developed countries is due to the presence of screening programs which can detect early invasive cancers (Parkin et al. 2002). Breast cancers do occur in males but they are only 1% of the total cancer cases and therefore very rare compared to females (Jemal et al. 2011). In Iceland 30% of diagnosed cancers in women are breast cancers. Since 1987, when organized screening was implemented, the number of women diagnosed with breast cancer in Iceland has increased but at the same time there has been a great improvement in survival rates (www.krabb.is).

1.2 Breast cancer pathobiology

Breast cancer classification is based on various aspects of the tumor including histopathology, receptor status, grade of the tumor, and gene expression profile (Viale 2012). Most breast cancers originate in the breast ducts and account for around 80% of all breast tumors. Around 10-15% of tumors develop inside the lobes and other subtypes exist and account for around 10% of breast tumors (Banin Hirata et al. 2014). The receptor status classification is based on the hormone receptors that the tumor expresses. The best studied hormone receptors in breast cancers are the estrogen receptor (ER), the progesterone receptor (PR) and the human epidermal growth factor receptor 2 (HER-2) (Almeida & Barry 2011). These receptors, if expressed, can have significant effect on the prognoses and are also important to determine the right treatment since the tumors respond

differently to treatments based on which hormone receptors they express (Banin Hirata et al. 2014). Based on micro array data of the gene expression, tumors are commonly classified into 5 subgroups luminal A, luminal B, basal-like tumors, ERBB2 over expressing tumors and normal breast-like tumors. Luminal A tumors have the best prognosis while basal-like tumors, luminal B and ERBB2 over expressing tumors have a rather poor prognosis (Hu et al. 2006).

The most commonly used strategy to diagnose early stage breast cancer is through regular screening. Most common screening methods are clinical breast examination and mammography (Almeida & Barry 2011). It is recommended that women past 40 undergo such screening every 2-3 years but if there is an underlying hereditary risk factor, screening should start earlier and be more frequent (Fakkert et al. 2011). Magnetic resonance imaging (MRI) has also been used for screening in younger patients with underlying hereditary risk factors because it is more sensitive than the mammography and it can better detect tumors in young dense breast tissue (Kriege & Brekelmans 2004; Morrow 2004). MRI is not used as a standard screening method because it is expensive, there is higher risk of false positives and it could increase unnecessary follow up examinations (Houssami N 2009). If possible tumors are detected by screening they are usually easy to confirm with microscopic analysis of a sample from the affected area of the breast (www.cancer.net).

The different types of breast tumors have effect on the possible therapeutic options after diagnosis. Surgery is commonly used to remove the tumor; either the whole breast is removed (mastectomy) or only the affected part of the breast (lumpectomy) (Almeida & Barry 2011; Veronesi & Cascinelli 2002). The surgical removal is usually combined with chemotherapy and/or radiation therapy which target all proliferating cells (Group 1995). Recent technological advances in high-throughput genomics, transcriptomics and proteomics have enabled us to more effectively find molecular factors that have effect on clinical outcome and drug response and thus given rise to personalized medicine (Sang-Hoon Cho, Jongsu Jeon 2012). Trastuzumab is a Her2 monoclonal antibody that inhibits cell growth by inhibiting Her2 from activating its intracellular tyrosine kinase (Goldenberg 1999). It is therefore commonly used as a treatment for tumors which have been shown to overexpress the Her2 receptor and has been shown to have a significant survival benefit (Goldhirsch et al. 2013). There are several drugs specialized for tumors overexpressing

ER and PR (Sang-Hoon Cho, Jongsu Jeon 2012). Tamoxifen is used a lot as treatment against tumors that have been characterized as ER+ by binding to the ER and preventing the increased cell growth effects from estrogen (Jordan 1993). Tamoxifen and aromatase inhibitors have also proven to be useful to prevent the formation of tumors in high risk individuals (Cuzick et al. 2003; Kalidas & Brown 2005).

1.3 Breast cancer cell biology

Since our cells have many check points to detect and respond to flaws in the cell cycle, no one mutation is enough to cause cancer. According to Hanahan´s and Weinberg´s Hallmarks of Cancer: The Next generation there are 8 requirements that the cell needs to fulfill to ultimately be able to form a malignant tumor (Figure 1, Hanahan & Weinberg, 2011).

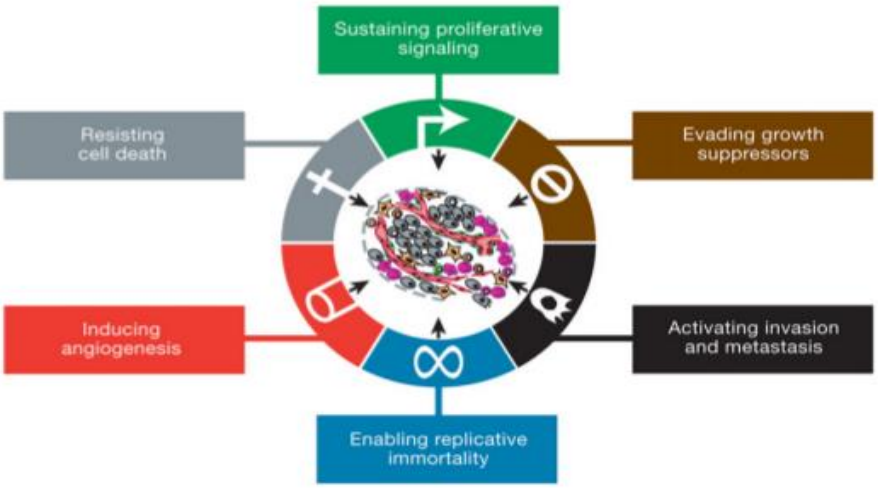

Figure 1. The six hallmarks of cancer proposed by Hanahan and Weinberg. *These six hallmarks were first proposed by Hanahan and Weinberg in 2000 as the requirements for cell to become a cancer cell. In 2011, Avoiding immune destruction and deregulating cellular energetics were added as emerging hallmarks* (Hanahan & Weinberg 2011).

1. Sustaining proliferative signaling: The growth of normal cells is held in check by carefully regulating growth promoting signals. Cancer cells need to somehow deregulate this system to be able to grow as they please. They can do this for example by over producing growth factor ligands or increasing the number of receptor proteins making

them hyper responsive.

2. Evading growth suppressors: There are really important factors in the cell that negatively regulate growth. TP53 is an example of a tumor-suppressor which is often defective in cancer cells which enables them to get past these negative growth regulations.

3. Resisting cell death: Apoptosis is one of the most important defenses against cells that are behaving irregularly. So cell proliferation is not enough for the cancer cells to grow in number, they also have to be able to get past programmed cell death, for example by inactivating the Beclin-1 gene.

4. Inducing Angiogenesis: For a tumor to grow they need to be able to access and transport nutrients and oxygen. They can facilitate the growth of new blood vessels through production of vascular endothelial growth factors.

5. Enabling replicative immortality: Most normal cells in the human body can only go through limited cell divisions and therefore they are in no way immortal. Cancer cells need to immortalize themselves and do so often by activating their Telomerase.

6. Activating invasion and metastasis: Normal cells are regularly kept in place by cell-cell adhesion. For them to move freely they have to free themselves from this cell-cell adhesion, by for example deregulating E-cadherin.

7. Reprogramming energy metabolism: The out of control proliferation and cell growth of tumors demands that the cells provide a lot of fuel under unusual circumstances.
To meet this demand the cancer cells need to alter they metabolic pathways.

8. Evading immune destruction: Our cells are constantly under the surveillance of the immune system which, under normal circumstances, recognize and eliminate developing cancer cells. So cancer cells have to be able to hide from this system or react when the system attacks them.

1.4 Gene alterations in breast cancer

Breast cancers, like other cancers, occur due to unregulated cell growth because of series of alterations in oncogenes, tumor suppressor genes, epigenetic factors and environmental factors. These mutations can be constitutional, inherited from parent, or somatic. Oncogenes affect cell growth and differentiation and are in normal cells either turned off or expressed at lower levels than they are in tumors where they have been altered. MYC, CCND1 and ERBB2 are examples of oncogenes that have been shown to play an important

role in early development of sporadic breast cancer when mutated. These early mutations will have dramatic effect on the development of the tumor and therefore different mutations lead to clinically different tumor types (Kenemans et al. 2004). Tumor suppressor genes are usually associated with regulation of the cell cycle or promotion of apoptosis in normal cells. Mutations in these genes often increase the risk of further mutations and through those mutations it increases the risk of cancer. It is estimated that around 5-10% of breast cancers are due to strong inherited components which are mostly tumor suppressor genes. The best known of those genes are BRCA1 and BRCA2 which both take part in DNA damage response (Lalloo & Evans 2012). It has been shown that environmental factors can also increase the risk of breast cancer development due to obesity, diet, alcohol intake and various chemicals (Jemal et al. 2011). Epigenetic factors, such as DNA methylation and histone modification, can affect gene expression by up-regulating them, down-regulating them or even turning them on/off without changing the gene primary nucleotide sequence (Wright & Saul 2013; Nowsheen et al. 2014). If these modifications of epigenetic factors affect certain pathways in the cell it can increase the risk of tumor development. These modifications are usually turning on oncogenes through acetylation or turning off tumor suppressor genes trough methylation in breast cancers (Nowsheen et al. 2014). Histone deacetylase (HDAC) usually causes genes to be down-regulated or turned off by removing acetyl groups from the histones. Flaws in HDAC have been connected with tumor development and there have even been successful treatments of breast cancer with drugs that inhibit the HDAC (Connolly & Stearns 2012; Thomas & Munster 2009).

1.5 Genomic aberrations in breast cancer

Genomic aberrations are very frequent in breast cancer and have been shown to have significant effect on the formation of tumors and the progression of the disease. These aberrations can be due to chromosomal instability, breakage and amplification. Chromosomal instability can cause unusual numbers of chromosomes through duplication or loss of whole chromosomes. Breakage of a chromosome can cause a loss or translocation of genetic material. Amplification occurs when multiple copies of the same region are formed (Chin et al. 2006). The amplification of 17q12 is well studied and is known to cause over expression of the ERBB2 gene. The amplification of 8p11-12 has

been found in 10-15% of breast tumors (Gelsi-Boyer et al. 2005) and because of the complex pattern of the amplification it is likely that there are more than one cancer susceptibility genes in the region (Reynisdottir et al. 2013). These amplified regions mainly affect the formation and progression of tumors by causing genes to be overexpressed or through the formation of fusion genes.

1.6 Fusion genes

Chromosomal aberrations can give rise to a new gene product which is a hybrid gene from two previously separated genes. These genes are called fusion genes and occur mainly due to chromosomal translocations, inversions and deletions.

The first chromosomal aberration which gave rise to a fusion gene was found with the discovery of the Philadelphia chromosome in leukemia. This translocation gave rise to the fusion of the BCR locus on chromosome 22 to the ABL tyrosine kinase on chromosome 9 (BCR-ABL). The fused protein product was shown to induce leukemia when expressed in bone marrow cells in mice (Daley et al. 1990). Later on there was a success in producing a drug that was a specific inhibitor of the BCR-ABL fusion (Fausel 2007). This discovery led to the search and discovery of the many other fusion genes known today.

Even though it has been known for many years that fusion genes play an important role in leukemia and sarcomas they were not until recently found in common carcinomas. The reason for this lack of understanding of fusion genes in common carcinoma is because they were thought to be not as important as point mutations and deletions and because of technical reasons it was difficult to detect these fusion genes. In 2005, when Tomlins et al. found that the TMPRSS2-ERG fusion gene was expressed in over 50% of all prostate cancers more focus was put into identifying fusion genes in common carcinomas (Tomlins et al. 2005).

Fusion genes have been discovered to be present in breast cancer cell lines and tumors but most of them have not been shown to be recurrent. The MDA-MB-175 cell line has a fusion of ODZ4-NRG1 that has been known since 1999 (Liu et al. 1999). Recently, there were found four expressed fusion genes in MCF7 with the use of high-throughput

sequencing studies (Hampton et al. 2009) and in 24 breast cancer cell lines and tumors there were found 21 fusion genes but none of them were recurrent (Campbell et al. 2008). Recently however a study found two classes of recurrent gene fusions using paired-end transcriptome sequencing to map gene fusions in breast cancer cell lines and tumors. These recurrent gene fusions included MAST and NOTCH family genes. Both of these recurrent gene fusions seemed to have phenotypic effects in breast epithelial cells (Robinson et al. 2011). This study shows that rare recurrent gene fusions occur in breast cancers and might be used as a target for personalized medicine.

These findings among others support that fusion genes in breast cancer cannot be ignored as rare and unimportant events. It seems that these fusion genes are not like the common recurrent fusion genes found in leukemia so they have to be approached in a different way. It is likely that fusions in breast cancer involve rarer fusion which influence genes in the same pathway and fusions where one important gene is recurrent with different fusion partners which all have the same effect on the important gene.

1.7 Paired end RNA sequencing

The Sanger method was the dominating method for sequencing for almost two decades and is referred to as the first generation of sequencing. In the past few years there has been a rapid shift from this classic sequencing method towards next generation sequencing (NGS) methods. All NGS methods today are based on shot gun sequencing where the DNA is sheered into small fragments which are then sequenced using one of the many platforms available today (Metzker 2010). Before NGS transcriptomics studies largely relied on hybridization-based microarray technologies which only gave a limited ability to understand the many complex factors of the transcriptome. RNA sequencing using NGS has enabled researchers to get a more complete view of the transcriptome for multiple organisms and has for example increased the understanding of transcription initiation, alternative splicing and improved detection of fusion genes (Ozsolak & Milos 2011).

The limiting factor to RNA sequencing with NGS is the short read length. The solution to this limitation is the widely used paired-end sequencing (Figure 2). Instead of producing small reads which are completely sequenced paired-end sequencing produces larger reads

and only the two ends of the read are sequenced. These read ends are unique enough so that they can be aligned against a reference genome and thus the whole read can be correctly placed. The ability of the paired-end RNA sequencing to reveal alignment between the two ends of a DNA fragment and the transcriptome allows us to detect fusion genes which could not be detected with the classic shotgun NGS sequencing technique (Fullwood et al. 2009; Martin & Wang 2011).

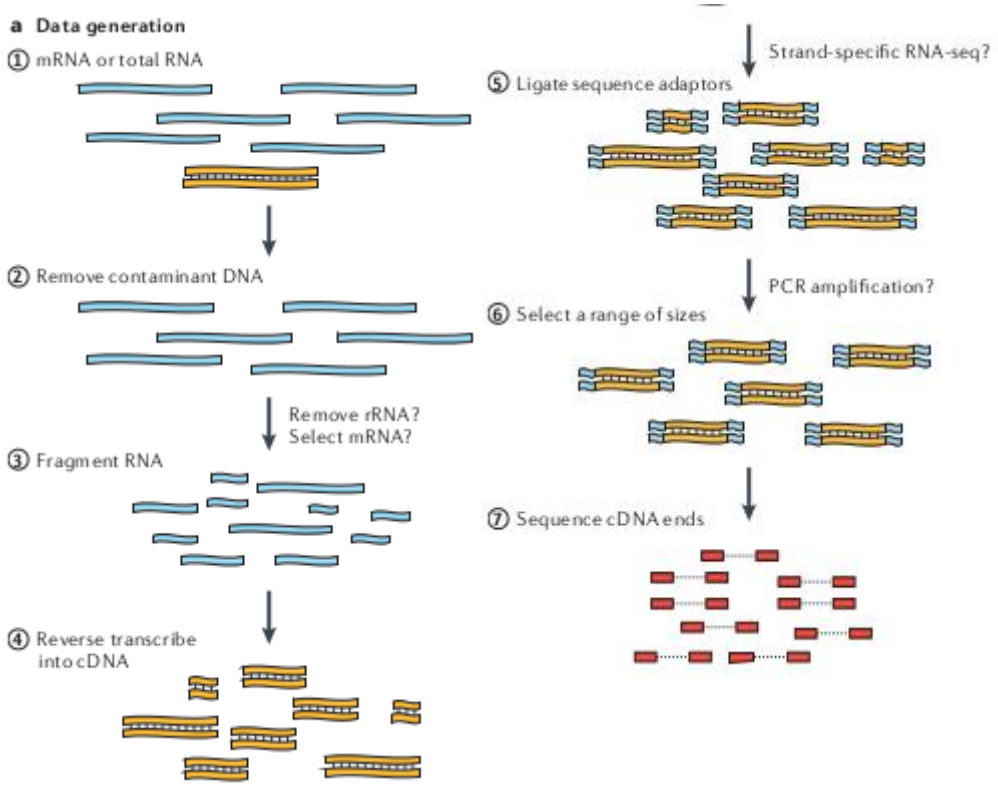

Figure 2. A schematic picture describing the process of paired-end next generation sequencing. *In short, all contamination is first removed from the RNA extraction which can include RNA which is not of interest (1-2). The RNA is then fragmented and reversed transcribed into cDNA (3-4). Sequencing adaptors are ligated onto the RNA fragment and specific range of fragment size chosen (5-6). Finally the cDNA ends are sequenced using NGS technologies to produce many short reads (7). The steps marked with question marks are optional but are crucial for some sequencing studies.* (Martin & Wang 2011).

For the study described herein paired-end sequencing data for chosen breast cancer cell lines and breast cancer tumors were obtained from the SRA database at NCBI (www.ncbi.nlm.nih.gov/sra).

2 Aims

The aim of this project is to search for fusion genes in various breast cancer cell lines and tumors, which have amplified regions, using high-throughput sequencing data obtained from the NCBI SRA database. There will be special emphasis on possible fusion genes at the 8p11-12 amplified regions due to its complex amplification pattern, which increases the likelihood of more than one oncogene residing there.

3 Methods and materials

3.1 Cell lines

Fusion gene predictions were done with SOAPfuse for 12 different cell lines. Seven of those cell lines had either amplification or loss at the 8p11-12 chromosome region; T-47D, MCF7, MDA-MB-134, ZR-75-1, CAMA-1, SUM-52, SUM-225 and MDA-MB-231. In this study there were only done further analysis on predicted fusions in MCF7 (ATCC HTB-22) and T-47D (ATCC HTB-133). T-47D is a luminal breast cancer cell line which has amplification in its 8p11-12 region. It has a loss at 34-35Mb at 8p12 and inversion at 35.4-38.15 Mb at 8p11-12. The inversion seems to be connected to a region on chromosome 14 (Kytölä et al. 2000; Pole et al. 2006). MCF7 is a luminal a breast cancer cell line which has a translocation from chromosome 8 to chromosome 14 as well as having a loss in the 8p11-12 chromosomal region (Kytölä et al. 2000). MCF10A is a cell line from a normal breast tissue that has been immortalized and fusion gene prediction results for MCF10A can be used for comparison of normal tissue vs tumor (Debnath et al. 2003). Other cancer cell lines which were analyzed with SOAPfuse but did not have any special characteristics at the 8p11-12 region were MDA-MB-231, SUM-229, SKBR3 and BT474.

3.2 RNA extraction

RNA was extracted from cell lines using RNeasy kit from Qiagen following the protocol. The medium was poured out of the cell culture tube and the cells washed twice with PBS. 600µl of RLT buffer was added to cells and pipetted to mix. When the cells had been lysed the extract was pipetted into a 1.5ml Eppendorf glass and 600µl of 70% ethanol added. The mixture was vortexed and then 700µl were loaded onto an RNeasy column. The column was placed in a 2ml collection tube, centrifuged for 15sec at 10.000rpm and the flow-through discarded. The rest was then loaded onto the column and the last step repeated. Then 700µl RW1 buffer was added to the column, centrifuged for 15sec at 10000rpm and flow-through discarded. Then 500µl RPE buffer was added onto column, centrifuged for 2min at 10000rpm. The column was then put into a clean collection tube and centrifuged for 1min at max speed. Next the column was put into a clean 1.5ml Eppendorf glass, 30µl RNase-free water loaded onto column and the column centrifuged for 1min at 10000rpm. The last step was then repeated with a clean Eppendorf glass to collect the rest of the RNA. RNA quantity was measured for each Eppendorf glass using Nanodrop ND-1000.

3.3 cDNA synthesis

cDNA was synthesized using a reverse transcription (RT) kit from Thermo Scientific. The RNA extraction was used to synthesis cDNA with the final concentration of around 100ng/µl. 11µl(192ng/µl) RNA extract and 1µl random primers were mixed and put into a PCR reaction to anneal the primers to the RNA.

PCR annealing reaction

- 65°C 1min
- 22°C 10min
- 4°C infinite

While the annealing reaction takes place the master mix for the reverse transcription reaction was mixed. 4µl 5xRT reaction buffer, 1µl RNase out, 2µl dNTP´s and 1µl RT enzyme were mixed for each RT reaction. 8µl of the master mix were added into each one of the products from the PCR annealing reaction.

PCR RT reaction

- 25°C 5min
- 42°C 60min
- 70°C 5min
- 4°C infinite

The synthesized cDNA was stored in a freezer at -20°C between uses.

3.4 Identifying possible fusion transcripts with SOAPfuse

There are several different programs designed to predict fusion genes; SOAPfuse, Tophat-Fusion, FusionHunter, deFuse, Chimerascan, SnowShoes-FTD and more (Asmann et al. 2011; Jia et al. 2013; Kim & Salzberg 2011; Li et al. 2011; Maher et al. 2009; McPherson et al. 2011). Most of these programs are built on the same principles to find fusion genes based on paired-end RNA sequencing data by aligning them to a reference genome. The pipeline for each program can vary, which effects how sensitive the program is and how much memory and CPU time it consumes. SOAPfuse was mainly picked because of its high sensitivity while not needing too much memory or CPU time compared to other programs (Jia et al. 2013). I also chose SOAPfuse because it is accessible for free and is built for Linux operating systems which I have a lot of experience with. To run SOAPfuse we used Linux clusters with 32 GB memory located in VR-III at the University of Iceland. Páll Melsted is the supervisor of these Linux clusters.

SOAPfuse is developed by Jia et al. as a method to detect fusion transcripts based on paired-end RNA sequencing data. The algorithm searches for two types of reads to identify fusion genes. Span-reads are paired-end reads that map to two different genes and junction-reads (junc-reads) are reads that map over the exact junction site of the two genes, giving a single base resolution of the junction site (Figure 3).

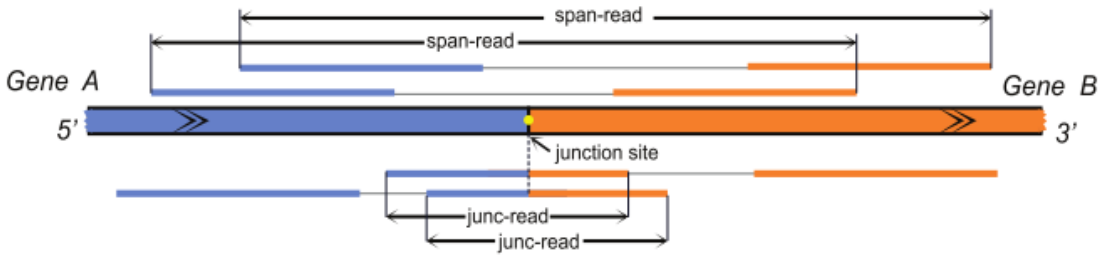

Figure 3. Model of gene fusion supported by span-reads and junc-reads. *Span-reads are paired-end RNA- sequences which ends align to different genes in the fusion gene. Junc-reads are paired-end RNA sequences where one end of the read aligns to the junction site of the fusion gene. SOAPfuse uses these two types of reads to detect fusion genes based on paired-end RNA sequences.* (Jia et al. 2013).

The program contains nine steps in its pipeline which can be split into four main parts that can be seen in an overview in Figure 4:

1. Read alignment (step S01 to S03): In this part SOAPfuse uses SOAP2 and BWA to align the paired-end reads to the human reference genome and annotated transcripts. At the end of this part the program has divided the paired-end reads into concordant reads, discordant reads and unmapped reads. Discordant reads are paired-end reads where only one end mapped against the reference genome or paired-end reads indicating an unusual insert size or mapped orientation.

2. Identifying candidate gene pairs (S04 and S05): Here the program searches from all discordant reads for span-reads that support candidate gene pairs. Here the program also applies some filters to ensure accurate prediction. The program excludes gene pairs from the same gene family as well as those with overlapped or homogenous exon regions.

3. Detection of predicted fusion (S06 and S07): In this part the program first identifies which gene of the candidate gene pairs is upstream and which is downstream based on the results from the paired-end alignment in part 1. Then filtered unmapped reads (FUM) are aligned against the candidate gene pairs, using SOAP2, to detect likely junc-reads. The aligned FUM along with span-reads that support the candidate fusion gene pairs are combined to detect the junction sites.

4. Filtering fusions (S08 and S09): Here the program filters out predicted fusions that are not supported by sufficient amount of span-reads, junc-reads and other criteria. In the end the program returns a list of high-confident fusions, predicted junction reads and SVG figures showing the alignment of reads that support the fusion transcript.

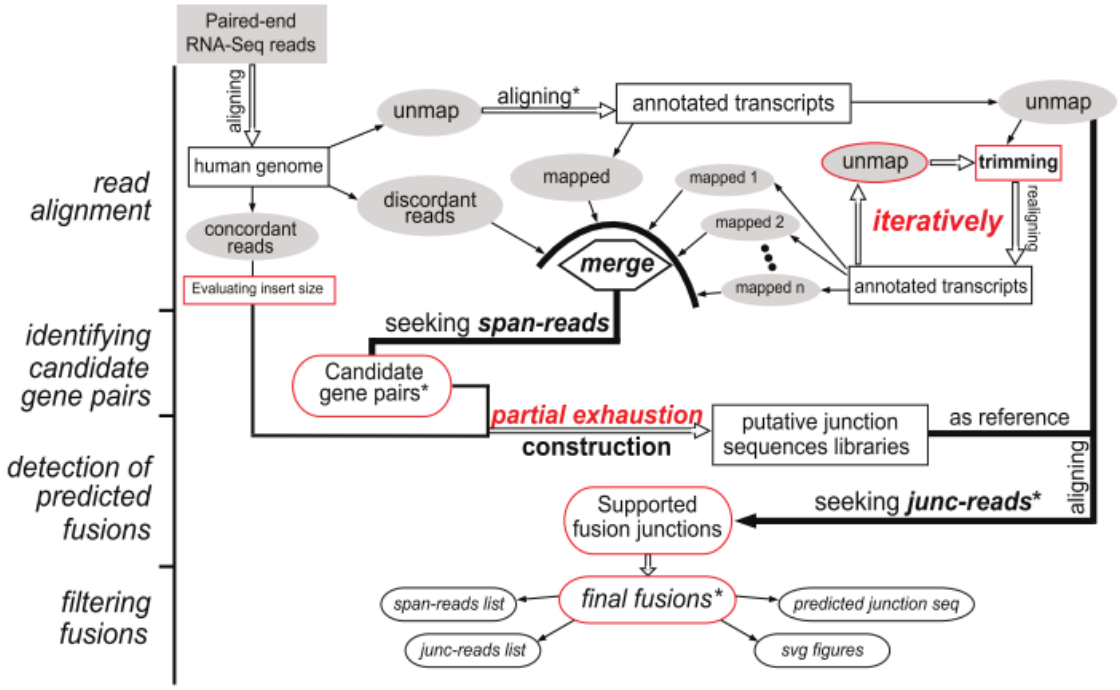

Figure 4. The four parts of SOAPfuse algorithm. *The algorithm is split into four main parts on the left site of the picture. Steps indicated in red play a central role in the algorithm and steps marked by an asterisk indicate key filtering steps* (Jia et al. 2013).

3.5 Validation of fusion genes

3.5.1 Designing primers

Primers are designed so that the junction site between the two genes gets amplified in the PCR reaction. To get a good sequencing of the junction site the primers are located ~90-100bp away from the fused region (Figure 5).

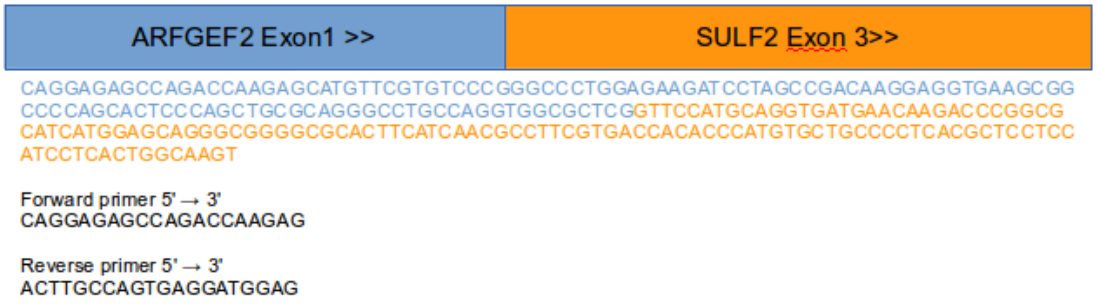

Figure 5. An example of a primer pair designed to detect fusion gene. *This primer pair was designed to validate the ARFGEF2-SULF2 fusion in the MCF7 breast cancer cell line. The fusion sequence was predicted by SOAPfuse and the primers are designed to amplify the fused product so that it can be sequenced with Sanger sequencing.*

3.5.2 PCR amplification

PCR amplification was optimized for each fusion gene primer pair to find the best PCR buffer, elongation reaction duration, primer annealing temperature and if betaine was needed in the master mix or not. The best PCR amplification for each fusion gene was then used to sequence the fusion junction using Sanger sequencing. Master mix for each PCR reaction is mixed for each reaction with the right buffer and with or without betaine depending on what mixture gives the best amplification for the fusion junction (Table 1).

Table 1. Master mix recipes for 1x PCR amplification reaction. *The mixes can either have the KCl or $(NH_4)_2SO_4$ buffers and they sometimes include betaine to ensure better binding between the cDNA and the primers.*

Reagent	1x [µl]	1x w/betaine [µl]
H_2O	6.6	4.6
Betaine	0	2
Buffer	1	1
$MgCl_2$	0.8	0.8
dNTPs	0.64	0.64
Forward primer (20pmol/µl)	0.2	0.2
Reverse primer (20pmol/µl)	0.2	0.2
cDNA (100ng/µl)	0.5	0.5
Taq polymerase	0.06	0.06

PCR amplification reaction

- 94°C 3min
- 35xCycles:
  - 94°C 30sek
  - 55/58/62°C 1min
  - 72°C 45sek
- 72°C 10min
- 4°C Infinite

The success and the quality of the PCR amplification was estimated by the quality of the band obtained when the PCR reaction products were run on an agarose gel. The agarose gel was used to optimize the PCR amplification procedure for each fusion gene and to check the quality of the PCR amplification before the Sanger sequencing.

3.5.3 Running agarose gel

2% agarose gel was prepared by mixing 60ml 1xTBE buffer and 1.2g agarose. Then the mix was heated in a microwave at max power for 70sec after which 1.8µl ethidium bromide (EtBr) were added into the mix. The mixture was given time to cool down before being poured into a cast. Combs were placed in the cast to create wells for loading samples and the gel allowed to set. Loading samples were mixed by taking 4µl PCR reaction samples and add 1µl 5xloading dye. The gel electrophoresis unit was filled with 1xTBE running buffer and the gel placed in the middle of it. All the loading samples, 1kb ladder and 50bp ladders were loaded into individual wells and the gel run at 100V until the dye line was approximately 75-80% of the way down the gel. Pictures were taken of the gels with Chemi XRS Documentation System from Bio-Rad by using UV-light. Under the UV-light the EtBr fluoresce, when intercalated with DNA. The pictures were used to estimate PCR amplification product quality and for publications.

3.5.4 Sanger sequencing

The most used Sanger method today is based on chain terminating dideoxynucleotides (ddNTPs). The ddNTPS cause the DNA replication to terminate which results in the production of fragments of different length. These fragments are passed through a narrow tube which contains a gel-like matrix which separates the fragments based on size. Each one of the four ddNTPs is labeled with different fluorescent dyes which emit light at

different wavelengths. By using the fluorescent label the ddNTPs that terminated the replication for each strand can be detected and its position can be determined based on the strands position in the gel matrix (Sanger & Coulson 1975; Smith et al. 1986).

First the PCR product was cleaned using exonucleases to get rid of all the unwanted DNA and primers. A PCR clean up prior to sequencing kit was used from Thermo Scientific. The 1x master mix for the cleaning reaction was made by mixing 4.25µl H_2O, 0.25µl ExoI and 0.5µl FastAP. PCR strips with 5µl exonuclease cleaning master mix in each well were prepared for each sample and then 2µl of PCR product was placed into each well. The PCR strip was then placed in a PCR machine to run the exonuclease cleaning program.

Exonuclease cleaning program

- 37°C 15min
- 85°C 15min

During the exonuclease cleaning program the master mix for the sequencing reaction was prepared. The sequencing reaction was done with a BigDye kit from applied Biosystems. There are two sequencing reactions for every sample, one with the forward PCR primer and another with the reverse PCR primer. The 1x master mix for the sequencing reaction was made by mixing 2.5µl H_2O, 1µl 5xSeq.buffer and 0.5µl BigDye. PCR strips with 4µl sequencing reaction master mix in each well were prepared and then 1µl of cleaned PCR product and 0.1µl(20pmol/µl) added into each well. The PCR strips were then placed in a PCR machine to run the PCR sequencing reaction.

PCR sequencing reaction

- x35 cycles:
  - 96°C 10sec
  - 50°C 5sec
  - 60°C 4min

After the PCR sequencing reaction 3.5µl CleanSeq(MCLAB) and 20µl 70% Ethanol were added into every sample from the PCR sequencing reaction. The solution was mixed by pipetting and then the PCR strip was placed in a magnetic plate for 3-5min. The ethanol mixture was then pipetted out of the strips and discarded. Another 50µl 70% ethanol were added to the strips and pipetted for cleaning before discarding the ethanol. The PCR strips were then taken off the magnetic plate, 50µl 1xElution buffer added into each sample.

After 3-5min wait the PCR strips are placed back on the magnetic plate for 3-5min. The solution was then pipetted into a sequencing plate.

The sequencing plate was placed into the sequencer (3130xl Genetic Analyzer, Applied Biosystems), the run given a proper name, the sequencer set on sequencing analysis, plate scheme filled out, instrument protocol set on Sequencing_BD1, analys,protoc set on POP7_BDv1.1_KB_36cm and then the run was executed. After the run, the data were analyzed using Seq.Analysis and the analyzed data saved onto a USB drive. The results of the sequencing analysis were then read using the Sequencher program (Gene Codes Corporation).

4 Results

4.1 Validation of methods

To test if our methods and the SOAPfuse algorithm were working as we hoped for we used the MCF7 breast cancer cell line which has three previously well characterized fusion genes (Edgren et al. 2011; Hampton et al. 2009). Two paired-end RNA sequencing datasets for the MCF7, from two different studies (Daemen et al. 2013; Edgren et al. 2011), were run though the SOAPfuse algorithm. SOAPfuse found all three fusion genes from both the datasets, see table 2.

Table 2. The result from SOAPfuse for the three previously known fusion genes in the MCF7 cell line. *All the previously validated fusions were found in both datasets. The table shows the number of span- and junc-reads, found by SOAPfuse, which support the fusion gene. Other possible fusion genes were found by SOAPfuse in the MCF7 cell line but they will not be listed here.*

Sample	5´gene	5´Chromosome	3´gene	3´chromosome	Span-reads	Junc-reads
SRR064286	BCAS4	Chr20	BCAS3	Chr17	121	128
SRR925723	BCAS4	Chr20	BCAS3	Chr17	450	270
SRR064286	ARFGEF2	Chr20	SULF2	Chr20	11	16
SRR925723	ARFGEF2	Chr20	SULF2	Chr20	176	119
SRR064286	RPS6KB1	Chr17	VMP1	Chr17	3	8
SRR925723	RPS6KB1	Chr17	VMP1	Chr17	45	58

Two of the three previously known fusion genes that SOAPfuse found in the MCF7 breast cancer cell line were amplified with PCR for validation with Sanger sequencing. The primers for RPS6KB1-VMP1 were designed based on the predicted junction site while the primers for the ARFGEF2-SULF2 were copied from another study (Hampton et al. 2009).

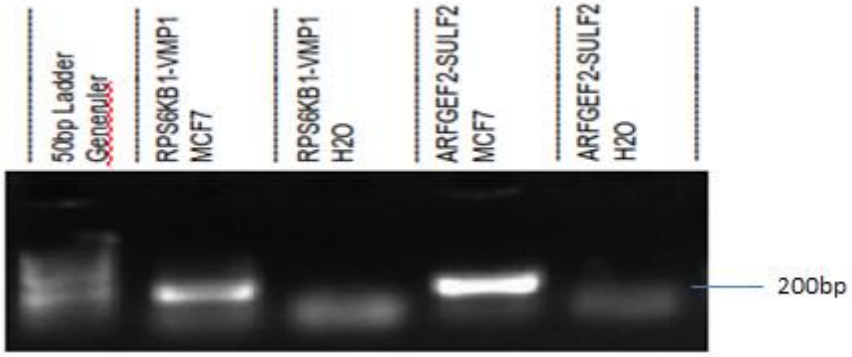

Figure 6. Amplified fusion genes in the MCF7 breast cancer cell line. *Picture of the PCR amplification products run on 2% agarose gel at 100V for 20min. The two previously known fusions were amplified in the MCF7 cancer cell line and H2O was used as a negative control for the amplification.*

The amplification was a success in the MCF7 cell line but there were some bands also visible in the negative controls which were smaller than expected band size, see Figure 6. These PCR products were sequenced using Sanger sequencing to confirm that these were the predicted fusion genes. The negative controls were also sequenced to make sure they were not amplifying the fusion genes as well.

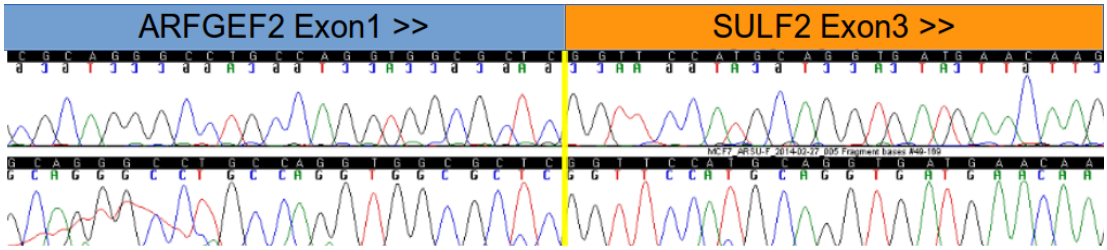

Figure 7. The sequencing results for the ARFGEF2-SULF2 fusion in the MCF7 breast cancer cell line*. The junction site is indicated with a yellow line. The fusion sequence predicted by SOAPfuse is the sequence with white letters on black background. The upper sequence is sequenced using the reverse primer while the lower one is sequenced using the forward primer.*

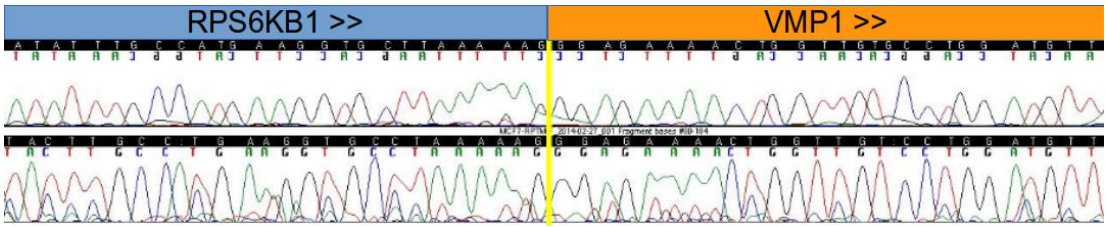

Figure 8. The sequencing results for the RPS6KB1-VMP1 fusion in the MCF7 breast cancer cell line. *The junction site is indicated with a yellow line. The fusion sequence predicted by SOAPfuse is the sequence with white letters on black background. The upper sequence is sequenced using the reverse primer while the lower one is sequenced using the forward primer.*

As seen in figures 7 and 8 both the fusion sequences predicted by SOAPfuse were confirmed in the MCF7 breast cancer cell line. Sequencing of the negative controls showed that nothing had been amplified in those samples. These results show that SOAPfuse can be used to successfully predict fusion genes based on paired-end RNA sequences from breast cancers and that based on these predictions primers can be designed to confirm the fusion genes using Sanger sequencing.

4.2 Scanning for possible fusion genes at the 8p11-12 amplified region in the T-47D cell line

At the start of the assignment there were issues running SOAPfuse due to lack of computer power. Therefore, before results were obtained from SOAPfuse attempts were made to scan for possible fusion genes at the 8p11-12 amplified region in the T-47D cell line using PCR amplification and Sanger sequencing. Two fusions between genes at 8p12-p11 have been detected (Wu et al. 2013) and thus primers were designed against one of the fusions and two potential fusion genes to test whether fusion genes could be detected in T-47D, see table appendix B. T-47D was used because it is known to have amplification in this region and that region was of special interest in this assignment due to its complex amplification pattern (Pole et al. 2006).

The scanning was made with forward primers for GPR-124, Erlin2 and PROSC paired against two different reverse primers for FGFR1, FGFR1-Rev1 and FGFR1-Rev2. FGFR1 was specifically targeted because it has previously been linked to tumor development and because a fusion between Erlin2 and FGFR1 has already been reported in breast cancer (Chin et al. 2006; Wu et al. 2013).

After extensive testing of the primers at different annealing temperatures, with/without betaine and at different elongation step duration there were only found bands that could possibly indicate gene fusion in T-47D for two of the primer pairs. To confirm that the primers were working as intended ERLIN2-F was paired against ERLIN2-R in the MCF7, see figure 9. We also previously amplified FGFR1 using FGFR1 paired with either FGFR1-Rev1 or FGFR1-Rev2 which both resulted in expected band size but were not sequenced. There was a strong band for GPR124-F paired with FGFR1-Rev1 and weak band for ERLIN2-F paired with FGFR1-Rev2 (Figure 9). Some primers showed no or weak bands and some were not specific enough to produce a clear band to allow use for further analyzes. The small bands which are in most wells are most likely primer dimers but they could also indicate contamination, see figure 9.

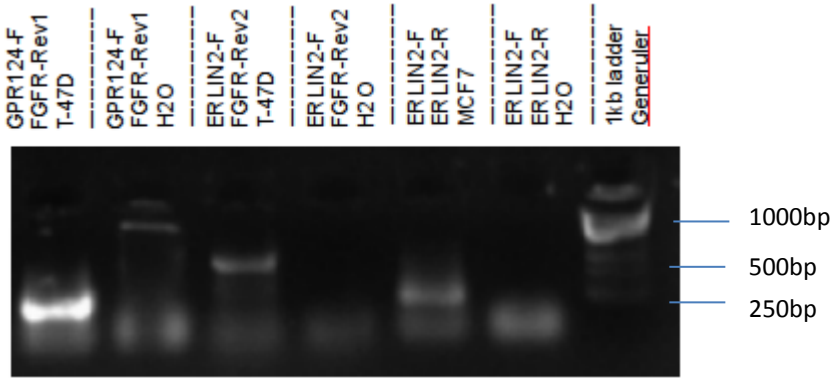

Figure 9. Bands from PCR amplification using primers for scanning for fusion genes in T-47D. *The PCR product was run on a 1.5% agarose gel at 100V for 25min. Erlin2 was amplified in MCF7 as a positive control and PCR reaction with water used as negative control.*

The PCR products which were produced with the scanning of the 8p11-12 region in T-47D were sequenced using Sanger sequencing along with the negative and positive controls.

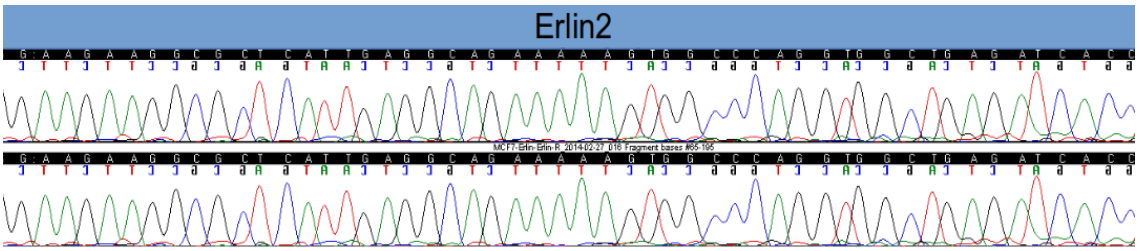

Figure 10. The sequencing results for Erlin2 in the MCF7 breast cancer cell line. *Both the forward and reverse primers gave clear and high quality sequences. The amplification was done with ERLIN2-F and ERLIN2-R, see figure 9.*

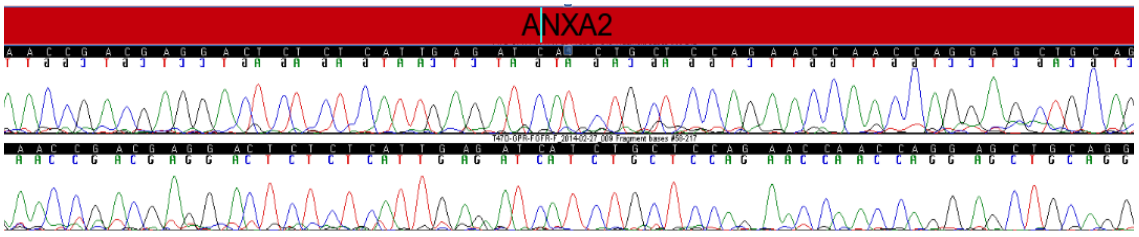

Figure 11. The sequencing results for the PCR product from the GPR124-F and FGFR1-Rev1 primer pair*. Both the forward and reverse primers gave clear and high quality sequences which could not be aligned to either GPR-124 or FGFR1. Blast revealed that the primers had amplified a part of the ANXA2 gene.*

The sequencing of the band produced by the ERLIN2-F and FGFR1-Rev2 primer pair was of low quality due to a lot of background and could not be aligned to any sequences related

to Erlin2 or FGFR1. The band produced by GPR124-F and FGFR1-Rev1 primer pair gave a high quality sequence for both the forward and the reverse primer but it could not be aligned with any parts of GPR-124 or FGFR1. It turned out that the primers had amplified the ANXA2 gene, see figure 11, even though they should not have any binding sits close to the genes according to primer blast (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). The amplification with ERLIN2-F and ERLIN2-R gave bands of expected band size and there was no observable band in the negative control except for primer dimers. The amplification was confirmed to be ERLIN2 with Sanger sequencing, see figure 10. The testing of FGFR1-F paired with either FGFR1-Rev1 or FGFR1-Rev2 in the MCF7 cell line both produced bands of expected size but were not sequenced for confirmation, data not shown. None of the negative controls indicated that any contamination could have been amplified. Sequencing of the negative controls indicated that the bands at the bottom of the gel were most likely due to primer dimers.

Later SOAPfuse was used to predict possible fusion genes in three different paired-end RNA-seq datasets for T-47D from two different studies (Daemen et al. 2013; Gertz et al. 2012). SOAPfuse returned no predicted fusions for the 8p11-12 amplified region in the T-47D. The results from the scanning and SOAPfuse predictions indicate that there is probably no fusion genes located in the 8p11-12 amplified region in the T-47D cell line.

4.3 Validation of predicted fusion genes in the T-47D cell line

SOAPfuse predicted a lot of possible fusion genes in T-47D cell line based on the three datasets used for the cell line, see appendix A. Some of the predicted fusion genes were found in all three datasets while others were specific to one or two of the datasets. Because of the large pool of possible fusion genes only four were chosen for further analyses (see Table 3). The fusion genes were chosen based on how likely they were to affect tumor development, how many datasets SOAPfuse predicted them in, the number of junc-reads and the number of span-reads.

Table 3. The four fusion genes predicted by SOAPfuse in T-47D cancer cell line chosen for further analyses. *The table shows the sample which SOAPfuse predicts the fusion genes to be found in and the number of span- and junc-reads which support the fusion gene. A lot of other fusion genes were predicted by SOAPfuse in the T-47D cell line but were not studied any further and will not be listed here.*

Sample	5´gene	5´Chromosome	3´gene	3´chromosome	Span-reads	Junc-reads
SRR500876	MECOM	Chr3	TTC18	Chr10	13	9
SRR925736	MECOM	Chr3	TTC18	Chr10	1	1
SRR500880	SMG5	Chr1	PAQR6	Chr1	6	14
SRR500876	SMG5	Chr1	PAQR6	Chr1	15	2
SRR925736	SMG5	Chr1	PAQR6	Chr1	9	12
SRR500876	NOTCH2NL	Chr1	NBPF10	Chr1	5	2
SRR500880	NOTCH2NL	Chr1	NBPF10	Chr1	4	6
SRR500880	VGLL4	Chr3	SH3BP5	Chr3	19	4
SRR925736	VGLL4	Chr3	SH3BP5	Chr3	7	8
SRR500876	VGLL4	Chr3	SH3BP5	Chr3	31	5

PCR primers were designed for the four chosen fusion genes, see table Appendix C, based on the fusion sequences predicted by SOAPfuse. After testing different annealing temperature and other factors sharp clear bands were produced with PCR amplification for the MECOM-TTC18 and NOTCH2NL-NBPF10 fusions. Weak and low quality bands were produced for the SMG5-PAQR6 and no band was produced for the VGLL4-SH3BP5 fusion see figure 12.

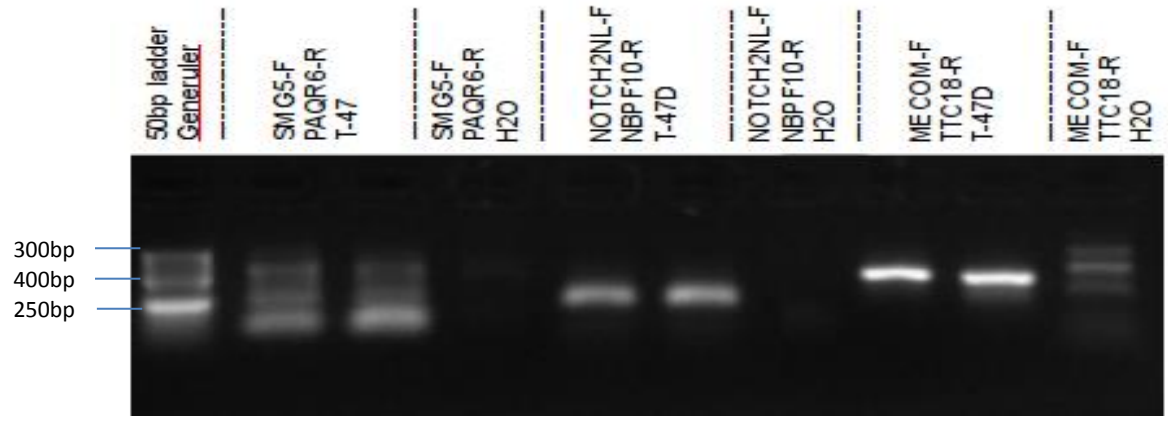

Figure 12. The PCR amplification products from the T-47D cell line which were further analyzed with Sanger sequencing*. The PCR products were run on 2% agarose gel at 100V for 45min. All fusions were amplified in two samples with T-47D and H2O was used as a negative control. Attempts to amplify the VGLL4-SH3BP5 fusion gene did not return any useable amplification products.*

The PCR amplification products were sequenced using Sanger sequencing and aligned to their predicted sequences using sequencher. The sequencing of SMG5-PAQR6 gave only the sequences for the primer pair which was not surprising due to the low quality of the PCR amplification. The sequencing of NOTCH2NL-NBPF10 and MECOM-TTC18 both gave low quality sequences. The low quality of the NOTCH2NL-NBPF10 sequences was surprising because there was no visible contamination in the NOTCH2NL-NBPF10 negative control so the reasons for this low quality are unknown. Sequencher was able to align the results for NOTCH2NL-NBPF10 to the fusion gene sequence predicted by SOAPfuse, see figure 13, but not the MECOM-TTC18 sequence. Sanger sequencing of the negative controls resulted in no sequences. In the negative control for the MECOM fusion there is some contamination which could be the reason for low quality sequences which we cannot align with the predicted fusion.

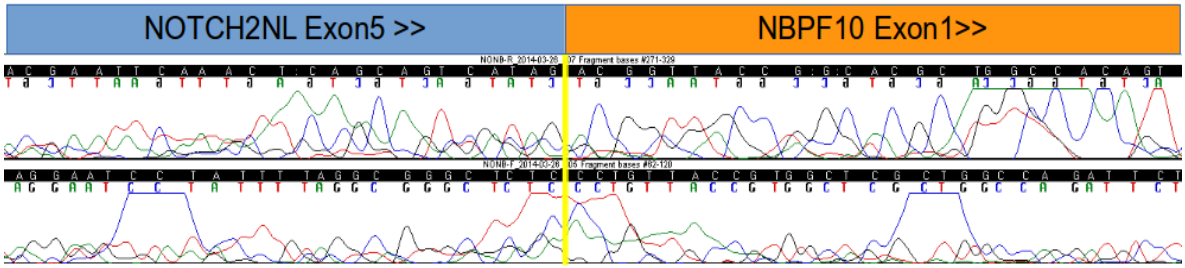

Figure 13. The sequencing results for the NOTCH2NL-NBPF10 fusion gene. *Sequences for both the forward and the reverse primers produced low quality sequences but they were however aligned to the fusion gene sequence predicted by SOAPfuse. The fusion junction is indicated in the figure as a yellow line and it matches to the junction predicted by SOAPfuse (blue and yellow bar above sequences).*

4.4 Other possible fusion genes

There were several other paired-end RNA-seq datasets analyzed with the SOAPfuse algorithm, see appendix A. These datasets included various cancer cell lines, breast cancer tumors and normal tissues, which can be used as a control. SOAPfuse returned several possible fusion genes for all the samples, some of which were recurrent in many different samples while others were limited to single or few samples. The NOTCH2NL-NBPF10 fusion which was sequenced in T-47D was predicted by SOAPfuse to be in 5/8 ER+ tumors but was not found by SOAPfuse in any of the normal breast tissue samples. SOAPfuse also predicted two fusion genes in the MDA-MB-134 breast cancer cell line which included genes that are located at the 8p11-12 region, see table 4.

Table 4. Fusion genes predicted by SOAPfuse in the MDA-MB-134 breast cancer cell line. *Both predicted fusions include a gene located at the 8p11-12 region which is amplified in the cell line. The table also shows the number of span-reads and junc-reads which were found by SOAPfuse to support the fusion gene.*

Sample	5´gene	5´Chromosome	3´gene	3´chromosome	Span-reads	Junc-reads
SRR925724	ANK1	Chr8	ZMAT4	Chr8	12	13
SRR925724	PROSC	Chr8	OR7E11P	Chr11	15	8

The interesting fusions predicted by SOAPfuse in MDA-MB-134 and other samples are not analyzed any further in this assignment but will be analyzed in more detail later.

5 Discussions

The results for the previously known fusion genes in the MCF7 breast cancer cell line show that SOAPfuse can be used to correctly predict possible gene fusions in breast cancer and that the predicted fusion gene sequences can be used to design primers for Sanger sequencing to verify the predicted fusion gene. It also verifies that the setup of the program on the Linux clusters was correct. The results however give no indication of the possible false positives and false negatives which SOAPfuse could return. According to testing done by the developers of SOAPfuse the false positive and false negative rate is lower than 5% (Jia et al. 2013) but it is most likely higher. Many of the genes reported by SOAPfuse were close together and it is likely that some of them were read-throughs but not true fusion genes. However some fusion genes are formed by two closely located genes due to microdeletions and the KANSL1-ARL17A/B fusion found by SOAPfuse is likely an example of such fusion since microdeletions which KANSL1 is affected by have previously been studied (Itsara et al. 2012). Therefore it is important not to discard all fusion genes with two genes which are close to each other.

Because of the large number of predicted fusion genes reported by SOAPfuse for each sample it was necessary to go through the list by hand. First we discarded fusion genes that had low numbers of junc-reads and span-reads because they were not supported well enough by the program. Fusion genes which had two genes that were close to each other and not a high number of junc-reads were also discarded as most likely false positives due to read-throughs. The rest of the predicted fusions were then looked into based on what gene families they belonged to and if they had been previously linked with tumor development. No fusion gene prediction program is 100% accurate so this is not an isolated issue with SOAPfuse.

It was never likely that scanning by amplifying with primers targeting genes in the 8p11-12 region in the T-47D cell line would result in the finding of fusion genes. The primers were designed without any idea of where the fusion junctions might be so even though a fusion gene which included the genes we scanned for was in the T-47D cell line it was unlikely that the right area would be amplified by the primers. One or two of the primers could bind to an exon which was not part of the fusion and even if they did they might be too far

away from each other for amplification. None of the attempted amplifications returned any fusion genes even though the primers had been verified as functional. After a SOAPfuse analyzes of RNA-seq from the T-47D cell line, which predicted no fusion genes in the 8p11-12 region, everything indicated that there were no fusion genes located at the 8p11-12 region in T-47D.

The low quality sequencing results for the MECOM-TTC18 and NOTCH2NL-NBPF10 fusion genes in T-47D is very likely due to contamination in the cDNA of T-47D which causes the background in the sequencing. It is also possible that the primers are not good enough and that could also be true for the SMG5-PAQR6 and VGLL4-SH3BP5, which were used to attempt to verify the fusion genes. None of these predicted fusion genes have therefore been 100% verified but there is a good indication that the NOTCH2NL-NBPF10 is actually present in the T-47D cell line. This NOTCH2NL-NBPF10 fusion gene is very interesting since there have already been fusion genes reported that include genes from the NOTCH gene family which have been shown to have effect on tumor development (Robinson et al. 2011). SOAPfuse also predicted the same fusion in 5/8 ER+ which indicates that the fusion could be recurrent.

There is still a lot of work left in verifying these predicted fusion genes in the T-47D cell line and in verifying other interesting fusions predicted by SOAPfuse in other breast cancer cell lines and tumors. After verifying these fusion genes there is still work left in studying the effects they have, if any, on tumor development and then finally scanning for them in breast tumor samples from Icelandic women to see if they are recurrent.

Bibliography

Almeida, C. & Barry, S., 2011. *Cancer: basic science and clinical aspects*, Available at: http://books.google.com/books?hl=en&lr=&id=j0RV27loexoC&oi=fnd&pg=PT14&dq=Cancer,+Basic+science+and+clinical+aspects&ots=hW41_xeGsd&sig=X_XiV9iNoYVJUz-hOUvtijtfiGw [Accessed May 4, 2014].

Asmann, Y.W. et al., 2011. A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. *Nucleic acids research*, 39(15), p.e100. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3159479&tool=pmcentrez&rendertype=abstract [Accessed April 28, 2014].

Banin Hirata, B.K. et al., 2014. Molecular Markers for Breast Cancer: Prediction on Tumor Behavior. *Disease markers*, 2014, p.513158. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3925609&tool=pmcentrez&rendertype=abstract [Accessed March 6, 2014].

Campbell, P., Stephens, P. & Pleasance, E., 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature …*, 40(6), pp.722–729. Available at: http://www.nature.com/ng/journal/v40/n6/abs/ng.128.html [Accessed April 14, 2014].

Chin, K. et al., 2006. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer cell*, 10(6), pp.529–41. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17157792 [Accessed March 20, 2014].

Connolly, R. & Stearns, V., 2012. Epigenetics as a therapeutic target in breast cancer. *Journal of mammary gland biology and neoplasia*, 17, pp.191–204. Available at: http://link.springer.com/article/10.1007/s10911-012-9263-3 [Accessed May 5, 2014].

Cuzick, J. et al., 2003. Overview of the main outcomes in breast-cancer prevention trials. *Lancet*, 361(9354), pp.296–300. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12559863.

Daemen, A. et al., 2013. Modeling precision treatment of breast cancer. *Genome biology*, 14(10), p.R110. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3937590&tool=pmcentrez&rendertype=abstract [Accessed May 7, 2014].

Daley, G., Van Etten, R. & Baltimore, D., 1990. Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome. *Science*, 247(4944), pp.824–830. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.2406902 [Accessed April 14, 2014].

Debnath, J., Muthuswamy, S.K. & Brugge, J.S., 2003. Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. *Methods*, 30(3), pp.256–268. Available at: http://linkinghub.elsevier.com/retrieve/pii/S104620230300032X [Accessed April 28, 2014].

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

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3091304&tool=pmcentrez&rendertype=abstract [Accessed January 9, 2014].

Fakkert, I.E. et al., 2011. Breast cancer screening in BRCA1 and BRCA2 mutation carriers after risk reducing salpingo-oophorectomy. *Breast cancer research and treatment*, 129(1), pp.157–64. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21373873 [Accessed May 4, 2014].

Fausel, C., 2007. Targeted chronic myeloid leukemia therapy: Seeking a cure. *American journal of health-system pharmacy*, 64(Cml). Available at: http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=10792082&AN=29297851&h=/TzfYb/Fn6IYshYkeZfeIKE+NCH5u4/R0eq3ELmSXqAdLoZyDA049ycGFJrXXT2ZbAGWCqUI+x6MqY+Jg5QV/g==&crl=c [Accessed April 14, 2014].

Fullwood, M.J. et al., 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. , pp.521–532.

Gelsi-Boyer, V. et al., 2005. Comprehensive profiling of 8p11-12 amplification in breast cancer. *Molecular cancer research : MCR*, 3(12), pp.655–67. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16380503 [Accessed April 14, 2014].

Gertz, J. et al., 2012. Genistein and bisphenol A exposure cause estrogen receptor 1 to bind thousands of sites in a cell type-specific manner. *Genome research*, 22(11), pp.2153–62. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3483545&tool=pmcentrez&rendertype=abstract [Accessed January 15, 2014].

Goldenberg, M.M., 1999. Trastuzumab, a recombinant DNA-derived humanized monoclonal antibody, a novel agent for the treatment of metastatic breast cancer. *Clinical therapeutics*, 21(2), pp.309–18. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10211534.

Goldhirsch, A. et al., 2013. 2 years versus 1 year of adjuvant trastuzumab for HER2-positive breast cancer (HERA): an open-label, randomised controlled trial. *Lancet*, 382(9897), pp.1021–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23871490 [Accessed May 3, 2014].

Group, E.B.C.T.C., 1995. Effects of radiotherapy and surgery in early breast cancer; An overview of the randomized trials. *N Engl j med*, pp.1444–1455. Available at: http://ci.nii.ac.jp/naid/30022546829/ [Accessed May 4, 2014].

Hampton, O. a et al., 2009. 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 research*, 19(2), pp.167–77. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2652200&tool=pmcentrez&rendertype=abstract [Accessed January 21, 2014].

Hanahan, D. & Weinberg, R. a, 2011. Hallmarks of cancer: the next generation. *Cell*, 144(5), pp.646–74. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21376230 [Accessed March 19, 2014].

Houssami N, H.D., 2009. Should MRI Be Performed on All Women with Newly Diagnosed, Early Stage Breast Cancer?

Hu, Z. et al., 2006. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC genomics*, 7, p.96. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1468408&tool=pmcentrez&rendertype=abstract [Accessed March 19, 2014].

Itsara, A. et al., 2012. Resolving the breakpoints of the 17q21.31 microdeletion syndrome with next-generation sequencing. *American journal of human genetics*, 90(4), pp.599–613. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3322237&tool=pmcentrez&rendertype=abstract [Accessed May 11, 2014].

Jemal, A., Bray, F. & Center, M., 2011. Global cancer statistics. *CA: a cancer journal …*, 61(2), pp.69–90. Available at: http://onlinelibrary.wiley.com/doi/10.3322/caac.20107/full?dmmsmid=71827&dmmspid=19396336&dmmsuid=1908926. [Accessed January 15, 2014].

Jia, W. et al., 2013. SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome biology*, 14(2), p.R12. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23409703 [Accessed January 15, 2014].

Jordan, V., 1993. A current view of tamoxifen for the treatment and prevention of breast cancer. *British journal of pharmacology*, 110(2), pp.507–517. Available at: http://doi.wiley.com/10.1111/j.1476-5381.1993.tb13840.x [Accessed May 4, 2014].

Kalidas, M. & Brown, P., 2005. Aromatase inhibitors for the treatment and prevention of breast cancer. *Clinical breast cancer*, 6(1), pp.27–37. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2228389&tool=pmcentrez&rendertype=abstract.

Kenemans, P., Verstraeten, R. a & Verheijen, R.H.M., 2004. Oncogenic pathways in hereditary and sporadic breast cancer. *Maturitas*, 49(1), pp.34–43. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15351094 [Accessed March 18, 2014].

Kim, D. & Salzberg, S.L., 2011. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome biology*, 12(8), p.R72. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245612&tool=pmcentrez&rendertype=abstract [Accessed April 28, 2014].

Kriege, M. & Brekelmans, C., 2004. Efficacy of MRI and mammography for breast-cancer screening in women with a familial or genetic predisposition. *… England Journal of …*, pp.427–437. Available at: http://www.nejm.org/doi/full/10.1056/NEJMoa031759 [Accessed May 4, 2014].

Kytölä, S. et al., 2000. Chromosomal alterations in 15 breast cancer cell lines by comparative genomic hybridization and spectral karyotyping. *Genes, chromosomes & cancer*, 28(3), pp.308–17. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10862037.

Lalloo, F. & Evans, D.G., 2012. Familial breast cancer. *Clinical genetics*, 82(2), pp.105–14. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22356477 [Accessed January 15, 2014].

Li, Y. et al., 2011. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics (Oxford, England)*, 27(12), pp.1708–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21546395 [Accessed April 28, 2014].

Liu, X. et al., 1999. Gamma-heregulin: a fusion gene of DOC-4 and neuregulin-1 derived from a chromosome translocation. *Oncogene*, 18(50), pp.7110–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10597312.

Maher, C. a et al., 2009. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(30), pp.12353–8. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2708976&tool=pmcentrez&rendertype=abstract.

Martin, J. a & Wang, Z., 2011. Next-generation transcriptome assembly. *Nature reviews. Genetics*, 12(10), pp.671–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21897427 [Accessed January 9, 2014].

McPherson, A. et al., 2011. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS computational biology*, 7(5), p.e1001138. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3098195&tool=pmcentrez&rendertype=abstract [Accessed April 28, 2014].

Metzker, M.L., 2010. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1), pp.31–46. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19997069 [Accessed March 19, 2014].

Morrow, M., 2004. Magnetic resonance imaging in breast cancer: one step forward, two steps back? *Jama*, 295(20). Available at: http://jama.jamanetwork.com/article.aspx?articleid=199931 [Accessed May 4, 2014].

Nowsheen, S. et al., 2014. Epigenetic inactivation of DNA repair in breast cancer. *Cancer letters*, 342(2), pp.213–22. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22634493 [Accessed April 28, 2014].

Ozsolak, F. & Milos, P.M., 2011. RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics*, 12(2), pp.87–98. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3031867&tool=pmcentrez&rendertype=abstract [Accessed March 19, 2014].

Parkin, D.M. et al., 2002. Global cancer statistics, 2002. *CA: a cancer journal for clinicians*, 55(2), pp.74–108. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15761078.

Pole, J.C.M. et al., 2006. High-resolution analysis of chromosome rearrangements on 8p in breast, colon and pancreatic cancer reveals a complex pattern of loss, gain and translocation. *Oncogene*, 25(41), pp.5693–706. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16636668 [Accessed January 15, 2014].

Reynisdottir, I. et al., 2013. High expression of ZNF703 independent of amplification indicates worse prognosis in patients with luminal B breast cancer. *Cancer medicine*, 2(4), pp.437–46. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3799278&tool=pmcentrez&rendertype=abstract [Accessed May 4, 2014].

Robinson, D.R. et al., 2011. Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nature medicine*, 17(12), pp.1646–51. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3233654&tool=pmcentrez&rendertype=abstract [Accessed January 15, 2014].

Sanger, F. & Coulson, A., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, p.4095. Available at: http://www.sciencedirect.com/science/article/pii/0022283675902132 [Accessed May 9, 2014].

Sang-Hoon Cho, Jongsu Jeon, S.K., 2012. Personalized Medicine in Breast Cancer: A SyStematic Review. *Journal of Breast Cancer*, 15(3), pp.265–272. Available at: http://ir.ymlib.yonsei.ac.kr/handle/22282913/21778 [Accessed May 4, 2014].

Smith, L. et al., 1986. Fluorescence detection in automated DNA sequence analysis. , p.2497. Available at: http://www.nature.com/nature/journal/v321/n6071/abs/321674a0.html [Accessed May 9, 2014].

Thomas, S. & Munster, P.N., 2009. Histone deacetylase inhibitor induced modulation of anti-estrogen therapy. *Cancer letters*, 280(2), pp.184–91. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19185986 [Accessed May 5, 2014].

Tomlins, S. a et al., 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science (New York, N.Y.)*, 310(5748), pp.644–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16254181 [Accessed March 19, 2014].

Veronesi, U. & Cascinelli, N., 2002. Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer. … *England Journal of* …, 347(16), pp.1227–1232. Available at: http://www.nejm.org/doi/full/10.1056/NEJMoa020989 [Accessed May 4, 2014].

Viale, G., 2012. The current state of breast cancer classification. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO*, 23 Suppl 1(Supplement 10), pp.x207–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22987963 [Accessed May 4, 2014].

Wright, R. & Saul, R. a, 2013. Epigenetics and primary care. *Pediatrics*, 132(Suppl 3), pp.S216–23. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24298130 [Accessed May 5, 2014].

Wu, Y.-M. et al., 2013. Identification of targetable FGFR gene fusions in diverse cancers. *Cancer discovery*, 3(6), pp.636–47. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3694764&tool=pmcentrez&rendertype=abstract [Accessed January 15, 2014].

Appendix A

SRA numbers for the paired-end RNA-seq of breast cancer cell lines which were analyzed with SOAPfuse.

Cell line	SRA number
T-47D	SRR500880
T-47D	SRR500876
T-47D	SRR925736
MCF7	SRR064286
MCF7	SRR925723
MCF10A	SRR925720
MDA-MB-134	SRR925724
MDA-MB-231	SRR925726
SKBR3	SRR925729
SUM-225	SRR934640
SUM-225	SRR934641
SUM-229	SRR925734
SUM-52	SRR925735
ZR-75-1	SRR925740
CAMA-1	SRR925698
BT474	SRR925695

SRA numbers for the paired-end RNA-seq of breast cancer tumors which were analyzed using SOAPfuse.

Sample type	SRA number
Triple negative tumor	SRR1027171
Triple negative tumor	SRR1027172
Triple negative tumor	SRR1027173
Triple negative tumor	SRR1027174
Triple negative tumor	SRR1027175
Normal breast tissue	SRR1027188
Normal breast tissue	SRR1027189
Normal breast tissue	SRR1027190
ER+ tumor	SRR791043
ER+ tumor	SRR791044

ER+ tumor	SRR791045
ER+ tumor	SRR791046
ER+ tumor	SRR791047
ER+ tumor	SRR791048
ER+ tumor	SRR791049
ER+ tumor	SRR791050

Appendix B

Primers designed to amplify and validated possible fusion genes as well as primers used as positive control.

Primer name	Sequence (5´ → 3´)	Tm [°C]
ERLIN2-F	CCGCAGAAACTACGAGTTGAT	57.9
ERLIN2-R	CATCTGCCTTTGCCTTCTCC	59.4
PROSC-F	GGTGATCGAGGCCTATGGAC	61.4
GPR124-F	AGGAACAACATCATCAGCAC	55.3
FGFR1-Rev1	CCATCTGGCTGTGGAAGTC	58.8
FGFR1-Rev2	TTGCCCTTGGAGGCATACTC	59.4
RPS6KB1_TMEM49_F	GAAACTAGTGTGAACAGAGG	55.3
RPS6KB1_TMEM49_R	CATAACTTTGTGCCATGGAG	55.3
ARFGEF2_SULF2_F	CAGGAGAGCCAGACCAAGAG	61.4
ARFGEF2_SULF2_R	ACTTGCCAGTGAGGATGGAG	59.4
SMG5-F	AGCAGAAAGGAGAAGCTCCT	59.4
PAQR6-R	ATGCCATCTTCCCAGAACAC	57.3
NOTCH2NL-F	TGAGTGCAACTGCCTTCCAG	59.4
NBPF10-R	AGGTGCCTCAACTCAGAGCT	59.4
VGLL4-F	CAAGAGGAAGTTCAGCATGG	57.3
SH3BP5-R	CAGTTCATCCAGTTTCACCG	57.3
MECOM-F	GCCAGTCAACCAGATGTTGG	59.4
TTC18-R	GAATGTGTGTTTGGGCCAGC	59.4

Appendix C

Junction sequences as predicted by SOAPfuse for possible fusion genes in MCF7, T-47D and MDA-MB-134.

5′Gene	Junction Sequence	3′Gene
RPS6KB1	CCATGAAGGTGCTTAAAAAG ::: GGAGAAAACTGGTTGTCCTG	VMP1
ARFGEF2	GGCCTGCCAGGTGGCGCTCG ::: GTTCCATGCAGGTGATGAAC	SULF2
NOTCH2NL	TCAAACTTCAGCAGTCATAG ::: ACGGTTACCTGGCACGCTGG	NBPF10
MECOM	CATGCCAGATAAATGATCAG ::: TATGTGCACAGAGTGCTTGC	TTC18
VGLL4	CGCATCTTCAACCCCCATCT ::: GGAGAACTGGAGAAGTTAAA	SH3BP5
SMG5	CTGGCTTTTGGAGTGTTGAG ::: GTCAACGTGGAGGTACCAGG	PAQR6
ANK1	CGACTACTCGCTGTCACCCTCCCAGATGAATG ::: GGAAGTGATGCCGACATGGTGGATAA	ZMAT4
PROSC	CTGTGGCGCGGCGGCCGCGG ::: AGGATCCAGAACTGCAGCCG	OR7E11P