Comparison of SNV Genotyping Sensitivity of Next-Generation Sequencing with Illumina's MiSeq and Quantitative Real-Time PCR with Fluidigm's BioMark HD

Master of Science Thesis

Heidi Bouquin

BioMediTech

University of Tampere

30.08.2016

PRO-GRADU TUTKIELMA

Paikka: Tampereen yliopisto

BioMediTech

Laskennallisen biologian ryhmä ja Eturauhassyövän molekyylibiologian ryhmä

Tekijä: Bouquin, Heidi Maria

Otsikko: SNV genotyypitysherkkyyksien vertailu Illuminan MiSeqin toisen sukupolven sekvensoinnin ja Fluidigmin BioMark HD:n kvantitatiivisen PCR:n välillä.

Sivumäärä: 77 sivua, sivut 33-77 liitesivuja

Ohjaajat: Heini Kallio (PhD), Prof. Matti Nykter, Prof. Tapio Visakorpi

Tarkastajat: Prof. Matti Nykter, Prof. Anne Kallioniemi

Aika: 30. elokuuta 2016

Tiivistelmä

Yhden nukleotidin variantti-genotyypitys (SNV-genotyypitys) on menetelmä, jota voisi soveltaa rutiininomaisesti syöpä-diagnostiikassa. Sitä ei käytetä, koska teknologiat joita menetelmä hyödyntää ovat suhteellisen uusia. Genotyypitys-laitteiden tulee olla luotettavia ja tarkkoja, koska syöpä-DNA:n määrä kudosnäytteissä voi olla hyvin vähäinen. Formaliinilla kiinnitetyt ja parafiiniin valetut näytteet (FFPE-näytteet) ovat esimerkki kudosnäytteistä, joissa on pieni määrä DNA:ta. Toisen sukupolven sekvensointia (NGS) ja reaali-aikaista kvantitatiivista PCR:ä (qPCR) voisi mahdollisesti käyttää SNV-genotyypityksessä, kun kudosnäytteissä on pieniä määriä DNA:ta. Tässä tutkimuksessa genotyypitettiin näytteitä, joissa oli eri osuuksia syöpä-DNA:ta kahdella eri menetelmällä: NGS:lla Illuminan MiSeqillä ja qPCR:lla Fluidigmin BioMark HD:llä. Tulokset osoittavat, että MiSeq kykenee systemaattisesti havaitsemaan yhden nukleotidin variantteja näytteistä, joissa on 10 % syöpä-DNA:ta, joka tarkoittaa 22,5 ng DNA:ta. BioMark HD kykenee havaitsemaan yhden nukleotidin variantteja näytteistä, joissa on 20 % syöpä-DNA:ta, joka tarkoittaa 12,0 ng DNA:ta. Näiden tulosten lisäksi, BioMark HD on luotettavampi, koska se havaitsi kahdeksan kymmenestä variantista, kun MiSeq havaitsi vain neljä.

MASTER'S THESIS

Place: University of Tampere
BioMediTech
Computational Biology Group and Molecular Biology of Prostate Cancer Group

Author: Bouquin, Heidi Maria

Title: Comparison of SNV Genotyping Sensitivity of Next-generation Sequencing with Illumina's MiSeq and qPCR with Fluidigm's BioMark HD

Pages: 77 pages, pages 33-77 appendices

Supervisors: Heini Kallio (PhD), Prof. Matti Nykter, Prof. Tapio Visakorpi

Reviewers: Prof. Matti Nykter, Prof. Anne Kallioniemi

Date: 30th of August 2016

Abstract

Single nucleotide variant genotyping (SNV genotyping) is a method which could be used routinely for cancer diagnostics. It is not, because the technologies utilized are relatively new. Genotyping instruments need to be reliable and precise because the amounts of cancer DNA found in tissue samples can be very small. Formalin fixed paraffin embedded samples (FFPE samples) are an example of tissue samples that contain small amounts of DNA. SNV genotyping with next-generation sequencing (NGS) and quantitative real-time PCR (qPCR) could possibly be used for tissue samples containing small amounts of DNA. In this study samples with different fractions of cancer DNA were genotyped by two different methods: NGS with Illumina's MiSeq and qPCR with Fluidigm's BioMark HD. The results show that MiSeq is able to systematically detect single nucleotide variants from samples with a 10% fraction of DNA, representing 22.5 ng of DNA. BioMark HD is able to detect single nucleotide variants from 20% DNA fractions, representing 12.0 ng of DNA. In addition to these results, BioMark HD is more reliable, because it detected eight out of ten variants, while MiSeq only detected four.

Acknowledgements

I would like to thank all of the people who made it possible for me to finish my master's thesis. It has been a long process, thanks to the surprises in life, but it's finally come to an end. And now it's time to move on to other things.

I want to thank Prof. Matti Nykter from the Computational Biology group and Prof. Tapio Visakorpi from the Molecular Biology of Prostate Cancer group, for giving me the opportunity to become more familiar with sequencing and genotyping during the process of my thesis. Thank you both for giving me a chance to use the resources you have in your research groups and thank you for your time and advice.

I want to thank my supervisor Heini Kallio (PhD) from the Molecular Biology of Prostate Cancer group, for all her time, patience and advice. Thank you for guiding me through it all.

I also want to thank Mauro Scaravilli, Tommi Rantarepo, Merja Helenius and Annika Kohvakka for the advice they gave me at different stages of my thesis.

Last, but certainly not least, I would like to thank my husband and daughter, and the rest of my family for all of their support and encouragement.

Table of Contents

Abbreviations

ASP	SNPtype assay allele-specific primer
CCLE	Cancer cell line encyclopedia
COSMIC	Catalogue of somatic mutations in cancer
CRT	Cyclic reversible termination
ECD	Enrichment control DNA
FRET	Fluorescent resonance energy transfer
FFPE	Formalin-fixed and paraffin embedded
IFC	Integrated fluid circuit
IGV	Integrative genomics viewer
MAF	Minor allele frequency
NGS	Next-generation sequencing
PSMA	Prostate-specific membrane antigen
qPCR	Quantitative real-time PCR
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant

1. Introduction

Nucleic acid-based diagnostic methods for finding biomarkers are being researched for possible clinical usage (1). Genomic alterations such as single nucleotide variants (SNVs), which may cause cancer and which are the most common genomic alterations found in cancer (2), need to be recognized in a simple manner. Tissue samples may have very low frequencies of genomic alterations (2), including variants such as single nucleotide polymorphisms (SNPs). This is partially because cancerous tissues also have non-cancerous cells in them. Some samples have small amounts of DNA in them to begin with.

Hospitals and research groups around the world have thousands of formalin-fixed and paraffin embedded (FFPE) samples, which contain small amounts of DNA that could potentially be used for studies (3, 4). Biopsy samples are commonly stored as FFPE samples because they can be stored for very long periods of time. The average age of FFPE samples used in hospitals is 20 years (1). This means that FFPE samples can be used in retrospective studies (3, 4), which are convenient for patients, because there is no need of new tissue. The amount of usable DNA in FFPE samples is small due to several reasons. The samples themselves are usually quite small, which directly reflects on the amount of DNA, but also DNA is always lost from such samples due to the method of sample preparation. FFPE preparation methods compromise the quality of the DNA because DNA-tissue protein cross-links form (5, 6). It appears then, that very sensitive instruments are required for cancer detection and the study of FFPE samples. Measurements done with FFPE derived DNA were not performed in the scope of this study however.

Instruments with different genotyping techniques are manufactured for finding genomic variants. There are different techniques for genotyping such as microarray techniques (7), sequencing (8), and polymerase chain reaction (PCR) (9), to name a few. However, some techniques are more suitable for some samples than others (8). Some instruments and methods require larger amounts of starting DNA, in which case they may not be suitable for genotyping small cancer tissue or FFPE samples. The instruments may not be able to detect genomic alterations present as minor allele fractions (MAFs).

How small are the amounts of DNA that can be extracted from FFPE samples for subsequent study? Gilbert *et al.* 2007 did a study where they compared the published methods of DNA

recovery related to FFPE samples. Their results showed that DNA extraction yielded mean and median values of 169 ng/µl and 54 ng/µl respectively. If most samples yield approximately 54 ng/µl of DNA, is it enough for genotyping with next-generation sequencing (NGS) or quantitative real-time PCR (qPCR)?

A study by Beltran *et al.* 2013 showed that NGS can be used on DNA from FFPE samples. According to them 55 ng of DNA was required in order to have deep sequence coverage. According to Swango *et al.* 2007, as little as 1 ng of DNA can be used for genotyping with qPCR (10). In short, both NGS and qPCR methods could be used for samples with small amounts of DNA.

NGS could have a large impact on cancer diagnoses. This form of sequencing has proven to be very useful when studying cancer genomes (2), and is the most commonly used type of sequencing for FFPE samples nowadays (4). The small amounts of cancer genomes in samples is not a problem if NGS is used (2). qPCR provides another way to genotype.

There are many genotyping methods involving PCR (11), but they do not give quantitative results (12). qPCR is able to give such results in real-time as the PCR reaction occurs (12). This makes qPCR a versatile tool. qPCR can be used as a tool for diagnosing cancer and it can be used to determine patients' prognoses (12).

Clinical use of sequencing (2) and qPCR technologies (12) could lead to early diagnoses of cancer, faster responses to treatment and better prognoses. Genotyping could also be used as a method to find biomarkers common to cancer and to ascertain whether dissemination has occurred (2). It could also be used for checking whether or not the cancer treatments used have diminished the cancers, if tissue samples were available after treatment. In the future, making therapeutic plans for cancer patients will be assisted by the use of genome-based methods (2). This however, is still in the future. To get to this future it is necessary to improve our understanding of the genotyping technology at our disposal at the moment.

In this study samples with different fractions of cancer DNA were genotyped by two different methods. We wanted to find out if one of the methods were more sensitive to small fractions of cancer DNA containing SNVs. The first method was sequencing with Illumina's MiSeq. Agilent's HaloPlex Cancer Research Panel was used for making a DNA library and a target

enrichment protocol was used for capturing the sequences of interest. These sequences were then sequenced with MiSeq Benchtop Sequencer, which uses NGS technology. The sequencer uses Illumina's own sequencing by synthesis technology for strand extension (http://www.illumina.com/documents/products/datasheets/datasheet_miseq.pdf, 16.06.2016), in combination with cyclic reversible termination, single molecule templates and real-time sequencing (8).

The second genotyping method was qPCR. Fluidigm's BioMark HD qPCR was used for SNV genotyping. The instrument uses integrated fluid circuits (IFCs) in its 96.96 well plates which allows for high throughput, samples consisting of nanoliter volumes and automation (https://www.fluidigm.com/ifcs, 27.06.2016).

2. Literature Review

2.1 Prostate cancer

Out of all of the cancer related deaths in the Western world, prostate cancer deaths are among the top of the list (13). Histological prostate cancer is a form of cancer in which the cancer may remain indolent (14). An estimation of 40% of men over 50 years of age have histological prostate cancer (14). Usually during the progression of prostate cancer, approximately 40% of those who have localized prostate cancer develop metastases (13). Castration resistant prostate cancer is the last phase of the spectrum of the disease (13). The progression of the cancer is accompanied with a myriad of genetic alterations.

Genomic alterations can affect genes or regulatory pathways involved in tumorigenesis and disease progression (13). Different kinds of genomic alterations, such as structural alterations, indels and substitutions, play a role in prostate cancer development and progression (13). Some alterations can be found in genes present in the germline, such as *MSR1* and *RNASEL* (*HPC1*) (14). The genes *AR*, *PTEN* and *p53* among many others, are frequently mutated in somatic prostate cancer (14). Copy number variation can also be commonly found in genes *KLF5* and *MYC*, to name a few (14). Depending on the type of prostate cancer, different therapies are available.

Therapies vary from surgery to pathways-based therapies. Several forms of surgery are used for the removal of prostate cancer, such as open surgery and minimally invasive radical surgery (15). Surgery and other forms of therapy are often combined. As an example, radical prostatectomy and radiotherapy can be combined in the curative treatment of localized prostate cancer (15). Chemotherapy is also used (15). High intensity focused ultrasound therapy has also been used in some cases of prostate cancer, but the efficacy of the treatment is still questionable (15). Pathway-based therapies are also used to alter cell signaling through pathways affecting tumor suppressor genes and oncogenes (14). An example of such therapies is the use of a therapeutic agent called rapamycin that inhibits mTOR protein kinase, which is involved in controlling the cell cycle (14). In the case of castration resistant prostate cancer, there is no curative therapy (13). In such cases, and also in cases with metastases, androgen ablation therapy is used (15). There are vaccine based immunotherapies being developed, but they are not in clinical use yet (15). In cases where the cancer is low-risk, active surveillance

and watchful waiting are common approaches (15, 16). Regardless of the type of therapy chosen, early diagnosis is important.

The sensitivity of prostate cancer detection has increased. This is due to the use of prostate-specific antigens (PSA) in prostate cancer detection and in following the progression of the cancer (17). In combination with histological methods, the Gleason Score is used for diagnosing prostate cancer (14). Genomic biomarkers are of great interest, since their use in the clinics could aid making diagnoses and prognoses (13). Studying FFPE samples of old prostate cancer biopsies could bring more understanding to the field.

2.2 Breast Cancer

Breast cancer related deaths are among the most common cancer related deaths in the world (http://www.who.int/mediacentre/factsheets/fs297/en/, 18.08.2016). The most common types of breast cancer are ductal or invasive and lobular breast cancer (18). Whether the diagnosis of cancer in both breasts is done at the same time or at different times determines whether the cancer is synchronous or metachronous, respectively (19). The risk of metachronous breast cancer is higher than that of synchronous breast cancer, 3-13% vs. 1-5% (20). Women with bilateral breast cancer have poorer prognoses compared to women with unilateral breast cancer (20). The prognosis is considered poor if metastases start to form (21). This is because the disseminated cell clones are considered aggressive because they have more genetic alterations compared to the cells found in the primary tumor (21).

The genetic landscape of breast cancer can be quite varied. In some cases the progression of the tumor from certain clones can be seen, but in other cases a variety of genetically divergent clones can be found in the tissue samples (22). One frequently found genetic alteration in breast cancer is loss of heterozygosity (22). Changes in the function of genes *p53* and *BRCA* are also found in the formation of breast cancer (21), among many other mutations. In cases of metastases genetic mutations affect many genes (e.g. proto-oncogenes and tumor suppressor genes) and pathways, such as DNA repair pathways (21).

Several treatment methods are available for breast cancer. Quandrantectomy and mammary segmental resection are types of surgery where only the affected regions of the breast are removed (18). Modified radical mastectomy, partial mastectomy and breast conserving

surgery are other types of surgery used for treating breast cancer (23). Other treatment methods also exist such as radiotherapy, chemotherapy and endocrine treatment (18, 23), often in combination with some type of surgery. The type of treatment depends on how early the cancer is detected and on the type of cancer.

Breast cancer is generally diagnosed by clinical examination (18, 23), mammography or ultrasound (18, 20). In some cases it is diagnosed with the help of magnetic resonance imaging (20), because tumors might not be visible with the former methods even though present. In cases where there is cause to believe the presence of a tumor, biopsies are taken for sampling (18). Genetic diagnostics could help in making a more accurate diagnose of the type and severity of the cancer.

2.3 FFPE samples

FFPE samples of tissues are frequently made for the purpose of studying tissues and their molecular make-up (1, 3) The DNA and RNA found in the samples are often of interest (4). Formalin-fixation is a very frequently used method for fixing samples in histopathology (3, 4). Why these samples are so widely used is explained by their many good characteristics such as the low cost of the method, the possibility of long storage time, ease of handling and keeping of the quality of the samples through time (3). But most importantly, formalin fixation keeps samples relatively close to their *in vivo* morphology (4).

Unfortunately formalin-fixation has some setbacks. Fixation of tissues with formalin creates DNA-tissue protein cross-links in the samples (6), which hinder amplification of the DNA (3). The cross-linking is reversible to some extent (4). The aging of FFPE samples and changes in the fixative pH cause fragmentation of nucleic acids (2), which results in poor quality of the DNA extracted from the samples. According to some researchers, for each decade of storage, FFPE samples go through 5-50% degradation of nucleic acids (1). Also, there is no standardization in the multi-stepped method of specimen preparation (6).

The way FFPE specimens are specifically made varies from laboratory to laboratory, but the general steps are the same. The first step of the process of creating FFPE samples is fixation. 10% formalin, which is a 2-phase fixative, is the most commonly used fixative. It consists of formaldehyde and water. The first phase of fixation occurs with alcohol. The second phase,

which is done with the assistance of aldehydes, is a cross-linking phase. After fixation comes paraffin embedding. (1)

There are several steps involved in paraffin embedding. The first step, dehydration, involves moving of the specimen from an aqueous environment to an environment with alcohol. Clearing, which is the next step, is a subsequent removal of the alcohol and replacement with xylene. After clearing, comes impregnation, during which the xylene is replaced with paraffin. After all of these different replacement steps the tissue sample is surrounded with paraffin in embedding. At this point the sample is ready for sectioning with a microtome and long term storage. (1)

The study of fragmented DNA from FFPE samples was problematic over 10 years ago (4). The fragmented DNA could not be examined in a reliable way. Now, during the '-omics' era, examination of such samples is easier because of the new techniques available (4). Therefore interest in FFPE samples has increased.

2.4 Other kinds of tissue samples

Different kinds of tissue samples are taken from patients for histological, molecular and genetic analyses. Larger samples contain more DNA to study, but often only small samples can be taken.

Biopsies are a common type of tissue sample taken from patients. There are several kinds of biopsies procured from tissue depending on the size and location of the atypical tissue: fine needle aspiration and core needle biopsies (18), and surgical biopsies (http://www.cancer.org/treatment/understandingyourdiagnosis/examsandtestdescriptions/forwomenfacingabreastbiopsy/breast-biopsy-biopsy-types, 24.08.2016). In fine needle aspiration a small tissue sample is aspirated into the syringe, while in core needle biopsies a hollow needle aspirates a narrow column of tissue into the syringe. Even though the sample volumes in needle biopsies are small, the samples have enough of DNA in them for analysis (13). Surgical biopsies require a surgical procedure, as the name suggests. Tissue samples taken by biopsy are sometimes frozen for later DNA studies (4). Unfortunately, most hospitals do not have the capacity to storage large numbers of frozen samples for the duration of years (1, 4). But some tissues are frozen explicitly.

Some tissue samples are frozen directly after their retrieval from the patient (http://www.amsbio.com/Tissues-Frozen-Tissue-Sections.aspx?cty=FINLAND&cur=EUR, 24.08.2016). Such samples are frozen tissue sections. The fresh tissue samples are frozen in liquid nitrogen before they are cut into sections, which are then viewed. Other sample types remain fluid.

Biofluids can be used for diagnostics. All fluids from the body, such as blood, urine and tears, can be studied and used for diagnostic purposes (24). The fluids and their molecular composition can be analyzed *in vivo* or uninvasively.

2.5 Next-Generation Sequencing

All useful techniques are improved in time, the same goes for sequencing. Sanger sequencing is referred to by the term first-generation sequencing (2, 8), and it is a form of sequencing that is still used today, but less. Sanger sequencing is an analogue form of sequencing (2).

In Sanger sequencing there are two phases which are performed separate from each other: first, the synthesis, and second, the electrophoresis. During synthesis, a complementary DNA strand is synthesized by DNA polymerase which incorporates deoxyribonucleotide triphosphates (dNTPs) and dideoxyribonucleotide triphosphates (ddNTPs) into the strand (25), in a random fashion. When a ddNTP is attached to the strand, the synthesis of the complimentary strand is terminated (25). As a result, many different strands with varying lengths are produced and these fragments are separated by gel electrophoresis, after which their analysis reveals the base order of the sequence (25). Because Sanger sequencing is not digital, there are things that cannot be done with the method.

First generation sequencing is much more limited in the information that it gives. Sanger sequencing cannot be used to detect all of the possible alterations found in cancer simultaneously, which are deletions, small insertions, nucleotide substitutions, copy number alterations and chromosomal rearrangements (2). It is why NGS was developed.

NGS is a digital form of sequencing (2). Single molecules of DNA are used in array-like amplification and subsequent recognition by computer (2). The method uses over-sampling (2), which refers to the reading of the DNA sequence multiple times and the production of

many reads. Over-sampling brings an increase of sequence read-out confidence (2). The technology has increased throughput (8), making it possible to sequence and analyze numerous samples at the same time. NGS also allows for complex analysis of DNA and genomes, since deletions, small insertions, nucleotide substitutions, copy number alterations and chromosomal rearrangements – all the things Sanger sequencing could not do simultaneously – can be studied at the same time. The technology has made it possible to study aligned sequencing reads anywhere in the genome (2). Next-generation sequencing can be used for many other applications as well. Different NGS methods are available on the market.

There are different methods of sequencing with NGS. Some of the methods used are picotitre-plate pyrosequencing, ligation-based sequencing and single-nucleotide fluorescent base extension (2). In general there is a set of main phases in all of the methods: template preparation, sequencing combined with imaging and the analysis of the sequencing data (8). A DNA library is made during template preparation.

Template preparation includes the creation of a DNA library, which is then used as a template for sequencing (25). The process requires the fragmentation of genomic DNA. There are two kinds of templates used, clonally amplified or single-molecule templates (25). In the former, a DNA library is made, denatured into ssDNA, attached to beads or a solid surface and then finally amplified in order to increase the fluorescent signal during imaging. In the latter, single molecules of template are attached to a solid surface and no PCR is required. Sequencing and imaging follow.

Sequencing and imaging of clonally amplified and single-molecule templates is different. After the addition of a single nucleotide or probe to the template during sequencing, the signal is treated differently, depending on which template is used (25). During the imaging of clonally amplified templates, the fluorescent signals from a batch of the same amplified DNA molecule are treated as one consensus signal. If a single-molecule template is used, then each fluorescing nucleotide or probe gives an independent fluorescent signal. Imaging of the fluorescing signals occurs after each sequencing cycle. Sequencing techniques differ in the sequencing platforms available.

The techniques used for sequencing vary in the chemistry used. Cyclic reversible termination (CRT), sequencing by ligation, pyrosequencing and real-time sequencing are techniques found on the NGS market (8). NGS that uses CRT, incorporates a single modified nucleotide to the growing complementary strand per cycle, after which imaging occurs (8). After imaging, the terminating or inhibiting group and the fluorescing dye are cleaved off. After cleavage, a new cycle can begin. The technique is very different compared to sequencing by ligation.

Labeled probes are used in sequencing by ligation (8). The probes attach by hybridization to complimentary sequences next to the primed template. The labeled probe and primer are attached by DNA ligase. Fluorescent imaging takes place, after which either the probes are removed or the primers are replaced, and a new cycle can occur.

Pyrosequencing is a bioluminescent method. In pyrosequencing, the termination of DNA synthesis is done by limiting the amount of a single type of dNTP added (8). The DNA polymerase adds a nucleotide to the primer and stops. When more dNTP is added, the synthesis continues. When a pyrosequencing reaction occurs and pyrophosphate is released, the light generated is measured by a camera. The intensities are recorded as flowgrams and give the order of the bases in the sequence.

During real-time sequencing, imaging occurs at the same time as DNA synthesis which is not stopped at any moment (8). In some real-time sequencing technologies, a fluorescing dye is used in signaling, while in others a dye-quencher group is used for emitting a signal (8). Once sequencing is performed, the data can be analyzed.

2.6 Genotyping with qPCR

There are different genotyping methods which utilize PCR such as long-distance PCR methods and inverse shifting PCR (11). Some of the methods are very labor-intensive and time consuming (11). The development of qPCR has made things quicker and more efficient. Real-time PCR techniques make it possible to follow the production of the PCR end product in real time. The techniques involved incorporate the use of detection probes (12).

The labeled detection probes used in qPCR make it possible to follow the production of the end product in the instant they are produced (12). The probes fluoresce when end product is formed. One of the labeling techniques uses the exonuclease functions of Taq DNA polymerase (12). A quencher dye is attached to the 5' end of the probe and a reported dye to the 3' end. When the Taq DNA polymerase begins elongation, it first cuts off the quencher dye with its exonuclease activity. This causes the reporter dye to fluoresce in ratio to the amount of produced end product. Fluorescent resonance energy transfer (FRET) probes are also used for end product detection.

When FRET is used, two probes are utilized; one upstream and another downstream. The probe upstream has an excitatory dye at its 3' end, while the probe downstream has a reporter dye at its 5' end (12). The two probes hybridize during the annealing phase of PCR when there is end product. After hybridization the excitatory dye gives an electron to the reporter dye, causing it to fluoresce. The intensity of the fluorescence is then measured. Molecular beacons are also used in quantitative real-time PCR.

There are three components to molecular beacons (12). The first component is the tagged probe, of which there are two. They are end-product specific and have a quencher and a reporter dye at opposing ends. The second component consists of two complimentary sequences in each probe, one on the 5' end and one on the 3' end, allowing for the formation of a "stem". The third component is in the loop which is formed in the probe: a target specific sequence. The molecular beacon has an "on" and "off" position. Initially the beacon is off and no signal is emitted. This is when the PCR cycle is at or below annealing temperatures and the beacon is in a stem and loop conformation. When the stem is formed, the quencher and reporter dye are close to one another, resulting in no signal. When end-product formation begins and when an end-product molecule hybridizes with the target-specific sequence in the loop the beacon is turned on, the dyes are removed and a signal emitted.

3. Objectives

There were two objectives to this experiment. The first objective was to test the sensitivities of Illumina's MiSeq Benchtop Sequencer and Fluidigm's BioMark HD qPCR. The second objective was to measure or to estimate the minimal fractions of cancer DNA that the two instruments could detect.

4. Materials and Methods

4.1 Cell lines

Two cells lines were selected for this study: prostate cancer cell line LNCaP clone FGC and breast cancer cell line MDA-MB-415. Both cells lines were from Tapio Visakorpi's Molecular Biology of Prostate Cancer group from the University of Tampere, Finland. These two specific cell lines were selected, because they both contained SNVs in certain genes which could be targeted by Agilent's HaloPlex Cancer Research Panel Kit (Agilent Technologies, Santa Clara, USA), see **Table 1** in **Appendix 1.**

Five different SNVs were selected for targeting from each cell line. Five variants per cell line were considered sufficient because most of the variants in the HaloPlex Cancer Research Panel were found in both cell lines, which would have been problematic. The SNVs were different, so that when SNV detection occurred, it would be clear in which cell line the mutation was found in. The online databases Catalogue of Somatic Mutations in Cancer (COSMIC) at (http://cancer.sanger.ac.uk/cosmic, 07.01.2014) and Cancer Cell Line Encyclopedia (CCLE) at (http://www.broadinstitute.org/ccle/home, 07.01.2014) were used to verify which mutated genes in the cancer panel were found in the cell lines used.

During culture, the growth medium used for the LNCaP cells was ATCC-formulated RPMI-1640 Medium, which was supplemented with fetal bovine serum (FBS) up to a concentration of 10% and 1% L-glutamine, see **Appendix 2** for a list of reagents and kits used. The cells were detached from the flask with trypsin for subculturing. All washes were done with phosphate-buffered saline (PBS). The cells were incubated at a temperature of 37 °C.

The MDA-MB-415 cell line used Leibovitz's L-15 medium with 2mM L-glutamine and was supplemented with 10 µg/ml insulin, 10 µg/ml glutathione and FBS, of which the last supplement had a final concentration of 15%. When subculturing, the cells were detached from the flask by scraping. All washes were done with PBS. The MDA-MB-415 cells were also incubated at 37 °C, but separately from the LNCaP cells in an incubator with only free gas exchange with the surrounding atmospheric air, because Leibovitz's L-15 medium is not suitable for cells in an environment with a CO_2 and air.

4.2 DNA extraction

Qiagen QIAamp DNA Mini kit was used for DNA extraction. The LNCaP and MDA-MB-415 cells were collected separately from their culture flasks according to the "Protocol for Cultured Cells in QIAamp DNA Mini and Blood Mini Handbook, third edition, June 2012". A cell count was performed. The cell count was done with two methods: manually by a hemocytometer and digitally with Moxi Z Mini Automated Cell Counter (ORFLO Technologies, Ketchum, USA). Their average was used for cell number estimation to be sure it did not exceed the maximum number specified by the protocol, which was 5 x 10^6 cells.

After the cell count, the protocol "DNA Purification from Blood or Body Fluids (Spin Protocol)" which was also in the QIAamp DNA Mini and Blood Mini Handbook was used for the extraction of DNA. The DNA was eluted into a buffer provided by the kit.

4.3 Measuring of DNA concentration

The concentration of extracted DNA was measured with Qubit 3 fluorometer (ThermoFisher Scientific, Waltham, USA). The manual "Measuring of DNA with Qubit" was followed.

4.4 Samples with varying fractions of DNA for sequencing

Fifteen samples with different fractions of LNCaP and MDA-MB-415 DNA were made, see **Table 1**. Each sample had a combined total DNA amount of 225 ng, which was required by Agilent's HaloPlex Target Enrichment System protocol. DNase-free water was used for making the dilutions.

4.5 Agilent HaloPlex Target Enrichment System

The protocol HaloPlex Target Enrichment System for Illumina Sequencing (Version D.5, May 2013) was used for making a sequencing library suitable for Illumina paired-end multiplex sequencing. The workflow for the protocol can be seen in **Appendix 3 Figure 1.** The protocol was followed step by step, using the reagents provided by HaloPlex Cancer Research Panel Kit.

4.5.1 Digestion

The 15 samples and one control were prepared for the protocol. The control was an Enrichment Control DNA (ECD) sample provided by Agilent, but nothing of its contents was disclosed.

Table 1. Cell line DNA mixtures for sequencing. The ratio of LNCaP and MDA-MB-415 DNA in the samples is shown. Samples 10-15 are internal replicates.

Sample	LNCaP-%	Amount of DNA (ng)	MDA-MB-415-%	Amount of DNA (ng)
1	0	0.0	100	225.0
2	10	22.5	90	202.5
3	20	45.0	80	180.0
4	30	67.5	70	157.5
5	50	112.5	50	112.5
6	70	157.5	30	67.5
7	80	180.0	20	45.0
8	90	202.5	10	22.5
9	100	225.0	0	0.0
10	0 replicate	0.0	100 replicate	225.0
11	10 replicate	22.5	90 replicate	202.5
12	20 replicate	45.0	80 replicate	180.0
13	80 replicate	180.0	20 replicate	45.0
14	90 replicate	202.5	10 replicate	22.5
15	100 replicate	225.0	0 replicate	0.0

The genomic DNA samples were digested in eight restriction reactions A-H, see **Figure 1**. Each reaction had two different restriction enzymes. The program for the thermal cycler (BioRad, Hercules, USA) during the digestion was according to the protocol.

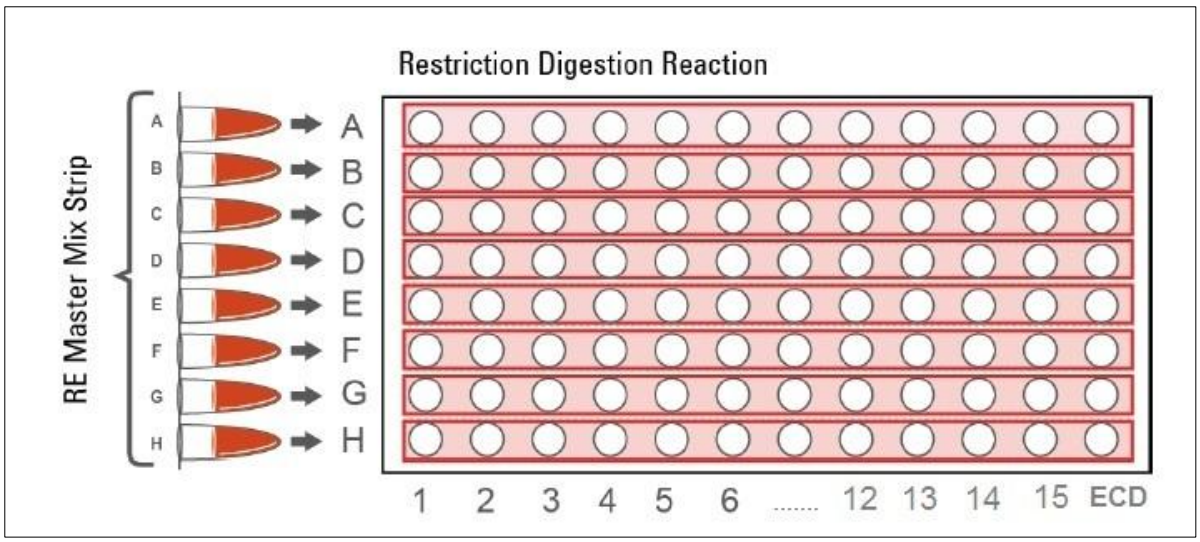

Figure 1. Restriction reactions for gDNA. The DNA samples 1-15 and the control ECD, were digested in eight restriction reactions A-H, each containing two unknown restriction enzymes. For simplification, only one 96-well plate with samples is shown, but two plates were used. Modified Figure from HaloPlex Target Enrichment System Protocol for Illumina Sequencing, Version D.5, May 2013.

4.5.2 Validation of ECD Restriction Digestion

Validation of the restriction digestion was done with 2100 Bioanalyzer (Agilent Technologies, Santa Clara, USA) and a High Sensitivity DNA Kit. The analysis with the Bioanalyzer was an electrophoretic analysis. The protocol used for the validation was "Agilent High Sensitivity DNA Kit Guide, G2938-90321 Rev. B, Edition 11/2013". Only the digested ECD reactions were analyzed in this validation.

4.5.3 Hybridization of DNA to HaloPlex Probes

Hybridization of the digested DNA to HaloPlex probes was done. At the same time, the samples were indexed for sequencing by adding Indexing Primer Cassettes provided by HaloPlex Cancer Research Panel Kit. Indexing was done according to sample number; sample 1 was given index #1, sample 2 was given index #2, and so forth for all 15 samples and the one control. During hybridization, sequencing motifs made by Illumina were automatically also added to the DNA fragments, see **Figure 2.** During hybridization the DNA probes directed the circularization of the targeted DNA fragments. Hybridization was done for 3 hours in a thermal cycler, according to the appropriate program indicated by the protocol.

4.5.4 Capturing the Target DNA

During the capture-phase of the protocol, the target DNA-HaloPlex probe hybrids were captured. The hybrids contained biotin, which made it possible to capture them with beads coated with streptavidin. The Agencourt AMPure XP Kit with its beads and reagents was used for the capture reaction. For an optimal capture reaction, a fresh and specifically diluted batch of NaOH was made. Therefore, specific guidelines were used, see **Appendix 4.**

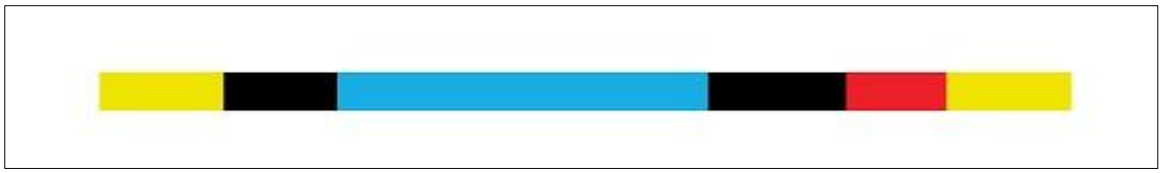

Figure 2. Content of HaloPlex-Enriched Target Amplicons. All amplicons contained the following parts: target insert (blue), Illumina's sequencing motifs (black), index (red) and library bridge PCR primers (yellow). Figure from HaloPlex Target Enrichment System Protocol for Illumina Sequencing, Version D.5, May 2013.

4.5.5 Ligation of Fragments

The nicks in the circularized HaloPlex probe-target DNA hybrids were closed using DNA ligase. The samples were incubated in a thermal cycler.

4.5.6 Preparation of PCR Master Mix

A master mix was made for the PCR reaction according to the protocol. Reagents not supplied by the kit are mentioned in **Appendix 2**.

4.5.7 Elution of Captured DNA

Elution of the captured DNA libraries was done with NaOH. The target DNA was released from the beads during this step.

4.5.8 Amplification of Captured Target Libraries

Amplification of the captured target libraries was done with PCR (BioRad). The program used is shown in **Table 2.**

Table 2. Amplification Program. The program used for the amplification of the captured target DNA. Segment 2 of the amplification consisted of 23 cycles.

Segment	Number of Cycles	Temperature (°C)	Time
1	1	98	2 minutes
2	23	98	30 seconds
2		60	30 seconds
2		72	1 minute
3	1	72	10 minutes
4	1	8	Hold

4.5.9 Purifying of the Target Libraries

The amplified target DNA was purified with the help of AMPure XP beads. 70% ethanol was used for the washes performed in this phase. Tris-acetate was used in the elution of the DNA. 4 µl of each library was set aside for the validation of the enrichment with Bioanalyzer.

4.5.10 Validation of Enriched Target DNA

The enriched target DNA was validated with two different devices. Originally the 2100 Bioanalyzer (Agilent Technologies) was supposed to be the only device used for validation.

The device did not work reliably, and so another device, LabChip GXI (PerkinElmer, Waltham, USA) was used to validate some of the samples.

There were two purposes for sample validation. The first one was to verify that there was a peak between 225 and 525 bp in the electropherograms representing the amplicon. The second was to determine the concentration of the enriched DNA by performing peak integration between peaks at 175 and 625 bp. For samples with too high a concentration (above 10 ng/µl), 1:10 dilutions were made with water and the samples were run again.

4.5.11 Pooling of DNA Samples

Equimolar amounts of indexed sample DNA had to be pooled for sequencing. The concentration values from the Bioanalyzer and LabChip measurements were used. Making a single equimolar DNA pool was impossible, because of the range of differences in molarity, so therefore two separate DNA pools were made, see **Table 3**. Those samples which had higher molarities were pooled into DNA Pool 1 and samples which had low molarities were pooled into DNA Pool 2. The samples in DNA Pool 2 happened to be the same ones that did not give reliable measurement values with the Bioanalyzer and were measured with LabChip GXI. See **Appendix 5** for an example of the calculations.

After pooling the samples into DNA Pool 1, the pool went through a round of AMPure XP bead purification. This additional purification was done, as was suggested in the protocol, if any of the samples had more than 10% molarity of adapter-dimer (at 125-150 bp) in the electropherograms compared to the peak value. The molarity of the adapter-dimer was more than 10% in most cases.

Pooling of samples into DNA Pool 2 was difficult. The required volume of DNA for each sample surpassed the amounts that were available. Since there was no time to grow more cells for the experiment, an improvisation was done. 5 µl of each sample was pooled together because their molarities were in the same range. Sample 12 was an exception; only 2.5 µl of DNA was pooled for the sample because it had twice the molarity of the other samples. In this way the pool had an average molarity of 8.3 nmol/l. The total volume for Pool 2 was 32.5 µl. The pool was not diluted by the addition of water.

Table 3. Pooling of DNA Samples. The DNA samples were pooled into two separate DNA pools prior to sequencing.

DNA Pool 1 Samples	DNA Pool 2 Samples
1, 4, 6, 9, 10, 11, 13, 14, 16	2, 3, 5, 7, 8, 12, 15

4.6 Sequencing with Illumina's MiSeq Benchtop Sequencer

A sample sheet with sample numbers and indexes was prepared using "Agilent's HaloPlex Target Enrichment System-ILM" protocol. After this the DNA library and PhiX Control were prepared for sequencing with Illumina's protocol "Preparing Libraries for Sequencing on the MiSeq, part # 15039740 Rev C August 2013". During the sequencing run, MiSeq automatically sent sequencing data to BaseSpace, a cloud-based genomics data hub.

DNA Pool 1 and 2 were handled separately. The DNA library was denatured and diluted to a final concentration of 2 nM according to the above mentioned protocol with HT1 Hybridization Buffer from MiSeq v2 Reagent Kit. Freshly diluted NaOH was used in all of the steps. Immediately before sequencing of the library an additional dilution to 6 pM was done according to the protocol.

A 5% PhiX spike was used as a control during sequencing. 30 µl of denatured and diluted PhiX control was added to 570 µl of 6 pM DNA library. Then the library was loaded into the MiSeq Reagent Cartridge and was ready for sequencing.

Illumina's protocol "MiSeq System User Guide Part # 15027617 Rev. H March 2013" was used during the setup with PR2 reagent and HT1 Buffer (from MiSeq v2 Reagent Kit). The above mentioned protocol was also used during the automated sequencing of the DNA library.

The first sequencing run of Pool 1 on Illumina's MiSeq failed. Since the run failed and no data could be obtained for the run, an assumption was made that the DNA library had too high a concentration, therefore perhaps causing over clustering of the flow cell. The assumption was made on the basis that the sequencing run could not be finished and because no reads were given by MiSeq. The ready DNA library was diluted for a second run. The dilution was a 1:10 dilution with water. Otherwise everything was done according to the protocol. The PhiX spike was kept the same, as a 5% spike.

The sequencing of Pool 2 failed. Since the pool was already very dilute and because there was no time to grow more cells for new samples, it was decided that the sequencing of Pool 1 was enough for this study.

4.7 Analysis of Sequencing Data

After sequencing, the data of the run was analyzed. For the data to be in such a form that it could be analyzed, several computational methods were used.

The Illumina adaptor sequences were removed from the ends of the fastq-files by trimming. Each read was trimmed by 30 bases from the 5' end to remove the adaptor. 50 bases were also removed from the 3' end in order to remove poor quality material. These values were chosen because the subsequent alignment worked properly. A tool called Pypette was used for trimming (https://github.com/annalam/pypette, 29.03.2016). See **Appendix 6**, for a list of the used scripts.

A program called Bowtie2 was then used for aligning the trimmed reads to the reference human genome (version 19). Default parameters were used with the program. At this point the files were compressed .gz files. All the reads were aligned at the same time as a batch.

A computational tool called SAMtools (Sequence Alignment/Map) was used for several computational steps prior to viewing the alignments with Integral Genome Viewer (IGV). This was necessary so that IGV could utilize the sorted bam-files. SAMtools View was used for the conversion of .sam files to .bam files. SAMtools Sort was then used to arrange the reads into order according to the reference genome coordinates. SAMtools Index was then used to index the .bam files. All SAMtools steps were combined to form a loop, in which each sample went through all of the different steps in an automated way. See **Appendix 6** for a list of the used scripts.

After all of the above mentioned computational processes the files were ready for viewing. The program IGV was used for viewing the sequencing reads. The program was downloaded from the internet website (https://www.broadinstitute.org/igv/; 12.08.2014). An analysis of all of the .bam files which contained all of the reads was done. Both .bam and .bai files were required for the viewing of reads. .bai files were created automatically when files were

converted to .bam format. The two file types were kept in the same folder even though only the .bam files were opened manually with IGV. IGV opens .bai files by itself at the same time when .bam files are manually opened.

4.8 qPCR with Fluidigm's BioMark HD

Fluidigm's BioMark HD quantitative real-time PCR (qPCR) was used for SNV genotyping the samples. The same SNVs that were looked for by sequencing were also searched for with qPCR genotyping. The following protocol was used for all steps: "Fluidigm Genotyping User Guide, SNPtype Assays for SNP Genotyping on the Dynamic Array IFCs, PN 68000098 Rev J1". All reagents used during the process can be found in **Appendix 2**.

Fluidigm's BioMark HD uses Integrated Fluid Circuits (IFCs), see **Figure 3**, which make it possible to run many assays at the same time. In this study a 96.96 IFC was used. The assays and samples were combined in 9216 separate reactions due to the network of microfluidic channels and valves placed in the center of the IFC. The assay and sample mixing is automated and occurs in the BioMark HD.

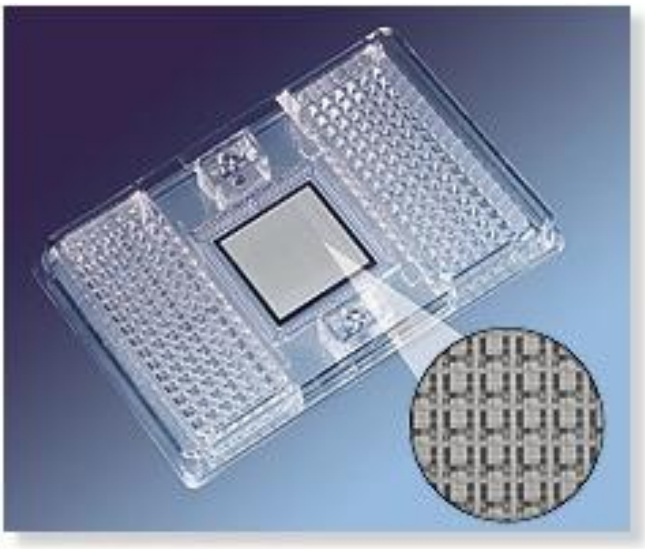

Figure 3. Integrated Fluid Circuit. On the left are the inlets for the tagged assays and on the right the inlets for the samples. The microfluidic channel and valve network is in the IFC's center. Figure from "Fluidigm Genotyping User Guide, SNPtype Assays for SNP Genotyping on the Dynamic Array IFCs, PN 68000098 Rev J1".

4.8.1 Making Primers for the SNPtype Genotyping Assay

The primers for the genotyping process were designed by Fluidigm's D3™ Assay Design. The manual "D3™ Assay Design, PN 100-6812 REV. A2" was used for making the allele-specific targets. The target sequences were given to Fluidigm in the form "80 bp + SNV + 80 bp", see **Appendix 7. Targets for Primers.** The finished primers included tags. Universal probes were used.

4.8.2 Samples with Varying Fractions of DNA for qPCR

The same samples that were used for sequencing were also used for the qPCR reactions. The samples contained 60 ng of DNA in a volume of 2.5 µl, according to the requirements of the protocol, see **Table 4**. For an example of the calculations refer to **Appendix 5**.

4.8.3 Preparing SNPtype Assay Mixes and Sample Mixes

Assay Mixes, which included SNPtype Assay Allele-Specific Primers (ASP) 1 and 2, were mixed with DNA Suspension Buffer for all of the 15 samples. A separate Assay Pre-Mix was made with 2X Assay Loading Reagent and PCR-certified water. These two mixes were combined according to the protocol to form a 10X Assay Mix.

Table 4. Cell line DNA mixtures for qPCR. The ratio of LNCaP and MDA-MB-415 DNA in the samples. Samples 10-15 are internal replicates.

Sample	LNCaP-%	Amount of DNA (ng)	MDA-MB-415-%	Amount of DNA (ng)
1	0	0.0	100	60.0
2	10	6.0	90	54.0
3	20	12.0	80	48.0
4	30	18.0	70	42.0
5	50	30.0	50	30.0
6	70	42.0	30	18.0
7	80	48.0	20	12.0
8	90	54.0	10	6.0
9	100	60.0	0	0.0
10	0 replicate	0.0	100 replicate	60.0
11	10 replicate	6.0	90 replicate	54.0
12	20 replicate	12.0	80 replicate	48.0
13	80 replicate	48.0	20 replicate	12.0
14	90 replicate	54.0	10 replicate	6.0
15	100 replicate	60.0	0 replicate	0.0

A Sample Pre-Mix was made according to the protocol with Biotium 2X Fast Probe Master Mix, SNPtype 20X Sample Loading Reagent, SNPtype Reagent, ROX and PCR-certified water. The Sample Pre-Mix was added to 2.5 µl of each DNA sample according to the protocol to form the Sample Mix.

4.8.4. Priming and Loading the Dynamic Array IFC

The 96.96 Dynamic Array IFC was placed into Fluidigm's IFC Controller HX for priming. The "Prime (138x)" script was run. After priming, 10X Assay Mix was dispensed in 4 µl aliquots with a multichannel pipette into the assay inlets on the IFC. 5 µl aliquots of each Sample Mix were dispensed on the IFC's sample inlets. Loading of the assays and samples into the IFC was done by placing the IFC into the IFC Controller HX and by selecting the "Load Mix (138x)" script.

4.8.5. Thermal Cycling Protocol

Once the assays and samples were loaded on the IFC, it was removed and transferred to Fluidigm's BioMark HD FC1 Cycler for PCR. The protocol chosen was "Thermal cycling protocol SNPtype 96x96 bv1", see **Appendix 8. Table 1**. The probes for the PCR reaction were selected: SNPtype-FAM and SNPtype-HEX. The following settings were also selected: genotyping application, ROX passive reference and auto exposure for the camera.

4.9 Analysis of PCR data

Fluidigm's SNP Genotyping Analysis software was used to analyze the qPCR run data. The genotyping protocol was followed for setting up sample and assay information. During the setup of assay information, the primers were set as follows: ASP-1 non-mutated primer tagged with FAM (X-axis) and ASP-2 SNP primer tagged with HEX (Y-axis). The default confidence threshold of 65 was used. The data normalization method chosen was SNPtype normalization. The computer software compared the relative fluorescence of the tagged samples. After the initial analysis, a second round of analysis was performed, in which the confidence threshold was decreased to 50 in order to make SNV detection more sensitive.

5. Results

5.1 Agilent HaloPlex Target Enrichment System

The enrichment system had several phases with results.

5.1.1 Validation of Restriction Digestion

The electrophoretic run of the validation of restriction digestion is summarized in the Figure below, see **Figure 4**. Due to technical problems, the ladder used for the analysis did not show up in the correct scale. The bands visible in the ECD restriction samples were almost as they should be, when compared to the example of a successful electrophoresis run in the protocol; 3 bands of varying size, which are the results of digestion by two restriction enzymes. It was not possible to determine the correct size of the bands, because of the erroneous ladder scaling. Samples 4 and 6 were not accurate because the higher marker used was not found. Sample 4 had the correct number of bands, but sample 6 did not. Lane 9 contained the same marker as lanes 10 and 11, but it should have contained undigested ECD.

5.1.2 Validation of Enriched Target DNA

The amplicon validation results from the Bioanalyzer can be seen in **Appendix 9. Figures 1-25**. Validations of samples 1, 4, 7, 12, 13 and 14 looked the best; see **Appendix 9. Figures 1**, **8, 12, 17, 19** and **21** respectively. Correct amplicon size was approximately between 225 and 525 bp. Otherwise the electropherograms did not have clear peaks in the correct range, indicating absence of material in the samples. Most samples had prominent peaks at 125 bp. These latter peaks indicated the formation of adapter-dimer.

See **Appendix 10. Table 1**, for the measured concentration for each sample. The values for those samples which were most reliable were accepted for further use, see **Appendix 10 Table 2**.

The amplicon validation results from LabChip GXI can be seen in **Appendix 11. Figures 1-7**. As with the Bioanalyzer results, the electropherograms produced by LabChip did not have clear peaks in the range of 225-525 bp. Each sample also had a peak at 125-150 bp, indicating an adapter-dimer. See **Appendix 12. Table 1** for the measured concentrations and molarities of the samples validated by LabChip.

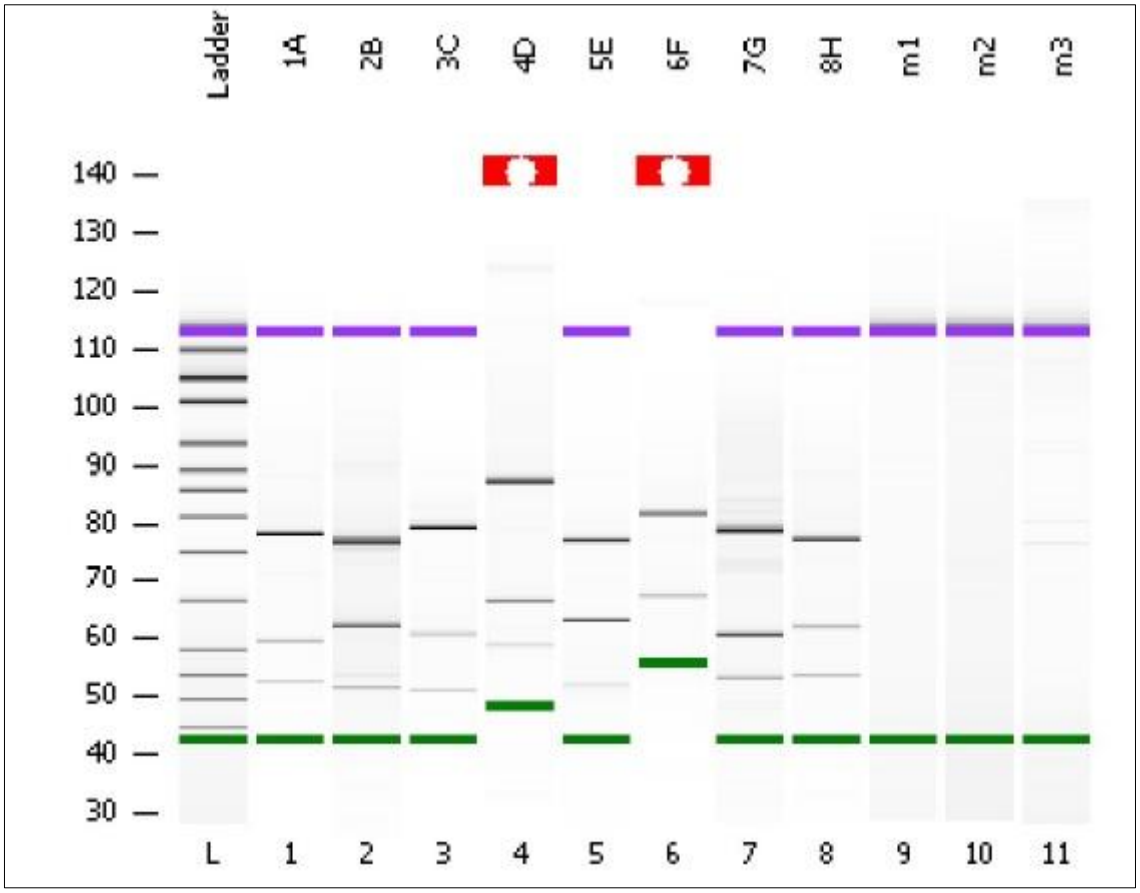

Figure 4. Enrichment control DNA electrophoresis results. The lane designated as L contained the 50 bp DNA ladder. Lanes 1-8 contained the 8 ECD digestion reactions A-H. Lanes 9-11 contained marker. The green bands represented the lower marker and the purple bands the higher marker. The higher markers for reactions D (lane 4) and F (lane 6) were not found.

5.2 Run Data from MiSeq Benchtop Sequencer

The run data sent to BaseSpace was used for assessing the performance of the sequencing run. **Appendix 13. Table 1** shows that 19, 141, 716.0 reads were produced during sequencing. 80% of the reads aligned with the samples. The number of aligned reads for each sample can also be seen in **Appendix 13. Figure 1**. According to the figure, sample 4 (30% LNCaP DNA and 70% MDA-MB-415 DNA) had the largest amount of reads align with it, 18.8% of all reads. Sample 16 (Control DNA) had the smallest amount of reads align with it, only 2.1%.

The QScore distribution plot shows the quality score distribution of the bases, see **Appendix 13. Figure 2**. 74.0% of the bases have a quality score of over Q30, meaning that 74% of bases have a 0.1% chance of error. The coverage (C) of the sequencing run was calculated to be 1 797. The calculations are seen in **Appendix 13**. Theoretically the coverage was 1 797, but some bases were covered more and some less.

5.3 Sequencing Results

After the computational manipulation of the raw sequencing data, the data from DNA Pool 1 was viewed with IGV. See **Appendix 14. Figure 1** for the sequencing call map. An example of viewing a sample and one of its SNVs with IGV, see **Appendix 14. Figure 2**.

Out of the 10 SNVs that were searched for, 4 were found: AR and PIK3R1 (from LNCaP), and MAP2K4 and CSF1R (from MDA-MB-415). Each SNV was found in the smallest fraction containing that particular cell line, which was a fraction of 10%. A 10% fraction contained 22.5 ng of DNA. The SNVs were also found in all consecutive fractions up to the 100% fraction, which contained 225.0 ng of DNA.

5.4 qPCR Results

See **Appendix 14. Figure 3**, for the call map of the SNV genotyping with Fluidigm. 8 out of the 10 SNVs were found: AR, PIK3R1 and SMO (from LNCaP) and ALK, BRAF, MAP2K4, CSF1R and ERBB4 (from MDA-MB-415). The smallest fraction with a SNV detected was 20%. A 20% fraction contained 12.0 ng of DNA.

Not all SNVs were detected systematically from 20% onward. Detection became more regular at 50% for the SNVs in the MDA-MB-415 cell line. A 50% fraction was equivalent to 30.0 ng of DNA. SNVs in the LNCaP cell line were not detected regularly until fractions with 70% of LNCaP DNA. A 70% fraction was equivalent to 42.0 ng of DNA. All of the above values were from analysis with the default confidence threshold of 65.

When the confidence threshold was decreased to 50, making detection more sensitive, the call information changed slightly. Some No Calls changed to SNVs, XX (homozygous non-mutated) or YY (homozygous mutated). These changes were distributed evenly in the call map, making no difference in the final detection sensitivity of the SNVs.

6. Discussion

The results from sequencing and qPCR were very different, see **Appendix 14**. Illumina's MiSeq was able to detect four SNVs out of ten. If a SNV was detected, it was detected in all the fractions with the cell line containing the mutation. The smallest fraction of DNA detected contained 22.5 ng of DNA. However, six remaining SNVs were not found in any of the samples. Fluidigm's BioMark HD was able to find a total of eight SNVs from the samples, but not systematically. The smallest fraction of DNA detected by Fluidigm contained 12.0 ng.

The results from sequencing with MiSeq are not encouraging. All ten SNVs searched for in this study are known variants and should have been found to at least some degree in the samples. Only four SNVs were detected from the following genes AR and PIK3R1 from the LNCaP cell line, and MAP2K4 and CSF1R from the MDA-MB-415 cell line, see **Figure 1** of **Appendix 14**. These four SNVs were visible in all samples, indicating that DNA from both cell lines was present during sequencing. The SNVs that are absent are systematically absent from all sequenced samples and produced no reads. This suggests that something happened before sequencing, during the enrichment and targeting phases of the HaloPlex Target Enrichment System protocol.

The HaloPlex Target Enrichment System protocol, which was used for making a DNA library, included many phases in which the samples were treated *en masse*, meaning they were handled as a group. No individual treatment was performed, such as adding primers for each target. This means that if a phase had gone wrong during the protocol, it would be visible in all of the samples. If the hybridization of the targets had been problematic in general, none of the targets would have hybridized or hybridization would have randomly occurred, but such is not the case. Four SNVs were systematically hybridized. The same logic goes for the targeting. A problem in targeting would have also been visible as no targeting occurring or something being targeted in random.

The probable absence of DNA from the samples is supported by the electropherograms of the samples, see **Appendices 9** and **11**. According to the electropherograms, samples 1, 4, 12, 13 and 14 contained some DNA, as seen in the amplicon peaks visible, unlike most of the other samples. The high peak at 125-150 bp, which is visible for almost all samples, is not an amplicon peak, but rather a peak indicating adapter dimer.

All of this suggests that the missing SNVs, were not targeted by the Agilent's HaloPlex Cancer Research Panel. According to Agilent and their enrichment protocol "HaloPlex probes hybridize selectively to fragments from target regions." Before beginning the enrichment protocol, an inquiry was made to Agilent concerning the content of the target regions: would the specific SNVs in this study be targeted? The answer given by Agilent was that the most common variants in cancer could be detected by the kit. When more information about the targeting of the specific SNVs was asked for, no information could be given because it would have been commercial secret infringement. It is possible that the variants searched for in this study, were not included as possible targets in the panel, if they were not part of the list of most common variants.

Another thought for why some variants were not visible, is that perhaps some genetic variation occurred during cell culture. However this idea does not hold, because the same absence of variants should have been visible in the qPCR genotyping results as well, since the method was used after sequencing.

It is not possible that the washing away of samples during the targeting protocol is a reason for the missing variants. If some DNA were washed away, it would be visible as entire samples missing all sequencing reads and variants, which is not the case. The results from genotyping with Fluidigm's BioMark HD were better.

The qPCR with BioMark HD was able to detect eight out of ten SNVs, see **Figure 3** in **Appendix 14**. The SNVs in ABL1 and NOTCH1 were not detected. It is possible that some pipetting errors occurred during the addition of these two particular assays onto the 96-well plate which was used during genotyping. This would seem likely, because the call map shows that No Call-results were very frequent in all samples, not just a few.

The detection of the found SNVs was not systematic throughout the different fractions. Even though 12.0 ng of DNA were detected in 20% DNA fractions, the detection occurred infrequently. Systematic detection was made only from samples with larger fractions with more than 30.0 ng of DNA.

Both genotyping methods were performed a single time, except for the actual sequencing of the DNA pool. This was because of a set deadline for the study. If there had not been such a

deadline, it would have been possible to optimize both methods, resulting in potentially better results.

When comparing NGS and PCR-based methods, the former is able to produce large quantities of data compared to the latter (5). The large quantity of reads for each target area in sequencing makes sequencing somewhat more reliable. This is because a SNV will show up frequently in the reads. All reads can be seen individually and they can be compared, which is something that cannot be done with the qPCR genotyping method. The qPCR method runs each assay a single time, unless there are internal replicates on the same IFC, and gives a single call result. The only comparison that can be made is comparing the automatic call to the final call made by the analysis software. The two are not necessarily the same. If they are not the same, the final call could change if the confidence threshold is decreased. In this way, an assay can have an increased rate confidence, but it cannot be compared to the confidence which thousands of reads bring from NGS sequencing.

The two SNV genotyping methods cannot be compared properly based on the results of this study, leaving this study inconclusive. Nothing can be said about the detection sensitivity of Illumina's MiSeq because the steps preceding sequencing were erroneous in some part. Fluidigm performed better, but even it did not give good results. With the results at hand however, it must be said that qPCR with Fluidigm's BioMark HD seems to be more sensitive and reliable than NGS with Illumina's MiSeq.

7. Conclusions

Fluidigm's genotyping with qPCR is more sensitive than sequencing with Illumina's MiSeq in the detection SNVs. Fluidigm's BioMark HD is able to detect SNVs from 20% DNA fractions, which represents 12.0 ng of DNA, but infrequently. More reliable detection occurs in DNA fractions of 70-80%, representing 42.0-48.0ng of DNA. MiSeq is able to detect SNVs from samples with a 10% fraction of DNA, which represents 22.5 ng of DNA, but the method only detected four out of ten SNVs.

8. References

(1) Hewitt SM, Lewis FA, Cao Y, Conrad RC, Cronin M, Danenberg KD, et al. Tissue handling and specimen preparation in surgical pathology: issues concerning the recovery of nucleic acids from formalin-fixed, paraffin-embedded tissue. Arch Pathol Lab Med 2008;132: 1929-1935.

(2) Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nature Reviews Genetics 2010;11: 685-696.

(3) Huijsmans CJ, Damen J, van der Linden JC, Savelkoul PH, Hermans MH. Comparative analysis of four methods to extract DNA from paraffin-embedded tissues: effect on downstream molecular applications. BMC Research Notes 2010;3: 239.

(4) Frankel A. Formalin fixation in the '-omics' era: a primer for the surgeon-scientist. ANZ J Surg 2012;82: 395-402.

(5) Wood HM, Belvedere O, Conway C, Daly C, Chalkley R, Bickerdike M, et al. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. Nucleic Acids Res 2010;38: e151.

(6) Gilbert MTP, Haselkorn T, Bunce M, Sanchez JJ, Lucas SB, Jewell LD, et al. The isolation of nucleic acids from fixed, paraffin-embedded tissues-which methods are useful when?. PLoS ONE [Electronic Resource] 2007;2: e537.

(7) Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS. A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet 2005;37: 549-554.

(8) Metzker ML. Sequencing technologies - the next generation. Nature Reviews Genetics 2010;11: 31-46.

(9) Hoelsch K, Lenggeler I, Pfannes W, Knabe H, Klein H, Woelpl A. Routine HLA-B genotyping with PCR-sequence-specific oligonucleotides detects a B*52 variant (B*5206). Tissue Antigens 2005;65: 488-492.

(10) Swango KL, Hudlow WR, Timken MD, Buoncristiani MR. Developmental validation of a multiplex qPCR assay for assessing the quantity and quality of nuclear DNA in forensic samples. Forensic Sci Int 2007;170: 35-45.

(11) Rossetti LC, Radic CP, Abelleyro MM, Larripa IB, De Brasi CD. Eighteen years of molecular genotyping the hemophilia inversion hotspot: from southern blot to inverse shifting-PCR. International Journal of Molecular Sciences 2011;12: 7271-7285.

(12) Jung R, Soondrum K, Neumaier M. Quantitative PCR. Clinical Chemistry & Laboratory Medicine 2000;38: 833-836.

(13) Beltran H, Yelensky R, Frampton GM, Park K, Downing SR, MacDonald TY, et al. Targeted next-generation sequencing of advanced prostate cancer identifies potential therapeutic targets and disease heterogeneity. Eur Urol 2013;63: 920-926.

(14) Dong J. Prevalent mutations in prostate cancer. J Cell Biochem 2006;97: 433-447.

(15) Kirby R, Challacombe B, Dasgupta P, Fitzpatrick JM. Prostate cancer treatment: the times they are a' changin'. BJU Int 2012;110: 1408-1411.

(16) Dahabreh IJ, Chung M, Balk EM, Yu WW, Mathew P, Lau J, et al. Active Surveillance in Men With Localized Prostate Cancer: A Systematic Review. Ann Intern Med 2012;156: 582-590.

(17) Akduman B, Crawford ED. Treatment of localized prostate cancer. Reviews in Urology 2006;8: S15-21.

(18) Nazario ACP, Facina G, Filassi JR. Breast cancer: news in diagnosis and treatment. Rev Assoc Med Bras 2015;61: 543-552.

(19) Imyanitov EN, Suspitsin EN, Grigoriev MY, Togo AV, Kuligina ES, Belogubova EV, et al. Concordance of allelic imbalance profiles in synchronous and metachronous bilateral breast carcinomas. International Journal of Cancer 2002;100: 557-564.

(20) Kim JY, Cho N, Koo HR, Yi A, Kim WH, Lee SH, et al. Unilateral Breast Cancer: Screening of Contralateral Breast by Using Preoperative MR Imaging Reduces Incidence of Metachronous Cancer. Radiology 2013;267: 57-66.

(21) Hampl M, Hampl JA, Reiss G, Schackert G, Saeger HD, Schackert HK. Loss of heterozygosity accumulation in primary breast carcinomas and additionally in corresponding distant metastases is associated with poor outcome. Clinical Cancer Research 1999;5: 1417-1425.

(22) Lichy JH, Dalbegue F, Zavar M, Washington C, Tsai MM, Sheng ZM, et al. Genetic heterogeneity in ductal carcinoma of the breast. Laboratory Investigation 2000;80: 291-301.

(23) Alkner S, Bendahl P, Ferno M, Manjer J, Ryden L. Prediction of outcome after diagnosis of metachronous contralateral breast cancer. BMC Cancer 2011;11: 114.

(24) Kong K, Kendall C, Stone N, Notingher I. Raman spectroscopy for medical diagnostics — From in-vitro biofluid assays to in-vivo cancer detection. Adv Drug Deliv Rev 2015;89: 121-134.

(25) Metzker ML. Emerging technologies in DNA sequencing. Genome Res 2005;15: 1767-1776.

9. Appendices

Appendix 1. HaloPlex Targeted Genes

Table 1. Targeted genes. Genes in cell lines LNCaP and MDA-MB-415, which were targeted by HaloPlex Cancer Panel Kit. Targeted SNVs are shown with their coordinates, along with affected amino acid and mutation type.

Prostate cancer cell line LNCaP with targeted SNVs	**Breast cancer cell line MDA-MB-415 with targeted SNVs**
ABL1 Coordinates: 9:133,759,986-133,759,986 N770S, A>G substitution	ALK Coordinates: 2:29,940,524-29,940,524 P236R, G>C substitution
AR Coordinates: X:66,943,552-66,943,552 T878A, A>G substitution	BRAF Coordinates: 7:140,549,931-140,549,931 P74A, G>C substitution
NOTCH1 Coordinates: 9:139,413,143-139,413,143 S333S, C<T substitution	MAP2K4 Coordinates: 17:12,028,636-12,028,636 S291*, C>A substitution (nonsense)
PIK3R1 Coordinates: 5:67,592,099-67,592,099 R639*, C>T substitution (nonsense)	CSF1R Coordinates: 5:149,433,643-149,433,643 Q970*, G>A substitution (nonsense)
SMO Coordinates: 7:128,845,520-128,845,520 C273R, T>C substitution	ERBB4 Coordinates: 2:212,522,511-212,522,511 W638*, C>T substitution (nonsense)

Appendix 2. List of Reagents and Kits Used

Cell culture

RPMI 1640 cell culture medium (Lonza, Basel, Switzerland)

Leibovitz's L-15 medium for cell culture (Sigma-Aldrich, St. Louis, USA)

DNA extraction

Qiagen QIAamp DNA Mini kit (Qiagen, Hilden, Germany)

Measuring of DNA concentrations with Qubit

dsDNA BR reagent (ThermoFisher Scientific, Waltham, USA)

dsDNA BR Buffer (ThermoFisher Scientific, Waltham, USA)

Standard 1 (ThermoFisher Scientific, Waltham, USA)

Standard 2 (ThermoFisher Scientific, Waltham, USA)

Agilent HaloPlex Target Enrichment System

Validation of Restriction Digestions

2100 Bioanalyzer Platform High Sensitivity DNA Kit (Agilent Technologies, Santa Clara, USA)

Capturing the Target DNA

Agencourt AMPure XP Kit, 60 ml (Beckman Coulter Genomics, Danvers, USA)

2M Acetic acid (Sigma-Aldrich, St. Louis, USA)

10 M NaOH, molecular biology grade (Sigma-Aldrich, St. Louis, USA)

PCR Master Mix

Herculase II Fusion Enzyme with dNTPs (100 mM; 25 mM for each nucleotide, 200 reactions), (Agilent Technologies, Santa Clara, USA)

2M Acetic acid (Sigma-Aldrich, St. Louis, USA)

Purifying of the Target Libraries

100% Ethanol, molecular biology grade (Sigma-Aldrich, St. Louis, USA)

10 mM Tris-acetate, pH 8,0 (Merck, Darmstadt, Germany)

Preparing Libraries for Sequencing on MiSeq

MiSeq v2 Reagent Kit (Illumina, San Diego, USA)

MiSeq v2 Reagent Kit 300 cycles PE-Box (Illumina, San Diego, USA)

10 M NaOH, molecular biology grade (Sigma-Aldrich, St. Louis, USA)

Tris-Cl 10 mM, pH 8.5 with 0.1% Tween 20 (Tris-Cl from Merck, Darmstadt, Germany Tween 20 from Sigma-Aldrich, St. Louis, USA)

Sequencing with MiSeq

MiSeq v2 Reagent Kit (Illumina, San Diego, USA)

MiSeq v2 Reagent Kit 300 cycles PE-Box (Illumina, San Diego, USA)

PhiX Control (Illumina, San Diego, USA)

qPCR with BioMark HD

Biotium Fast Probe Master Mix (Biotium, Hayward, USA)

Qiagen 2x Multiplex PCR Master Mix (Qiagen, Hilden, Germany)

DNA Suspension Buffer, 10 mM Tris, pH 8.0, 0.1 mM EDTA (TEKnova, Helsinki, Finland)

SNPtype Genotyping Reagent Kit 96.96 (Fluidigm, South San Francisco, USA)

SNPtype Assay Allele-Specific Primers (ASP) Plate, 100 µM ASP1/100 µM ASP2 (Fluidigm, South San Francisco, USA)

SNPtype Assay Locus-Specific Primer (LSP) Plate, 100 µM (Fluidigm, South San Francisco, USA)

50X ROX (ThermoFisher Scientific, Waltham, USA)

Appendix 3. Workflow of Sample Preparation for HaloPlex Target-Enrichment Protocol

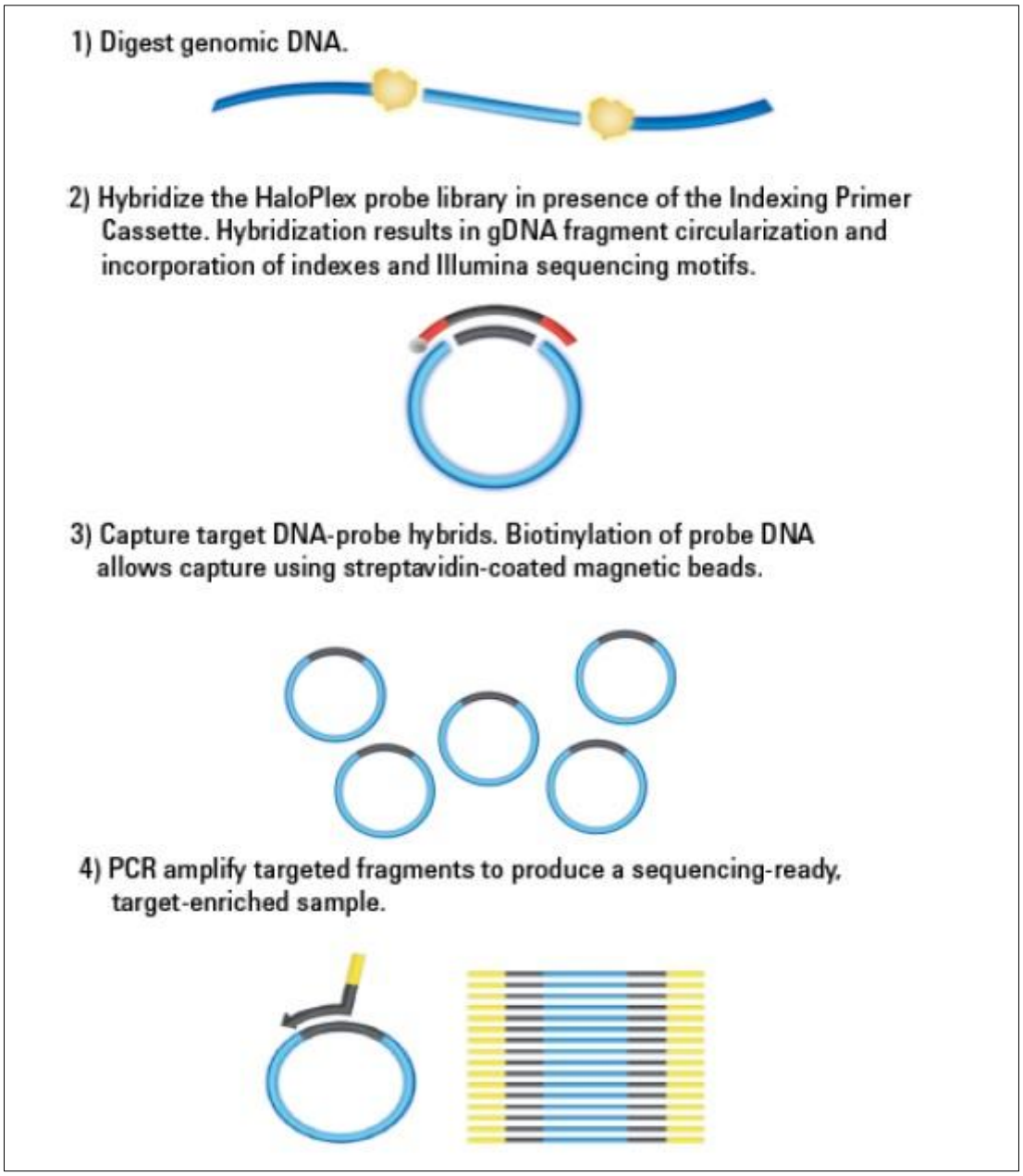

Figure 1. Workflow of sample preparation for HaloPlex target-enrichment sequencing. An indexed library was made for all the samples with the use of Illumina paired-end sequencing motifs. gDNA fragment circularization occurred as the result of hybridization between the gDNA and Illumina's motifs. The biotinylated probe DNA was captured with the help of streptavidin-coated magnetic beads, after which amplification of DNA was performed. The target-enriched samples were then ready for sequencing. Figure from HaloPlex Target Enrichment System Protocol for Illumina Sequencing, Version D.5, May 2013.

Appendix 4. Dilution Series for NaOH

<u>10 M NaOH to 1 M NaOH</u>

10 mol/l x X = 1 mol/l x 1 ml

X = $\frac{1\ x\ 1}{10} ml$

X = 0.1 ml = 100 µl

100 µl 10 M NaOH + 900 µl H_2O

<u>1 M NaOH to 100 mM NaOH</u>

1 mol/l x X = 0.1 mol/l x 1 ml

X = $\frac{0.1\ x\ 1}{1} ml$

X = 0.1 ml = 100 µl

100 µl 1 M NaOH + 900 µl H_2O

<u>100 mM NaOH to 50 mM NaOH</u>

100 mmol/l x X = 50 mmol/l x 1 ml

X = $\frac{50\ x\ 1}{100} ml$

X = 0.5 ml = 500 µl

500 µl 100 mM NaOH + 500 µl

Appendix 5. Calculations

Sample preparation for sequencing

Example of calculations for sample 2 preparation, with 10% LNCaP DNA and 90% MDA-MB-415 DNA.

22.5 ng of LNCaP DNA and 202.5 ng DNA were required for sample 2.

<u>LNCaP</u>

C (LNCaP) = 49 000 ng/ml

$$\frac{49\,000\,ng}{22.5\,ng} = \frac{1\,ml}{x}$$

X = 0.000459 ml = 0.459 µl = 0.5 µl contained 22.5 ng of LNCaP DNA

<u>MDA-MB-415</u>

C (MDA-MB-415) = 8 940 ng/ml

$$\frac{8\,940\,ng}{202.5\,ng} = \frac{1\,ml}{x}$$

X = 0.0226 ml = 22.6 µl contained 202.5 ng of MDA-MB-415 DNA

<u>DNA dilution</u>

$C_1V_1 = C_2V_2$

V_1 = 0.5 µl LNCaP + 22.6 µl MDA-MB-415 = 23.5 µl

$C_1 = \frac{225.0\,ng}{23.5\,\mu l}$ = 9.5 ng/µl

C_2 = 5.0 ng/µl (Defined by HaloPlex Target Enrichment System protocol)

$V_2 = \frac{9.5\frac{ng}{\mu l} x\ 23.5\,\mu l}{5.0\,ng/\mu l}$ = 44.65 µl end volume

Addition of water:

44.65 µl – 23.5 µl = 21.0 µl of water was added

Pooling of DNA Samples

Example of calculations for DNA Pool 1 with sample 1 with 0% of LNCaP DNA and 100% MDA-MB-415 DNA:

$C_1V_1 = C_2V_2$

C_2 = 2 nM (according to Illumina sequencing protocol)

V_2 = 240 µl (the final volume of DNA Pool 1)

$V_1 = \frac{C2V2}{C1} = \frac{2\ nM\ x\ 240\ \mu l}{38.9\ nM} = 12.33\ \mu l = 12.3\ \mu l$

The total volume of DNA Pool 1 samples added together was 170.7 µl. Distilled water was added up to the final volume of 240 µl.

Sample preparation for qPCR

Example of calculations for sample 2 with 10% LNCaP DNA and 90% of MDA-MB-415 DNA:

$C_1V_1 = C_2V_2$

C_2 = 60 ng/2.5 µl = 24 ng/µl

C_2 (LNCaP) = 0.10 x 24 ng/µl = 2.4 ng/µl

C_2 (MDA-MB-415) = 0.90 x 24 ng/µl = 21.6 ng/µl

C_1 (LNCaP) = 49.0 ng/µl

$V_1 = \frac{C2V2}{C1} = \frac{2.4\frac{ng}{\mu l} x\ 2.5\ \mu l}{49.0\ ng/\mu l}$ = 0.122 µl LNCaP DNA

C_1 (MDA-MB-415) = 30.0 ng/µl

$V_1 = \frac{C2V2}{C1} = \frac{21.6\frac{ng}{\mu l} x\ 2.5\ \mu l}{30.0\ ng/\mu l}$ = 1.8 µl MDA-MB-415 DNA

0.12 µl LNCaP DNA + 1.8 µl MDA-MB-415 DNA + 0.58 µl H_2O = 2.5 µl sample volume

For pipetting ease the volumes above were multiplied by 3.5:

0.42 µl LNCaP DNA + 6.3 µl MDA-MB-415 DNA + 2.03 µl H_2O

Appendix 6. Scripts Used During Computational Modification of Sequencing Data.

Trimming

The following scripts were used during trimming:

Example of trimming with sample 4 (30% LNCaP DNA and 70% MDA-MB-415 DNA):

Trimming of 3' end:

-3 50 -1 LNCaP-30-MDA-70.fastq

Trimming of 5' end:

-5 30 -1 LNCaP-30-MDA-70.fastq

fasta trim LNCaP-30-MDA-70.fastq100

SAMtools

The following scripts were part of a loop using different SAMtools commands:

Example with X denoting a single file representing one sample:

SAMtools View:

```
#|/bin/bash
for X in *.SAM
 do
  samtools view –b –h –S $X > ${X/.sam/.bam}
   done
```

SAMtools Sort:

```
for X in *.bam
 do
  samtools sort $X ${X/.bam/.sorted}
   done
```

SAMtools Index:

```
for X in *.sorted.bam
 do
  samtools index $X
   done
```

Appendix 7. Targets for Primers

LNCaP cell line

Gene: ABL1, coordinate 9:133759986-133759986, A>G substitution

Sequence for coordinates chr9:133759906-133760066
GGGAAGACAGTTTGACTCGTCCACATTTGGAGGGCACAAAAGTGAGAAGCCGGC
TCTGCCTCGGAAGAGGGCAGGGGAGA**[A/G]**CAGGTCTGACCAGGTGACCCGAGG
CACAGTAACGCCTCCCCCCAGGCTGGTGAAAAAGAATGAGGAAGCTGCTGATGA
GG

Gene: AR, coordinate X:66943552-66943552, A>G substitution

Sequence for coordinates chrX:66943472-66943632
AGCAGAGGCCACCTCCTTGTCAACCCTGTTTTTCTCCCTCTTATTGTT
CCCTACAGATTGCGAGAGAGCTGCATCAGTTC**[A/G]**CTTTTGACCTGCTAATCAAG
TCACACATGGTGAGCGTGGACTTTCCGGAAATGATGGCAGAGATCATCTCTGTGC
AAGTG

Gene: NOTCH1, coordinate 9:139413143-139413143, G>A substitution.

Sequence for coordinates chr9:139413063-139413223
TCGCAGTAGAAGGAGGCCACACGGTCATGGCAGGTGGCGCCGTGGA
AGCAGGCGGCGCTGGCACAGTCATCAATGTTCTC**[G/A]**CTGCAGTCCTCACCAGTC
CAGCCGTTGACACACACGCAGTTGTAGCCACCGTGGGTGTTGTGGCAGGTCCCGC
CGTTCTG

Gene: PIK3R1, coordinate 5:67592099-67592099, C>T substitution

Sequence for coordinate chr5:67592019-67592179
ATGATGAAGATTTGCCCCATCATGATGAGAAGACATGGAATGTTGGA
AGCAGCAACCGAAACAAAGCTGAAAACCTGTTG**[C/T]**GAGGGAAGCGAGATGGC

ACTTTTCTTGTCCGGGAGAGCAGTAAACAGGGCTGCTATGCCTGCTCTGTAGTGT
ATGTATCT

Gene SMO, coordinate 7:128845520-128845520, T>C substitution

Sequence for coordinates chr7:128845440-128845600

CTCTCTTCTAGGCCACATTCGTGGCTGACTGGCGGAACTCGAATCGC
TACCCTGCTGTTATTCTCTTCTACGTCAATGCG**[T/C]**GCTTCTTTGTGGGCAGCATT
GGCTGGCTGGCCCAGTTCATGGATGGTGCCCGCCGAGAGATCGTCTGCCGTGCAG
ATGGC

<u>MDA-MB-415 cell line</u>

Gene: ALK, coordinate chr2:29940524-29940524, G>C substitution

Sequence for coordinates chr2:29940444-29940604

CATATCGGCTGCGATGAGACAGGAAAGGGAAGGAGTCTTTCATTATC
CAGGTGAGATTCCATGTAAAATAATCAGGAGAA**[G/C]**GAGATGGCATGTTTGTTG
GTGATTCCAAGGAGCTATGACCTGGACATAAAAATAAAGAAAACACTGATCCAT
GTGCTTGG

Gene: BRAF, coordinate chr7:140549931-140549931, G>C substitution

Sequence for coordinates chr7:140549851-140550011

TTTTTATAAGTTCATTTTTTTTCTTTTCAAAATTACTAGATATGATACT
CAAAAGCTTACCTCCAGATATATTGATGGTG**[G/C]**ATTATGCTCCCCACCAAATTT
GTCCAATAGGGCCTCTATATGTTCCTGTGTCAACTTAATCATTTGTTTGATATTCC
ACA

Gene: MAP2K4, coordinate chr17:12028636-12028636, C>A substitution

Sequence for coordinates chr17:12028556-12028716

GCCTATTCCTTGAGTGTAAGGCAATTAATAACTTACACTTGTCTTTAT
GTTCCAGCCTGAAAGAATAGACCCAAGCGCAT**[C/A]**ACGACAAGGATATGATGTC

CGCTCTGATGTCTGGAGTTTGGGGATCACATTGGTATGTTTATGCTGATTCAACCT
TGCCA

Gene: CSF1R, coordinate chr5:149433643-149433643, G>A substitution

Sequence for coordinates chr5:149433563-149433723

GTGTCGCCCCATCCATGGAGGAGTTGAAGTTTGGAGGAGGGGAGAG
TGGTACTCCCTGTCGTCAACTCCTCAGCAGAACT**[G/A]**ATAGTTGTTGGGCTGCAG
CAAGGGCTGGGCGATATCCCCTTGCTCGCAGCAGGTCAGGTGCTCACTAGAGCTC
TCCTCCT

Gene: ERBB4, coordinate chr2:212522511-212522511, C>T substitution

Sequence for coordinates chr2:212522431-212522591

AACTAGGAAAGGATTTGAGCGACAAAATGGAAACATGGTAGATGTT
ACCTAGCATGTTGTGGTAAAGTGGAATGGCCCGT**[C/T]**CATGGGTAGTAAATGCA
GTCATGACTAGTGGGACCGTTACACCTGCAGGCAATTACAGAACAGAAAACATC
ATTCTCCAT

Appendix 8. PCR Program for Genotyping

The genotyping program used by Fluidigm's BioMark HD had the phases specified in **Table 1.**

Table 1. PCR program. The SNP-genotyping PCR program used during SNP genotyping.

Thermal Cycling Conditions	**Cycles**	**Temperature**	**Time**
Thermal Mix	1 cycle of:	70 °C 25°C	30 min 10 min
Hot Start	1 cycle of:	95 °C	5 min
Touchdown (from 64.0-61.0 °C, dropping 1 °C per cycle)	1 cycle of: 1 cycle of: 1 cycle of: 1 cycle of:	95 °C 64 °C 72 °C 95 °C 63 °C 72 °C 95 °C 62 °C 72 °C 95 °C 61 °C 72 °C	15 sec 45 sec 15 sec 15 sec 45 sec 15 sec 15 sec 45 sec 15 sec 15 sec 45 sec 15 sec
Additional PCR cycles	34 cycles of:	95 °C 60 °C 72 °C	15 sec 45 sec 15 sec
Cool	1 cycle of:	25 °C	10 sec

Appendix 9. Validation of Amplicon Size with 2100 Bioanalyzer

Figures 1-25 show all the electropherograms and electrophoresis run results for samples 1 to 16. Any additional runs are also shown directly after the initial runs. The peak values are indicated for each electropherogram. Correct amplicon size was approximately between 225 and 525 bp. The lower marker (green) is at 35 bp and the higher marker (purple) is at 10 380 bp. Necessary dilutions were done in a 1:10 ratio.

Sample 1

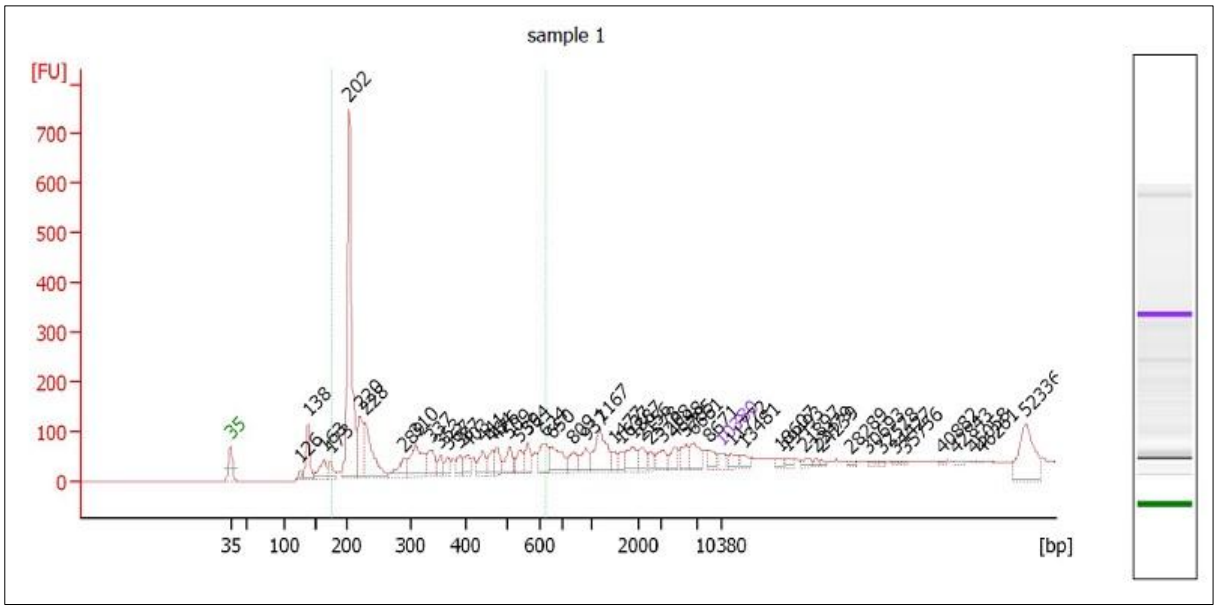

Figure 1. Sample 1 electropherogram. Peak Value: 202 bp. The electrophoretic run indicates that most of the product is the amplicon of 202 bp.

Sample 2

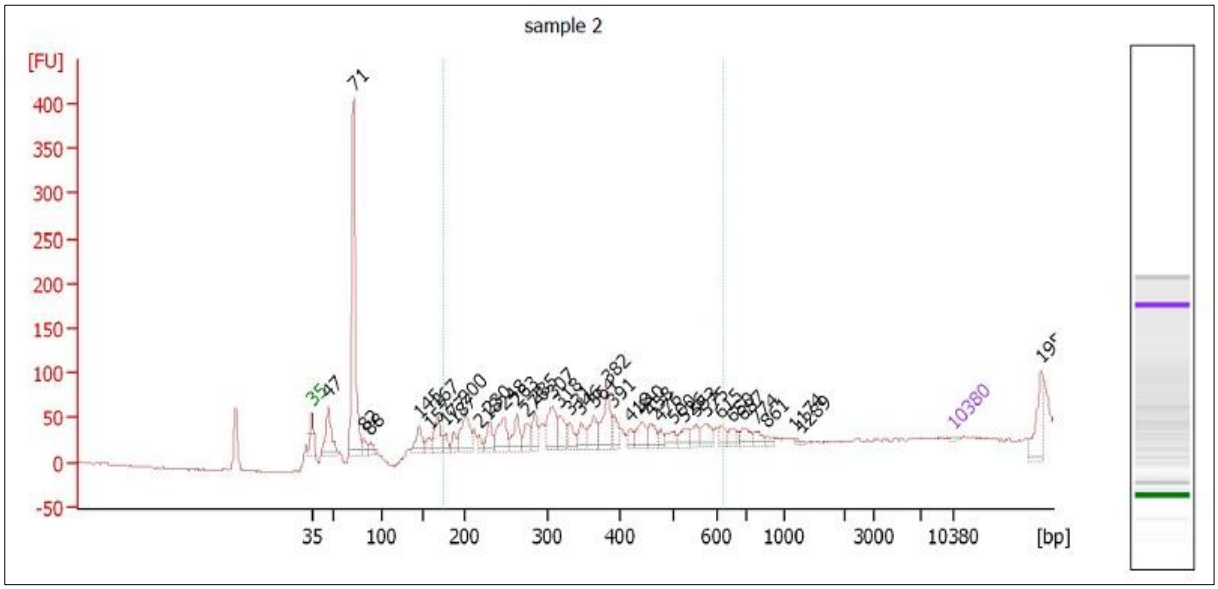

Figure 2. Sample 2 electropherogram. Peak Value: 307 bp. The peak at 71 bp is adapter-dimer. The baseline of the run was not normal and the lower marker had to be manually set at 35 bp. No clear band is visible.

Sample 2 diluted

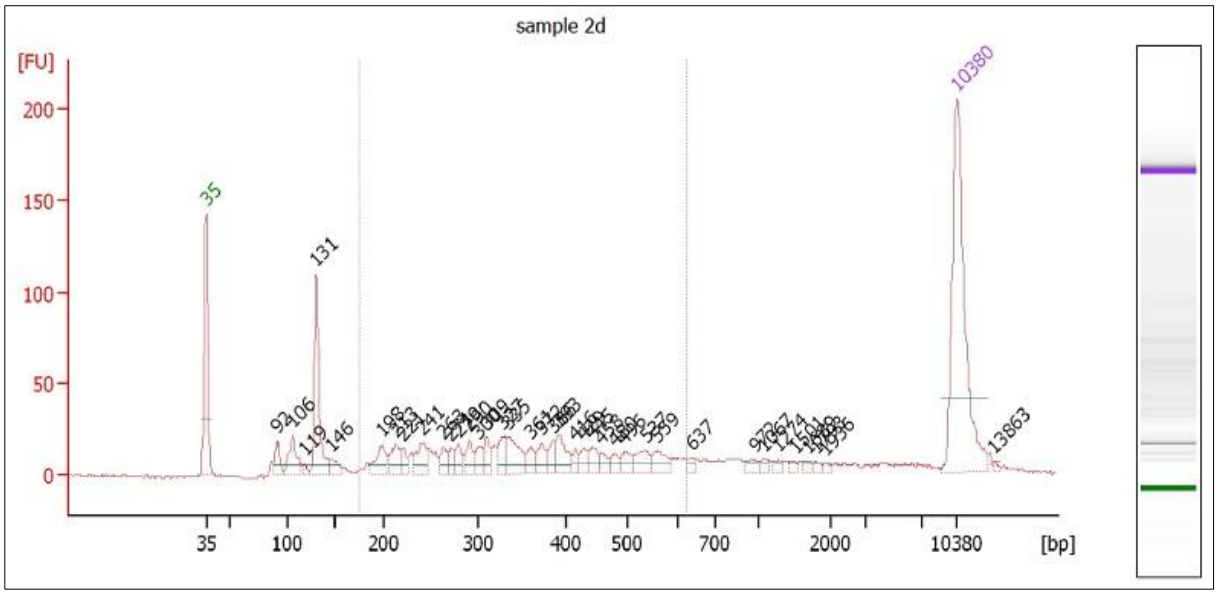

Figure 3. Sample 2 diluted electropherogram. Peak value: 335 bp. The peak at 131 bp is adapter-dimer. The amplicon is not clear enough in the electrophoretic run.

Sample 2 undiluted

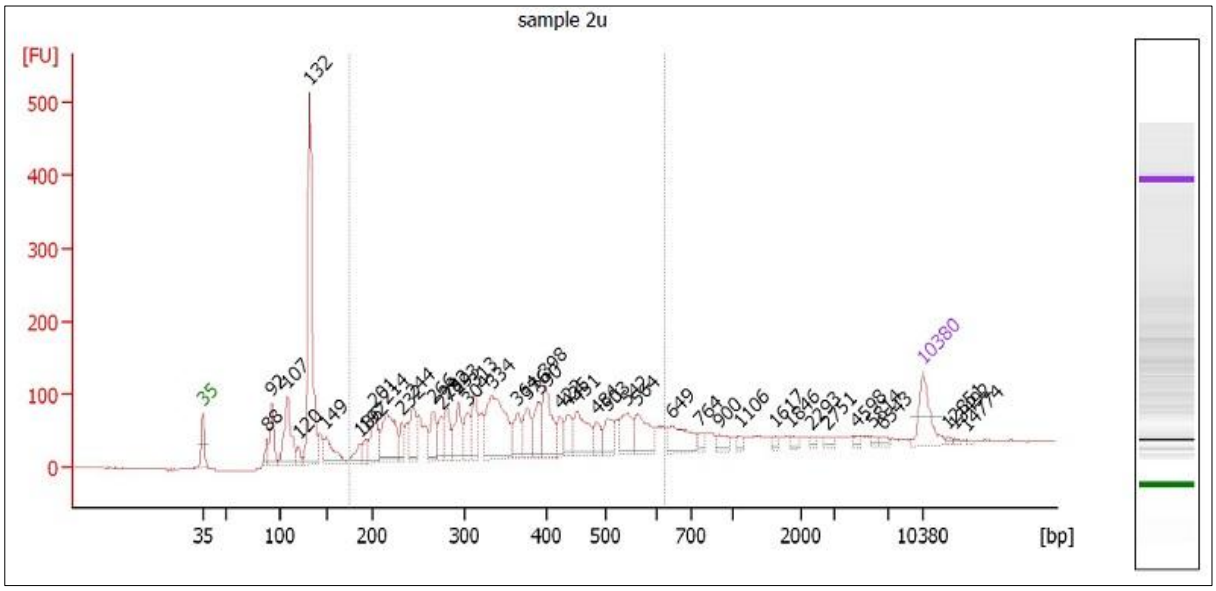

Figure 4. Sample 2 undiluted electropherogram. Peak value: 334 bp. The peak at 132 bp is adapter-dimer. The electrophoretic run shows a single dark band, but it is not the amplicon, but the adapter-dimer.

Sample 3

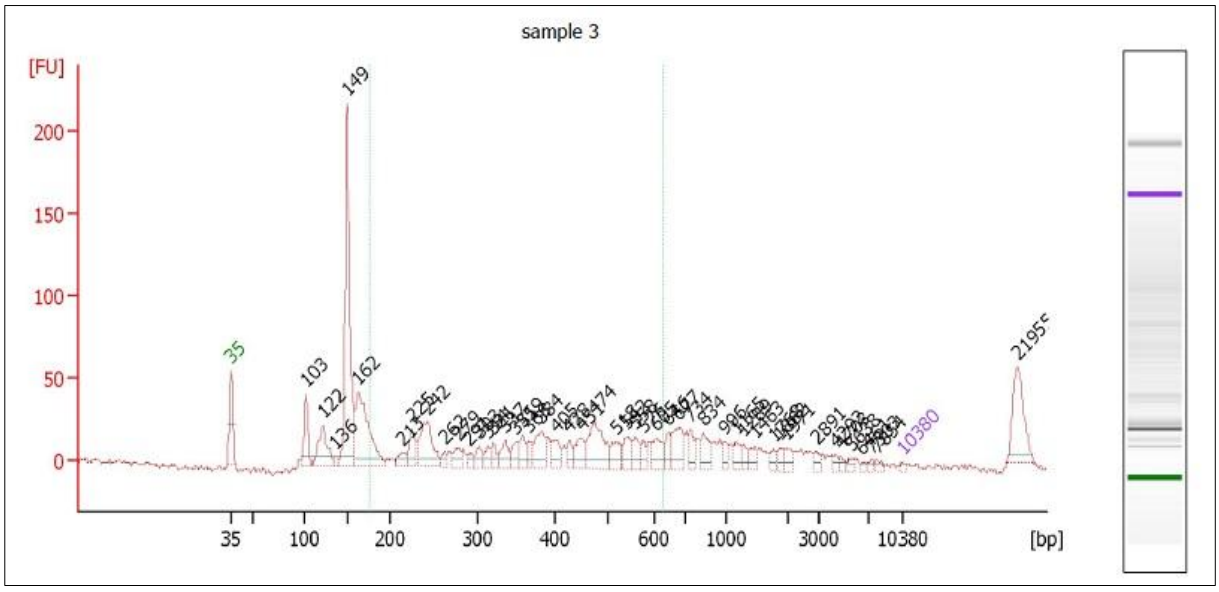

Figure 5. Sample 3 electropherogram. Peak value: 242 bp. The peak at 149 bp is adapter-dimer. The lower marker had to be set manually. The dark band in the electrophoretic run is adapter-dimer.

Sample 3 diluted (2nd run)

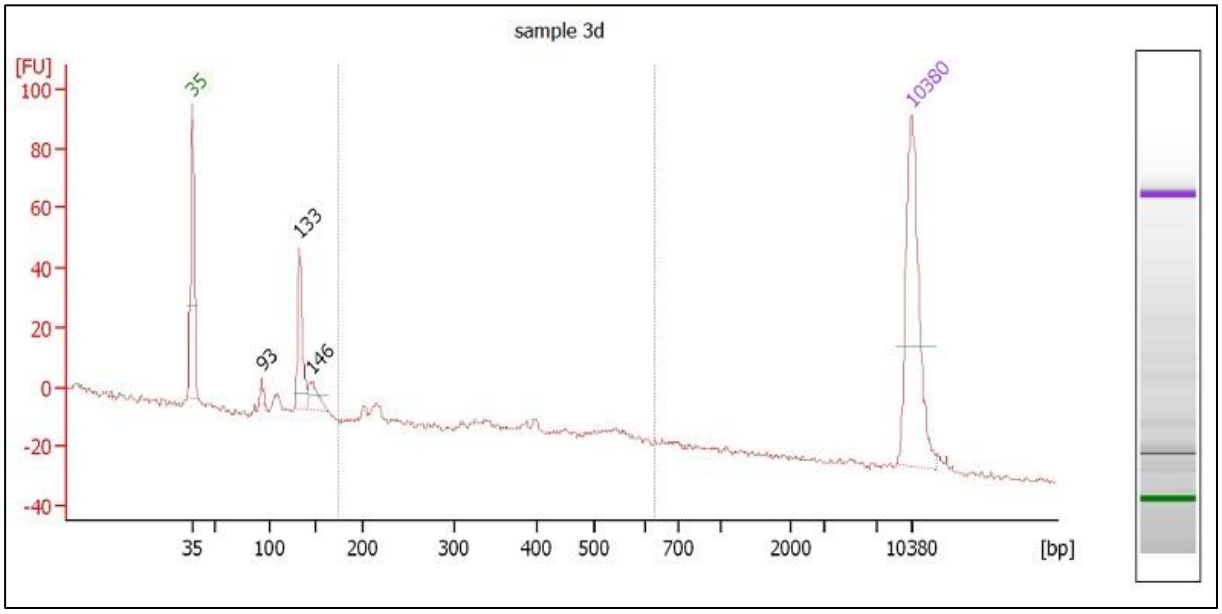

Figure 6. Sample 3 diluted electropherogram. The sample cannot be detected. The peak at 133 is adapter-dimer. The lower marker had to be set manually. The baseline is off. The validation was not successful.

Sample 3 diluted (3rd run)

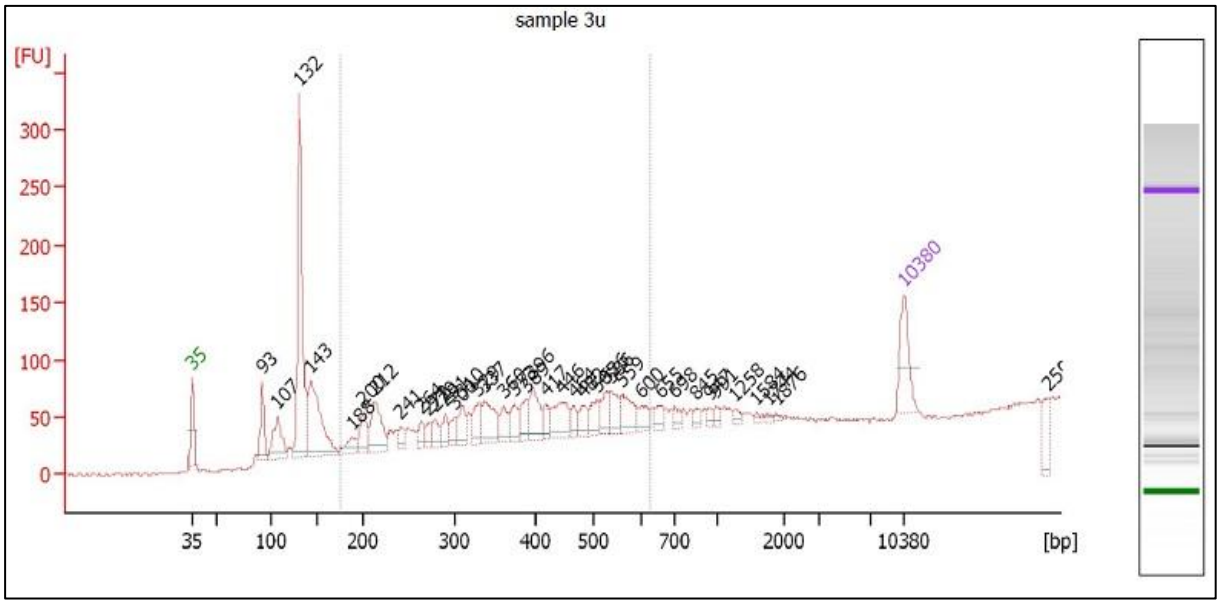

Figure 7. Sample 3 undiluted electropherogram. Peak value: 212 bp. The peak at 132 bp is adapter-dimer. The baseline is off. The band in the electrophoretic run is adapter-dimer. The validation was not successful.

Sample 4

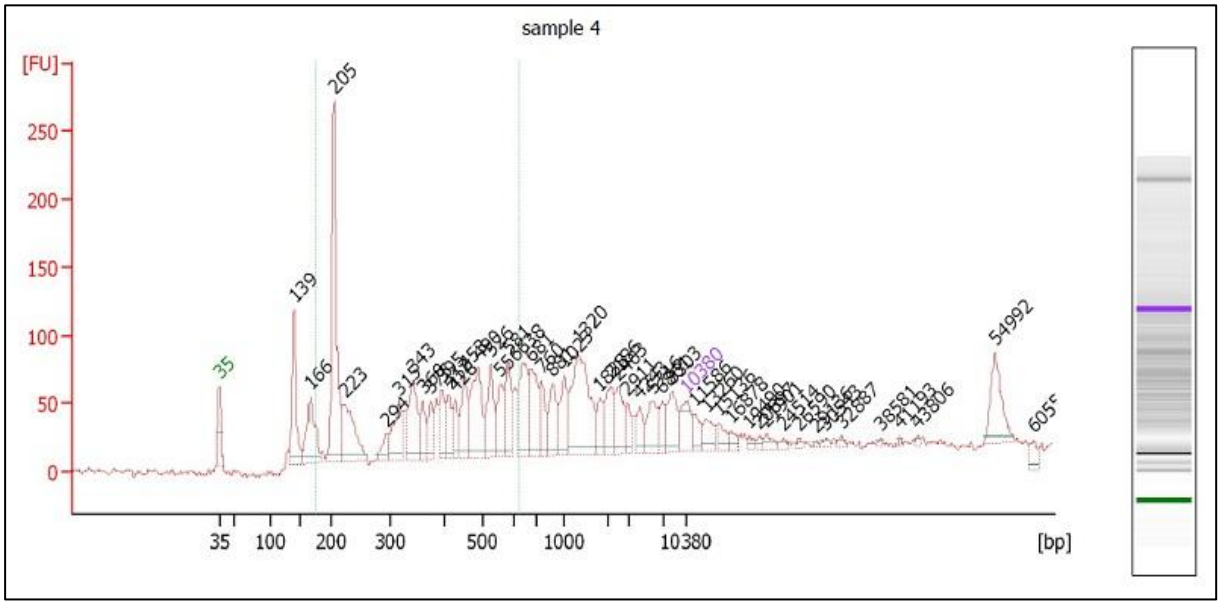

Figure 8. Sample 4 electropherogram. Peak value: 205 bp. The peak at 139 bp is adapter-dimer. The amplicon is visible on the electrophoretic run.

Sample 5

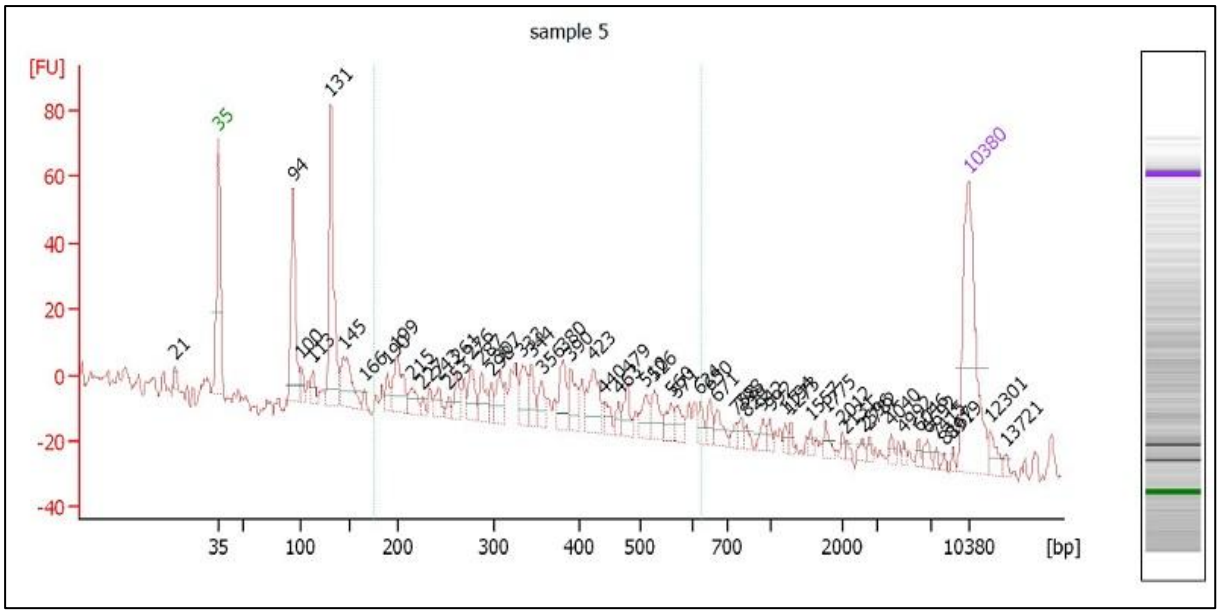

Figure 9. Sample 5 electropherogram. Peak value: 423 bp. The peak at 131 bp is adapter-dimer. The baseline of the electropherogram was off. The validation was not successful.

Sample 5 undiluted

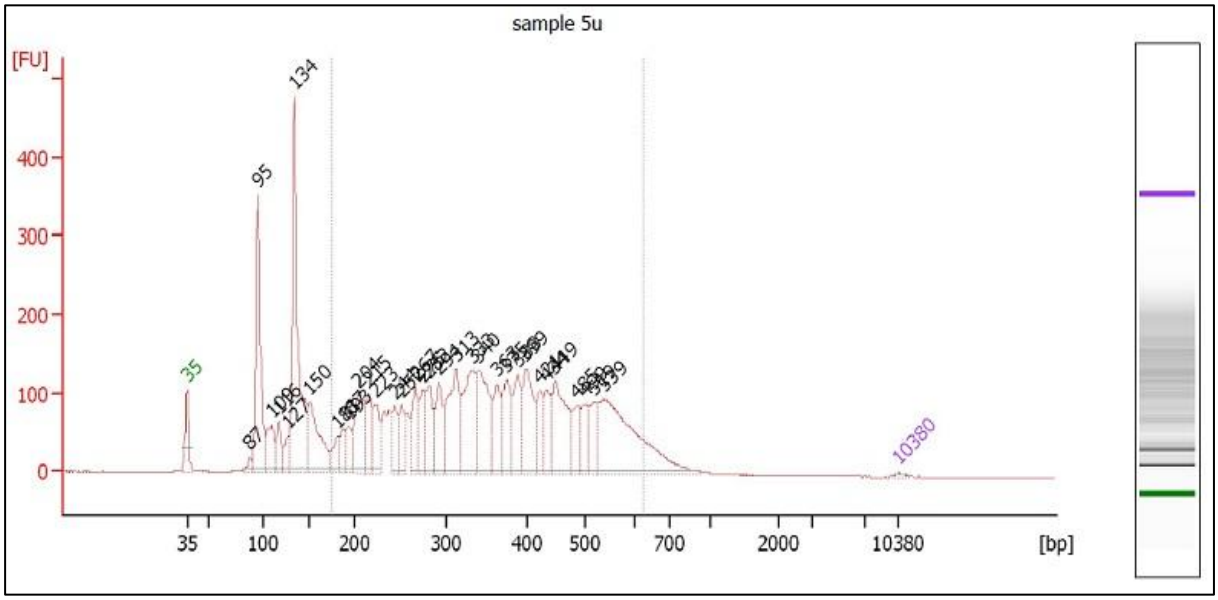

Figure 10. Sample 5 undiluted electropherogram. Peak value: 340 bp. The peak at 134 bp is adapter-dimer. Neither of the dark bands in the electrophoretic run are amplicons.

Sample 6

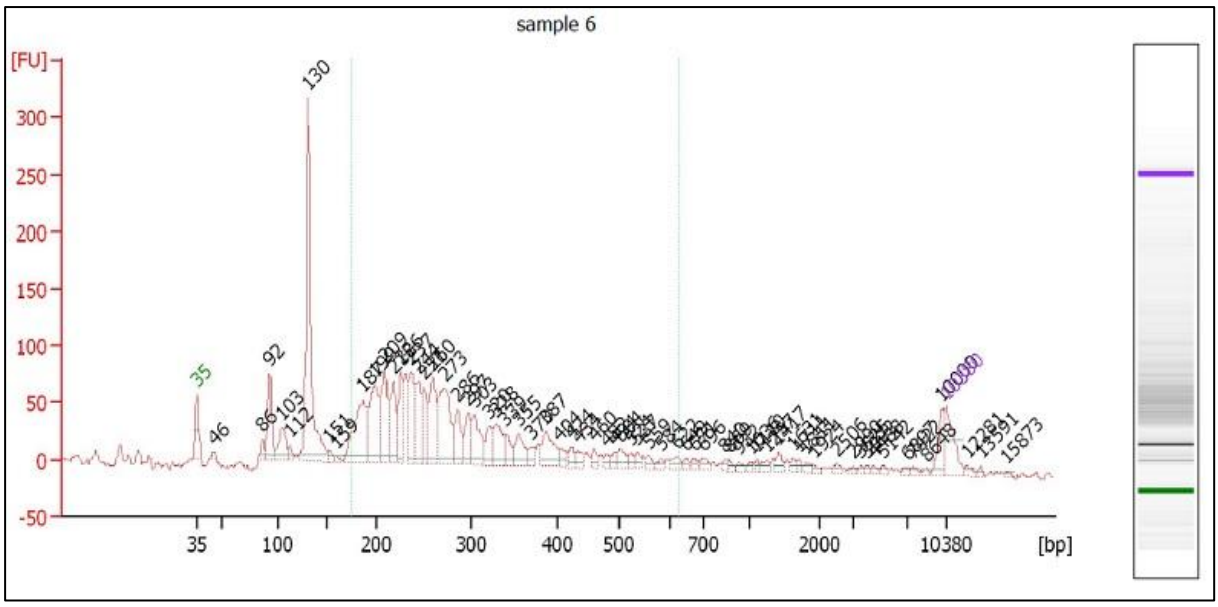

Figure 11. Sample 6 electropherogram. Peak value: 273bp. The peak at 130 bp is adapter-dimer. The band in the electrophoretic run represents the dimer.

Sample 7

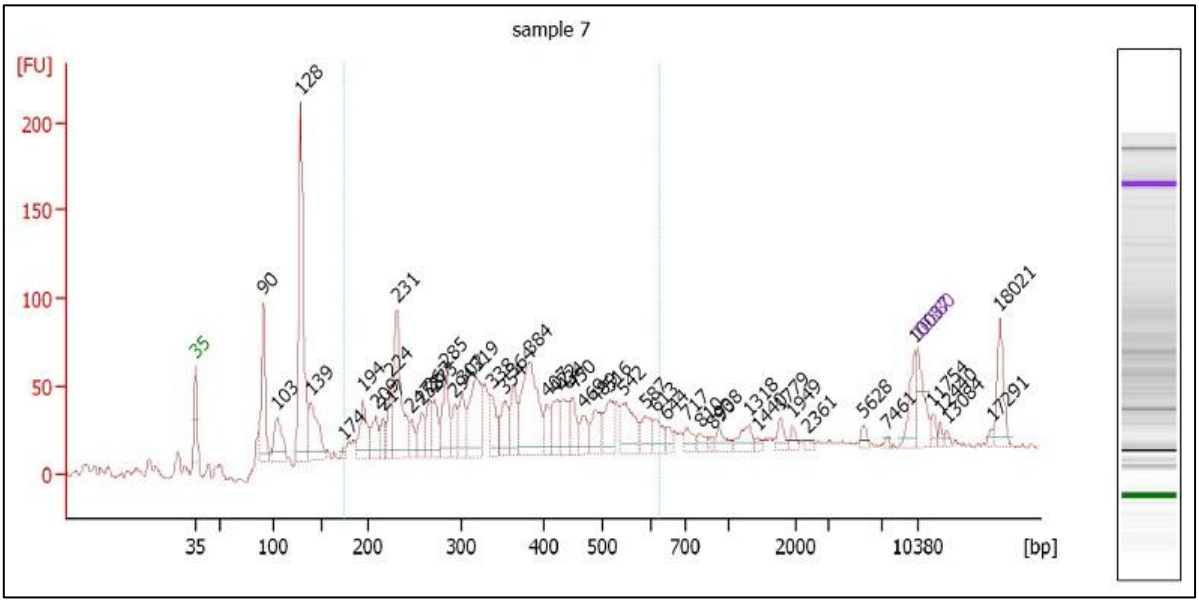

Figure 12. Sample 7 electropherogram. Peak value: 231 bp. The peak at 128 bp is adapter-dimer. Both dimer and amplicon are visible in the electrophoretic run.

Sample 8

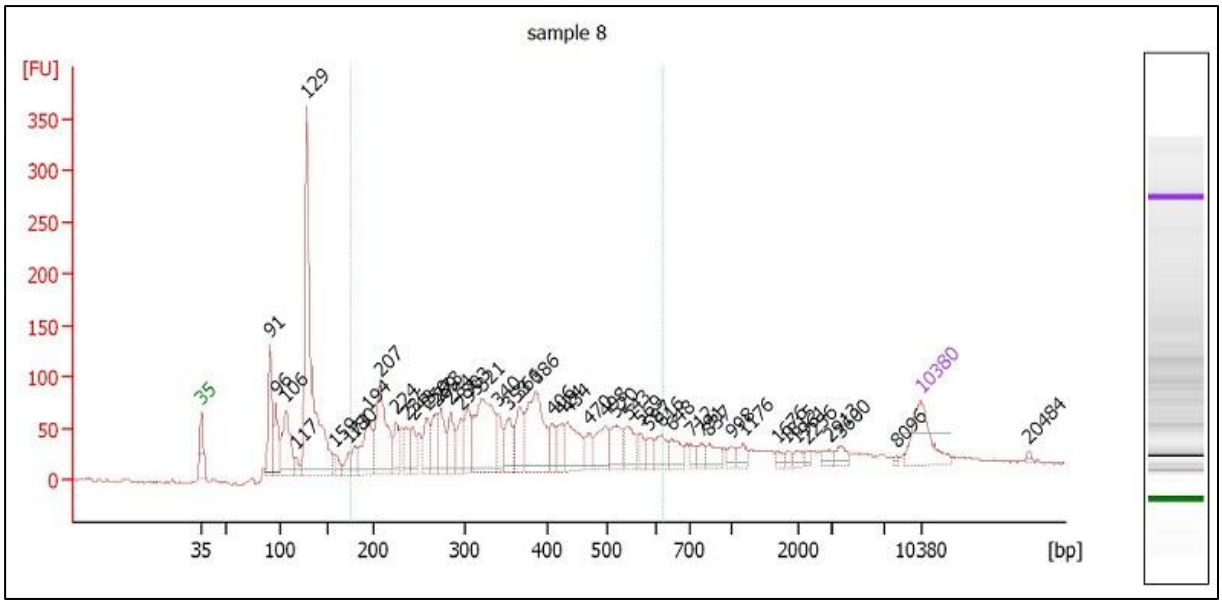

Figure 13. Sample 8 electropherogram. Peak value: 321 bp. The peak at 129 bp is adapter-dimer. The electrophoretic run indicates that there is mostly adapter-dimer.

Sample 9

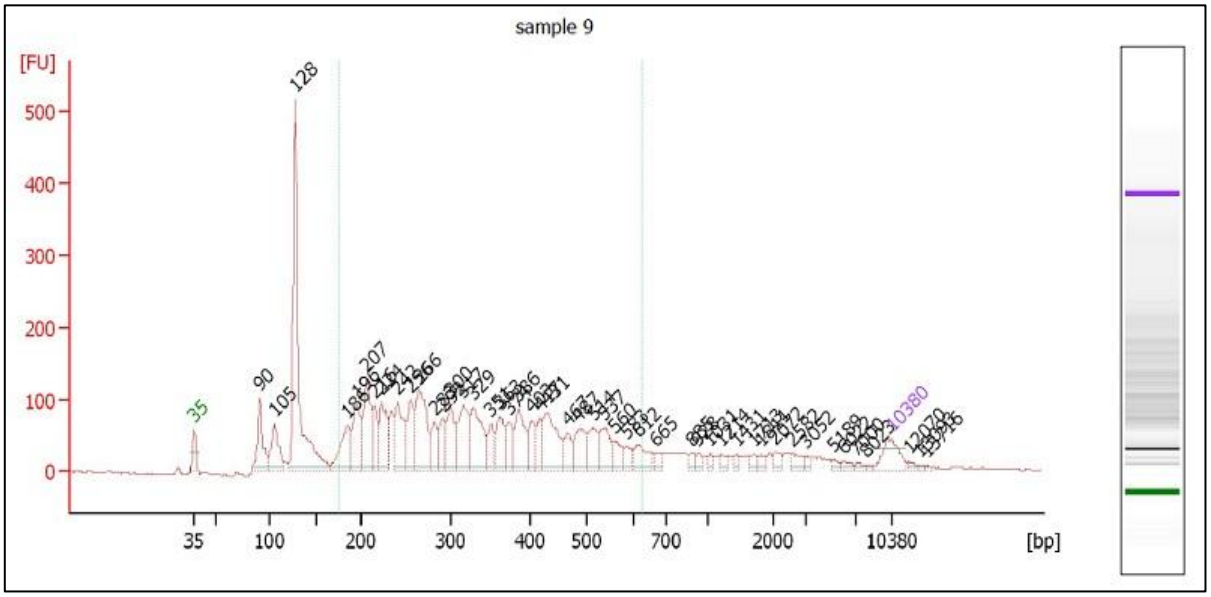

Figure 14. Sample 9 electropherogram. Peak value: 266 bp. The peak at 128 bp is adapter-dimer. The electrophoretic run indicates that there is mostly adapter-dimer.

Sample 10

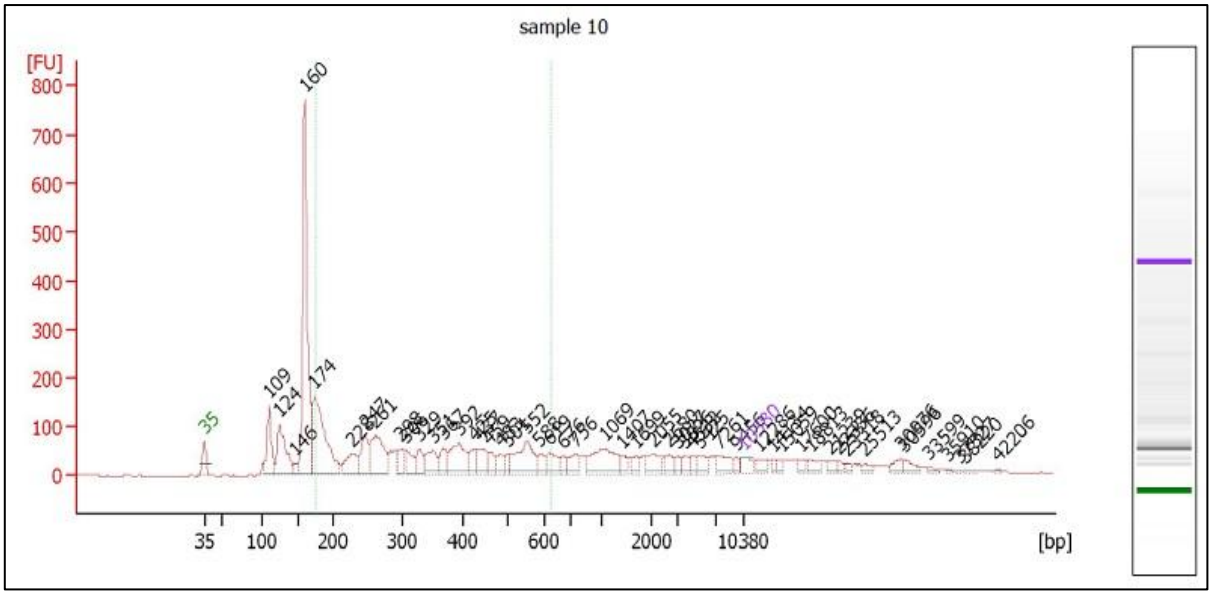

Figure 15. Sample 10 electropherogram. Peak value: 261 bp. The peak at 160 bp is adapter-dimer. The electrophoretic run indicates that there is mostly adapter-dimer.

Sample 11

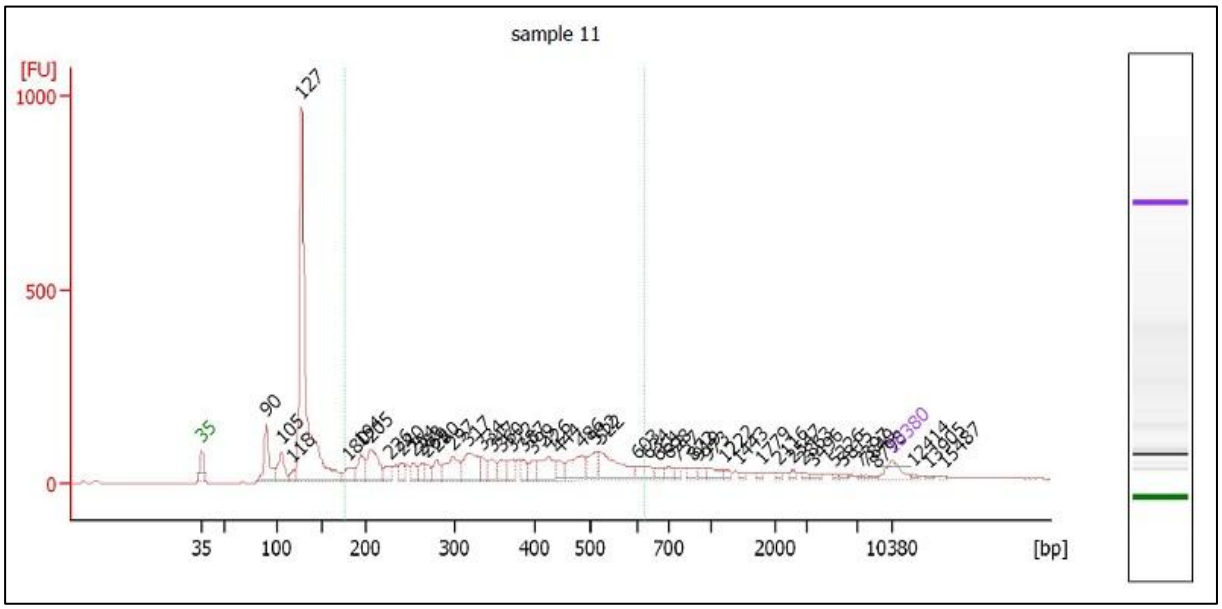

Figure 16. Sample 11 electropherogram. Peak value: 522 bp. The peak at 127 bp is adapter-dimer. The electrophoretic run indicates that there is mostly adapter-dimer.

Sample 12

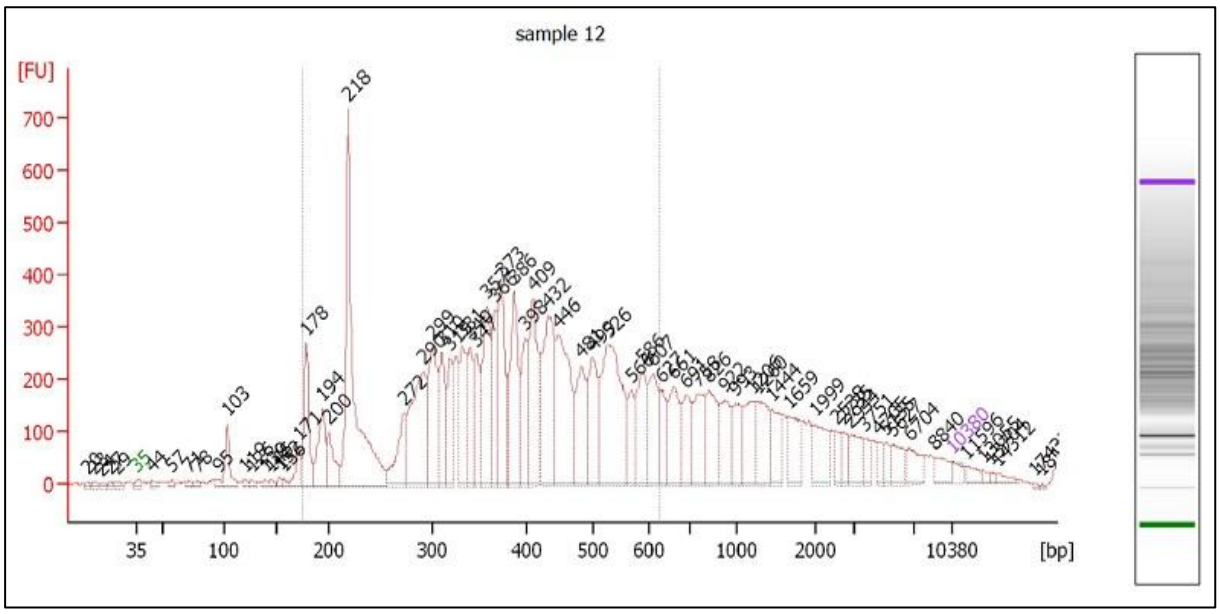

Figure 17. Sample 12 electropherogram. Peak value: 218 bp. The amplicon at 218 bp is visible in the electrophoretic run.

Sample 12 diluted

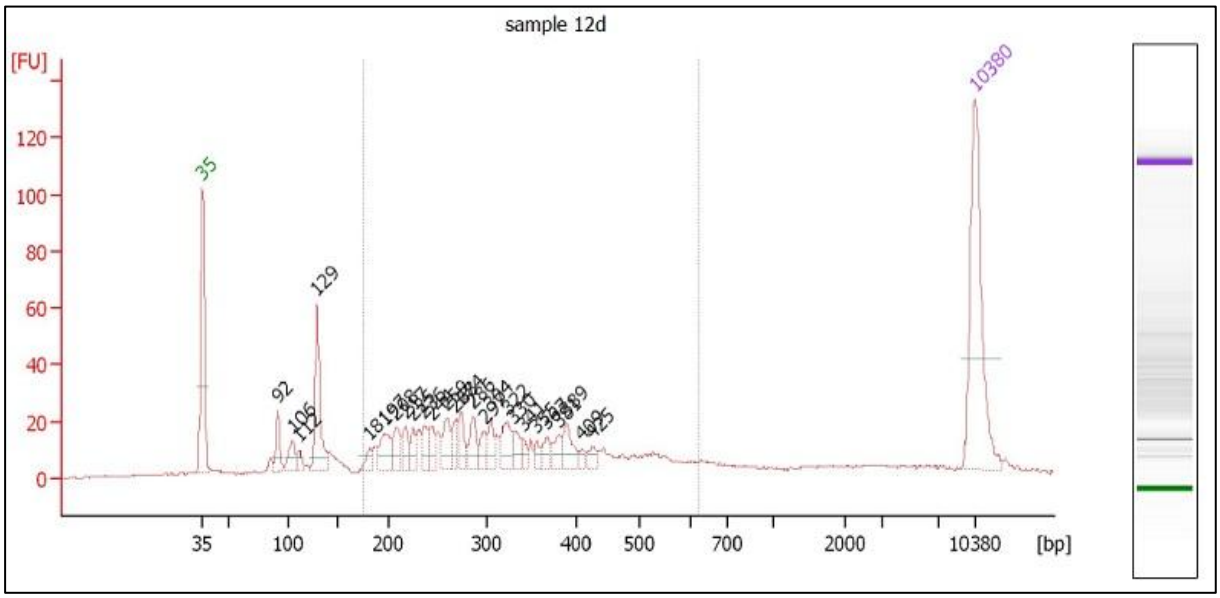

Figure 18. Sample 12 diluted electropherogram. Peak value: 425 bp. The peak at 129 bp is adapter-dimer. The higher marker had to be set manually. The adapter-dimer is primarily visible in the electrophoretic run.

Sample 13

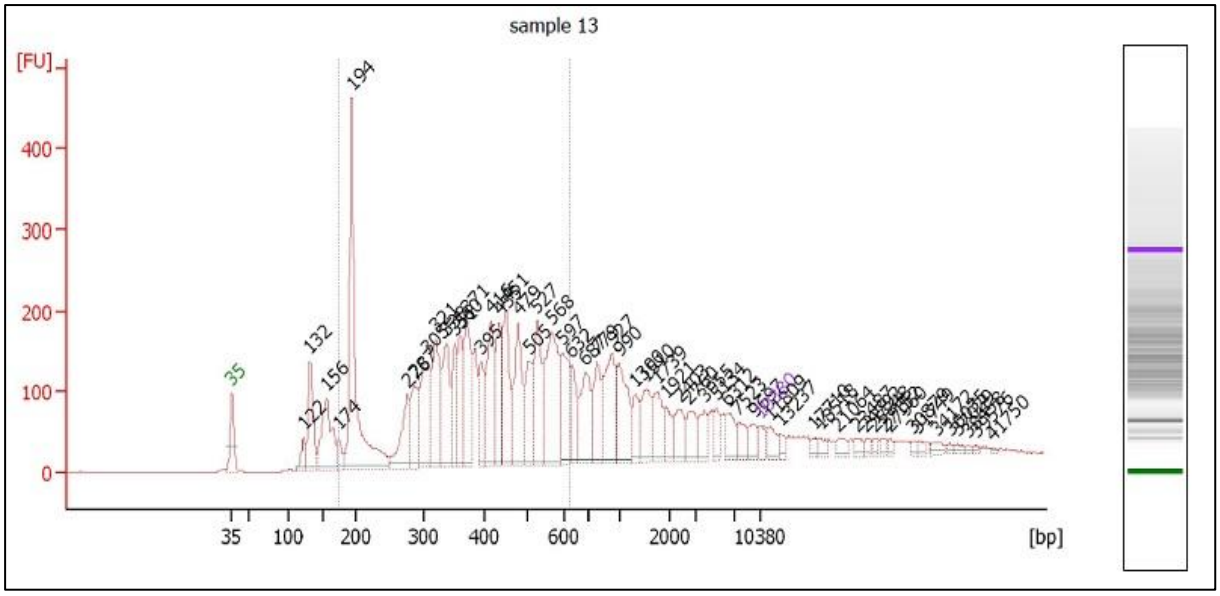

Figure 19. Sample 13 electropherogram. Peak value: 194 bp. The peak at 156 bp is adapter-dimer. The amplicon is visible as the darker band in the electrophoretic run.

Sample 13 diluted

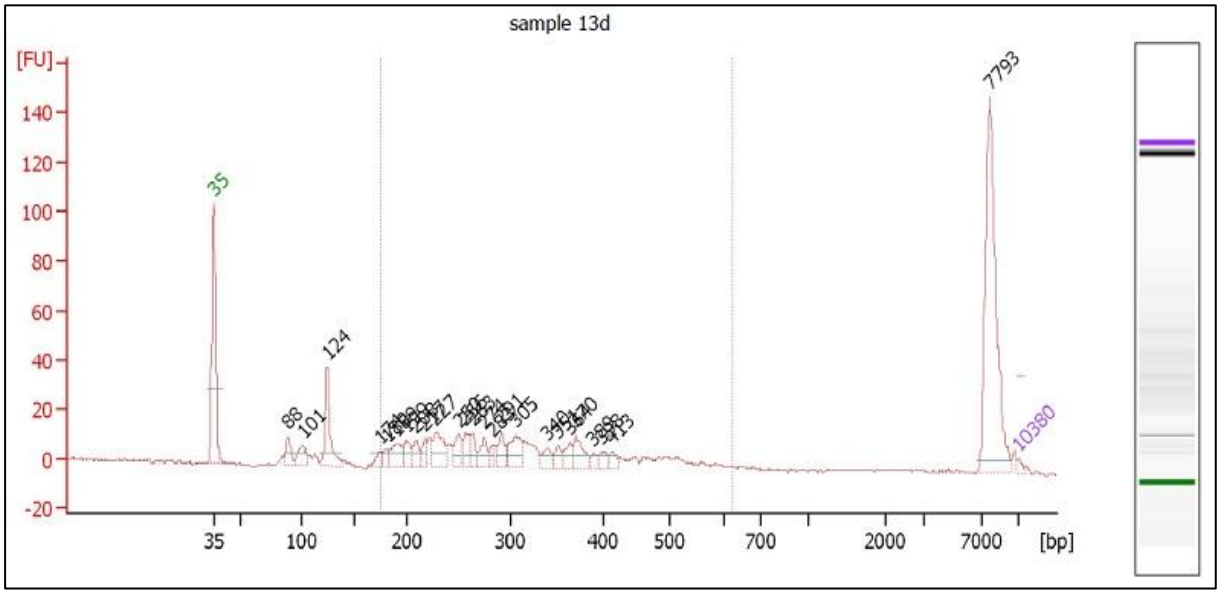

Figure 20. Sample 13 diluted electropherogram. Peak value: 227 bp. The peak at 124 bp is adapter-dimer. Both lower and higher markers had to be set manually. The electrophoretic run indicates that there is a single product, but it is too large to be an acceptable amplicon size.

Sample 14

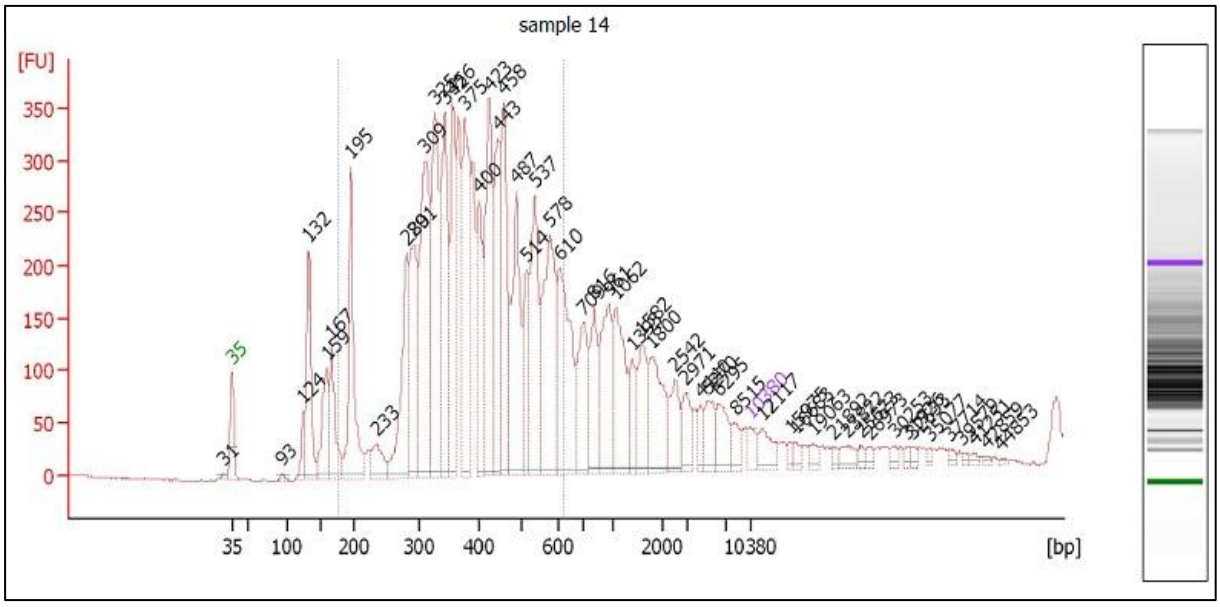

Figure 21. Sample 14 electropherogram. Peak value: 309 bp. The peak at 132 bp is adapter-dimer.

Sample 14 diluted

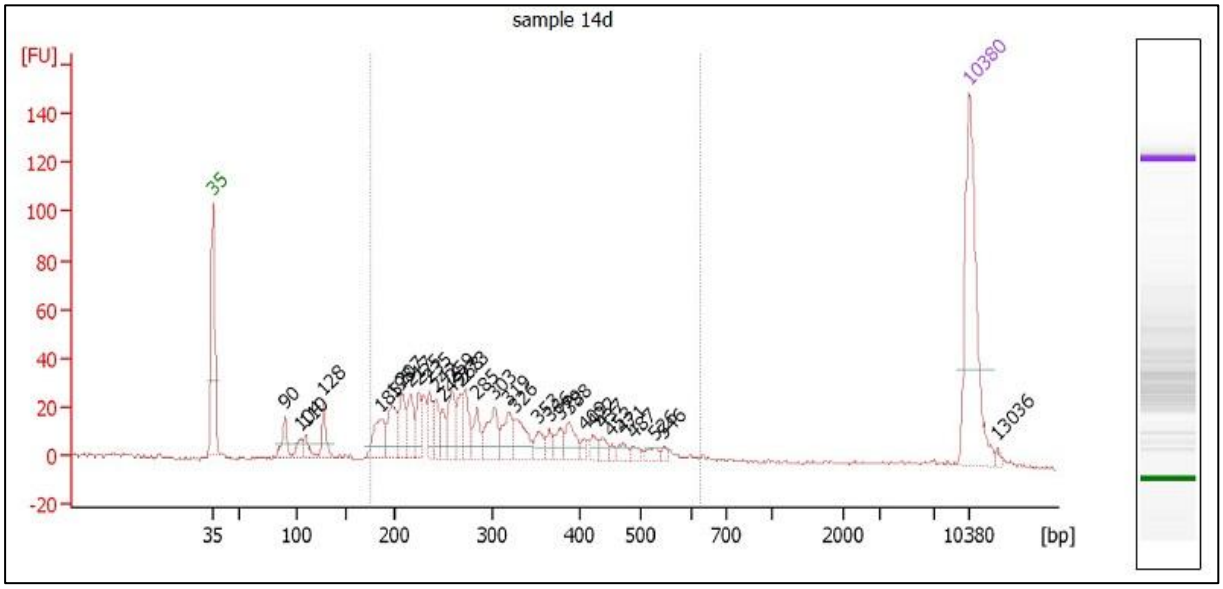

Figure 22. Sample 14 diluted electropherogram. Peak value: 303 bp. The peak at 128 bp is adapter-dimer. No clear bands are visible on the electrophoretic run.

Sample 15

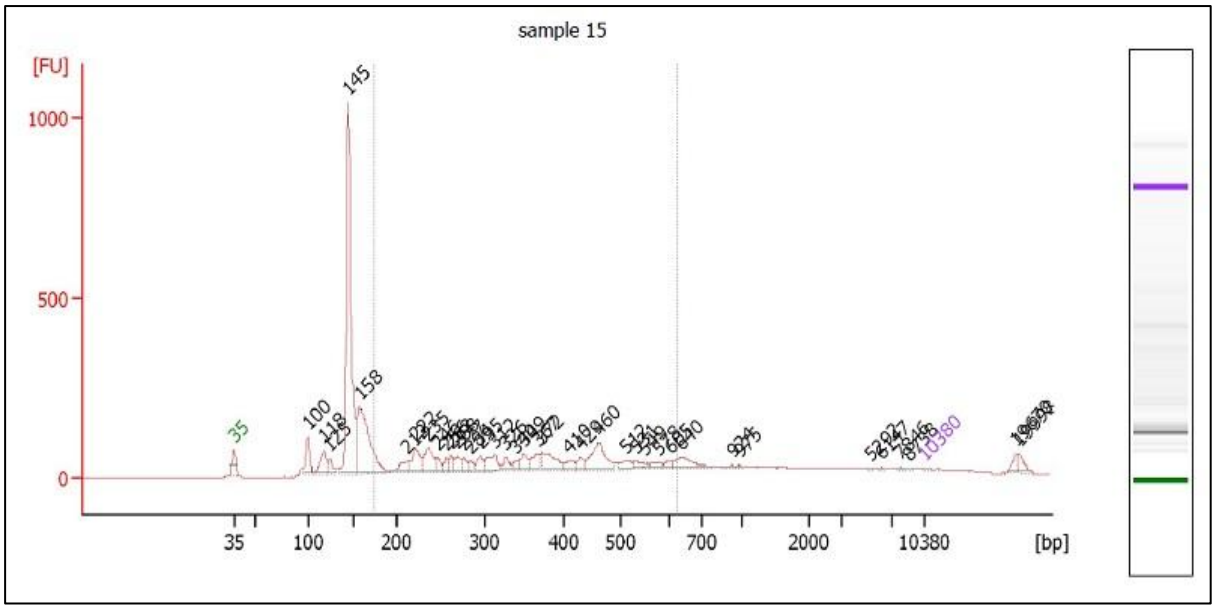

Figure 23. Sample 15 electropherogram. Peak value: 460 bp. The peak at 145 bp is adapter-dimer. The dimer is visible on the electrophoretic run.

Sample 15 diluted

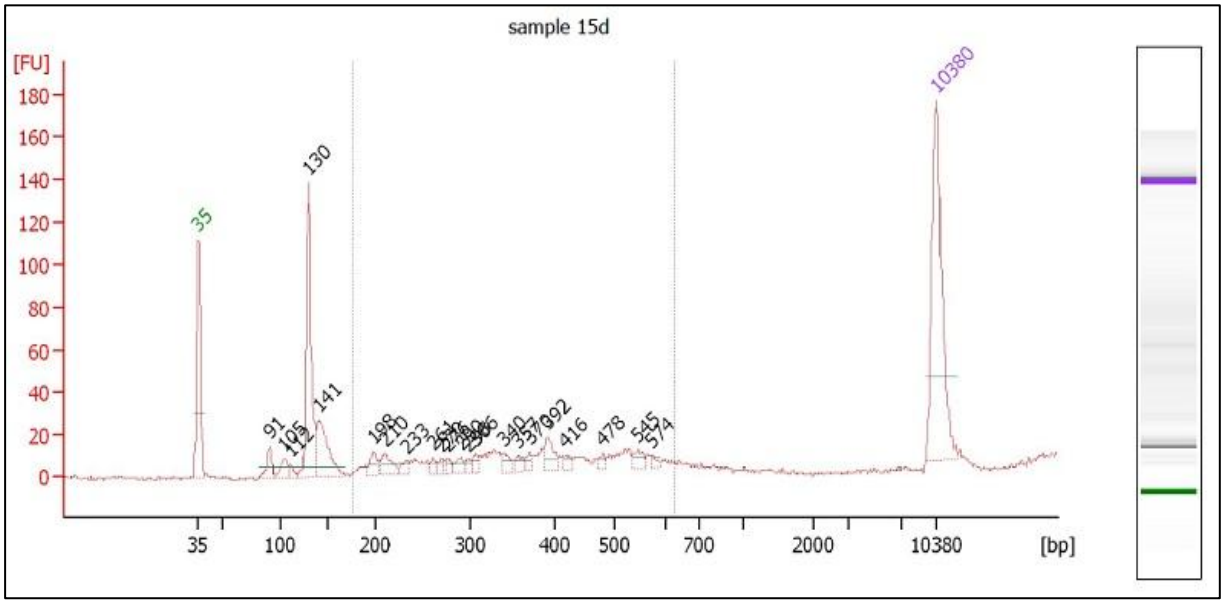

Figure 24. Sample 15 diluted electropherogram. Peak value: 210 bp. The peak at 130 bp is adapter-dimer. Only the dimer is visible on the electrophoretic run.

Sample 16

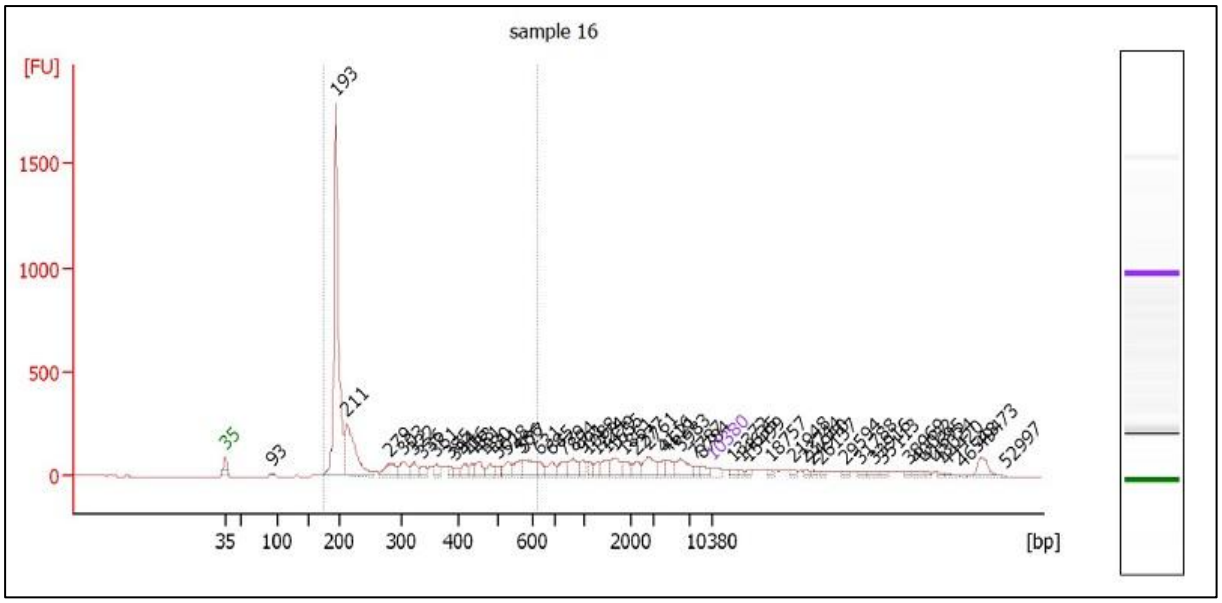

Figure 25. Sample 16 electropherogram. Control ECD. Peak value: 211 bp. The peak at 193 bp is not adapter-dimer, but a product. The amplicon product is visible on the electrophoretic run.

Appendix 10. Concentrations of Samples Measured with 2100 Bioanalyzer

Tables 1 and **2** show the concentrations of the samples as measured with Bioanalyzer and also the values chosen for subsequent use in the study because of their reliability.

Table 1. Measured sample concentrations with Bioanalyzer. All of the sample concentrations measured with Bioanalyzer. Additional measurement results are indicated directly after the initial measurement values. Key: "d" sample was diluted for measurement, ≯ over the range.

Sample #	Concentrations with possible reruns (ng/µl)		
1	6.73	-	-
2	8.51	2.13	2.4 d
3	16.89 ≯	0.19 d	1.27
4	3.26	-	-
5	0.41	101.26 ≯	-
6	3.37	-	-
7	3.58	-	-
8	2.06	-	-
9	4.42	-	-
10	4.91	-	-
11	3.36	-	-
12	26.89 ≯	4.2 d	-
13	14.64 ≯	140.3 d ≯	-
14	15.65 ≯	4.6 d	-
15	46.49 ≯	1.1 d	-
16	7.77	-	-

Table 2. Usable final concentration values. The concentrations and molarities of samples which had consistent results, measured with Bioanalyzer.

Sample #	Concentration (ng/µl)	Molarity (region 175-625 bp) (nmol/l)
1	6.73	38.9
4	3.26	15.8
6	3.37	19.8
9	4.42	23.0
10	4.91	26.1
11	3.36	16.5
13	14.64	66.0
14	4.6	26.1
16	7.77	49.5

Appendix 11. Validation of Amplicon Size with LabChip GXI

LabChip GXI was used to revalidate the amplicon size of several samples. **Figures 1-7** show the electropherograms of samples 2, 3, 5, 7, 8, 12 and 15.

Sample 2

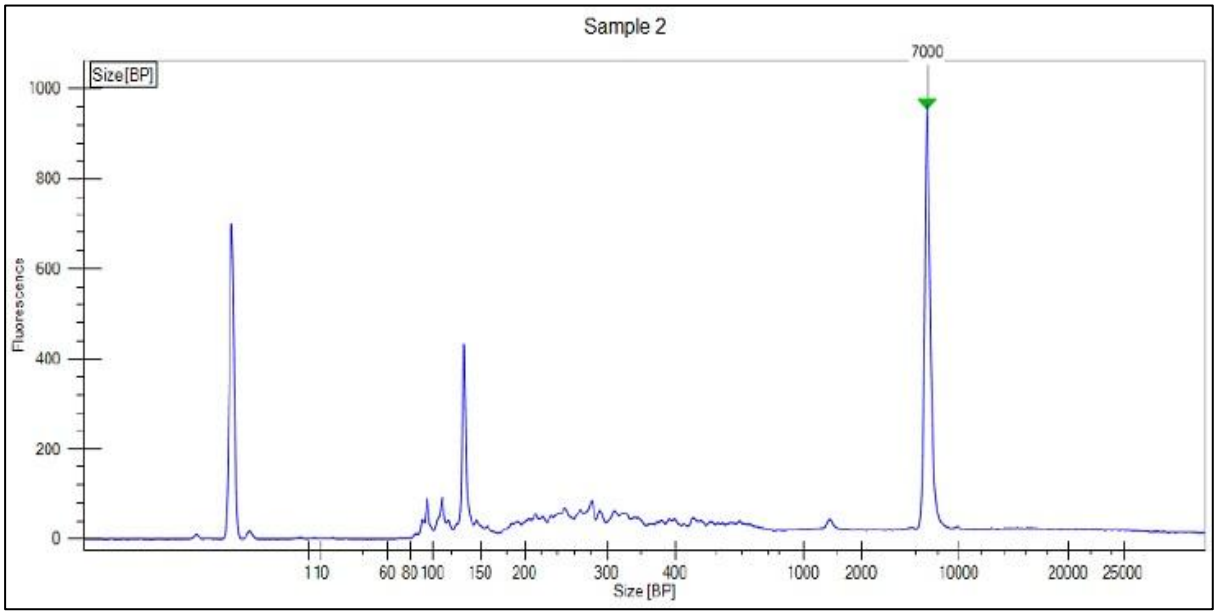

Figure 1. Sample 2 electropherogram. Peak value: 281 bp. The peak at 133 bp is adapter-dimer.

Sample 3

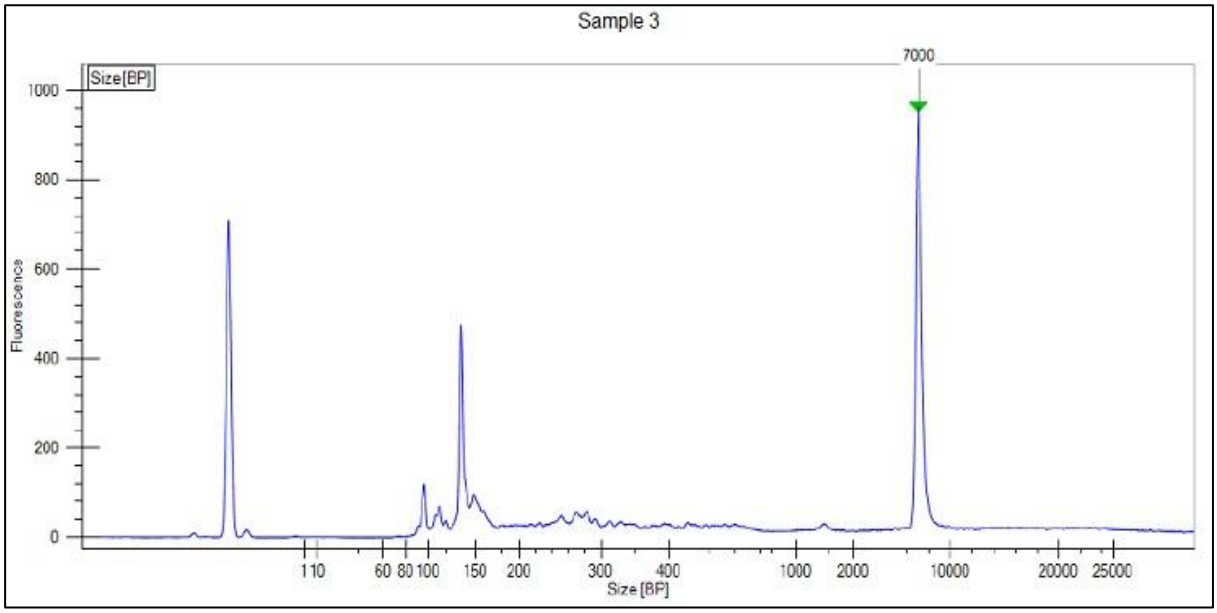

Figure 2. Sample 3 electropherogram. Peak value: 270 bp. The peak at 135 bp is adapter-dimer.

Sample 5

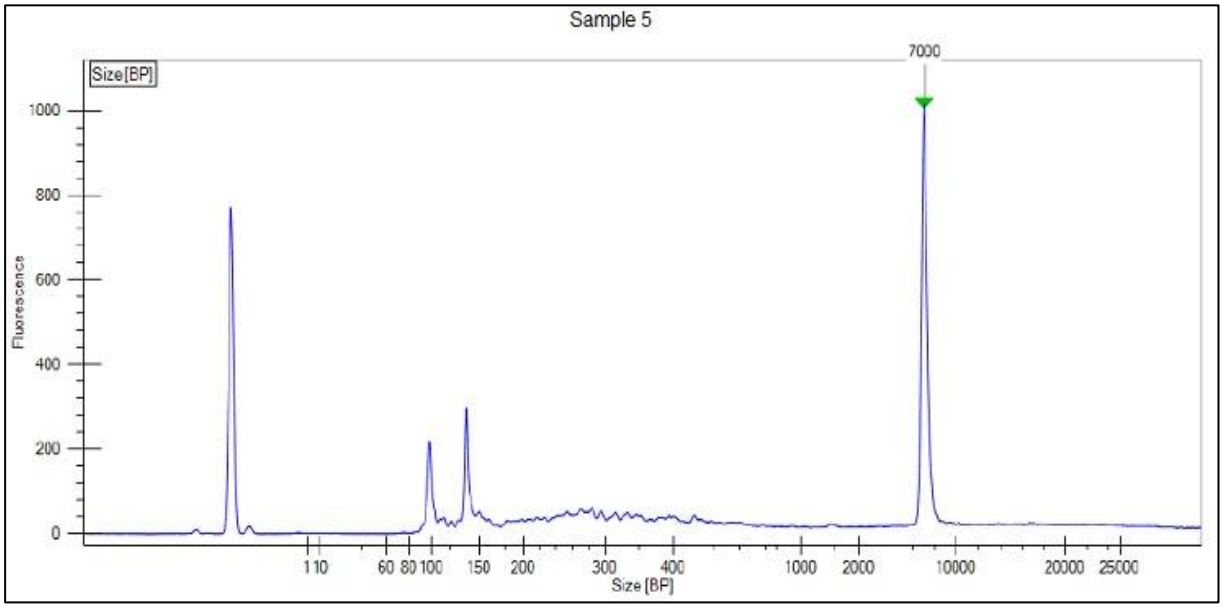

Figure 3. Sample 5 electropherogram. Peak value: 315 bp. The peak at 137 bp is adapter-dimer.

Sample 7

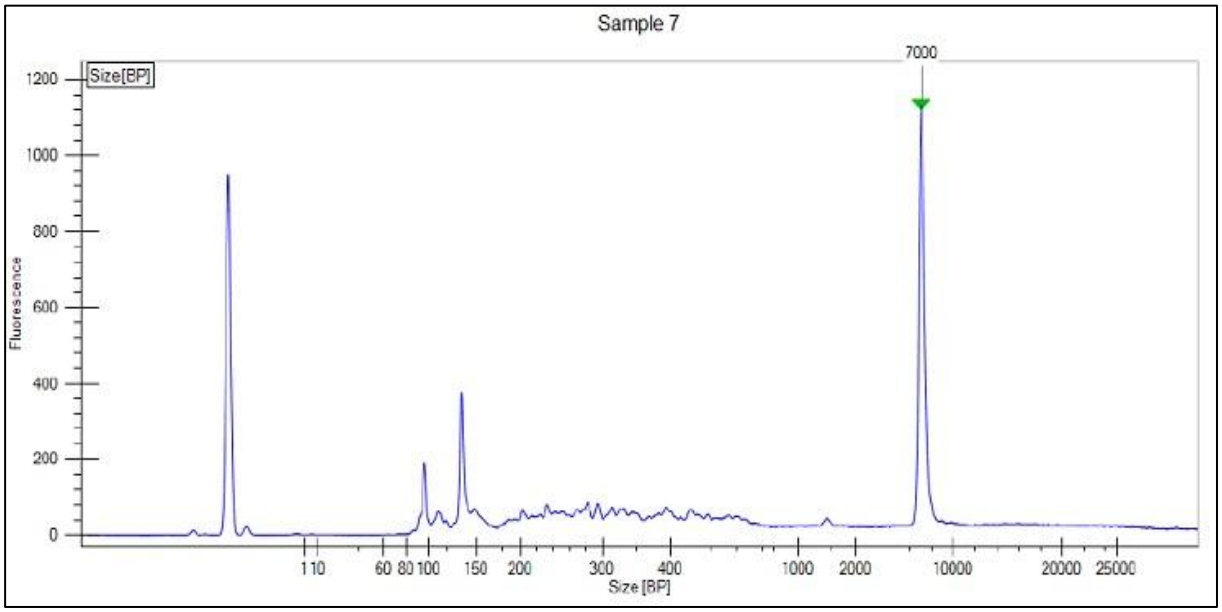

Figure 4. Sample 7 electropherogram. Peak value: 393 bp. The peak at 135 bp is adapter-dimer.

Sample 8

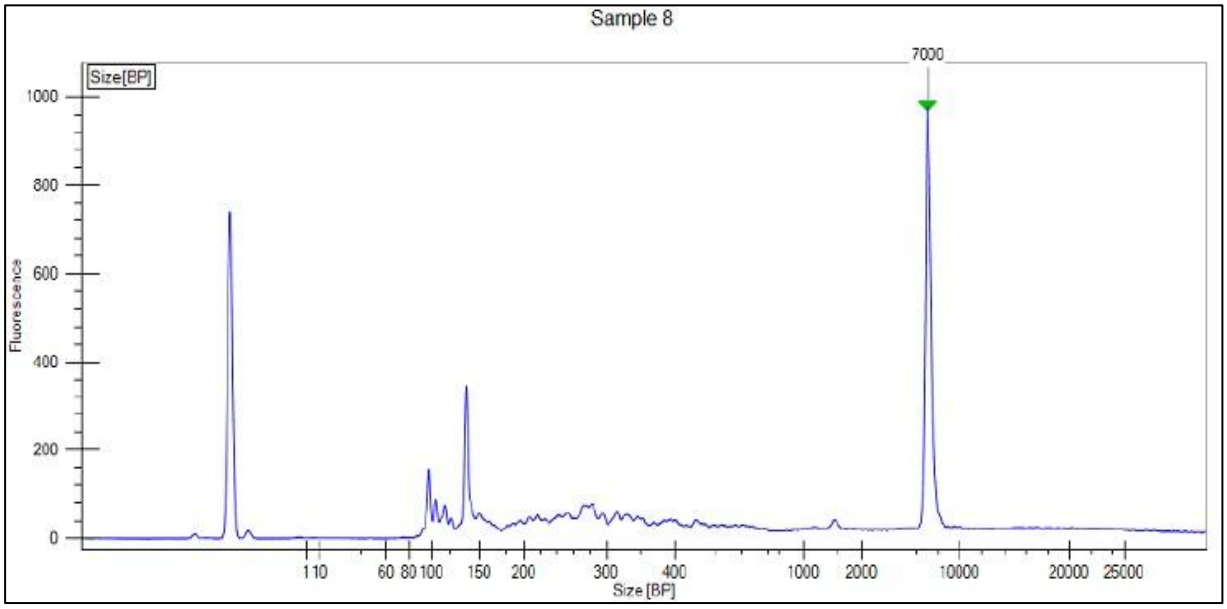

Figure 5. Sample 8 electropherogram. Peak value: 274 bp. The peak at 136 bp is adapter-dimer.

Sample 12

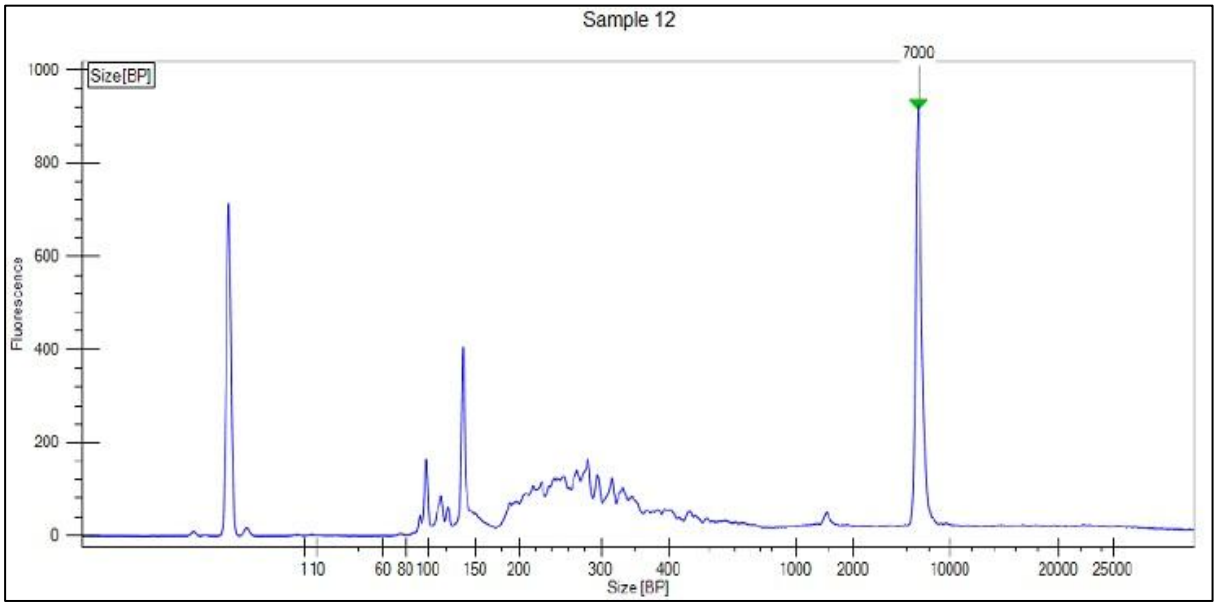

Figure 6. Sample 12 electropherogram. Peak value: 284 bp. The peak at 137 bp is adapter-dimer.

Sample 15

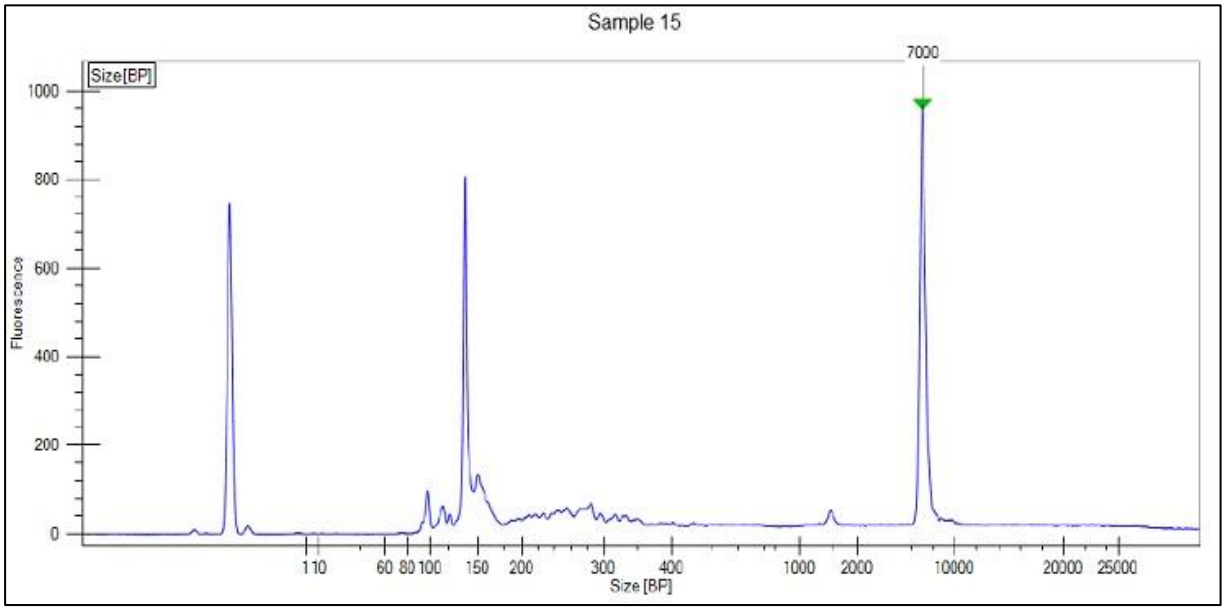

Figure 7. Sample 15 electropherogram. Peak value: 285 bp. The peak at 137 bp is adapter-dimer.

Appendix 12. Concentrations of Samples Measured with LabChip

Table 1. Concentrations and molarities of samples. Measured concentrations and molarities with LabChip.

Sample #	**Concentration (ng/µl)**	**Molarity (region 175-625 bp) (nmol/l)**
2	1.746	9.049
3	1.223	6.352
5	1.472	7.607
7	1.933	9.737
8	1.737	9.071
12	3.499	19.074
15	1.443	7.842

Appendix 13. Sequencing Run Data

Run data from MiSeq is shown in **Table 1.** and **Figures 1-2** below.

Table 1. Reads Mapped to Index ID. The table shows the total number of reads, the percentage of aligned reads and shows what percentages of reads have aligned to which samples (sample ID).

TOTAL READS	PF READS	% READS IDENTIFIED (PF)	CV	MIN	MAX
19141716	17568612	79.9615	0.5686	2.1173	18.8430

INDEX NUMBER	SAMPLE ID	PROJECT	INDEX 1 (I7)	INDEX 2 (I5)	% READS IDENTIFIED (PF)
1	1	NA	AACGTGAT		6.0416
2	4	NA	AGTGGTCA		18.8430
3	6	NA	ACATTGGC		10.0535
4	9	NA	CGCTGATC		9.6627
5	10	NA	ACAAGCTA		5.7167
6	11	NA	CTGTAGCC		11.1405
7	13	NA	AACAACCA		4.1277
8	14	NA	AACCGAGA		12.2583
9	16	NA	AAGACGGA		2.1173

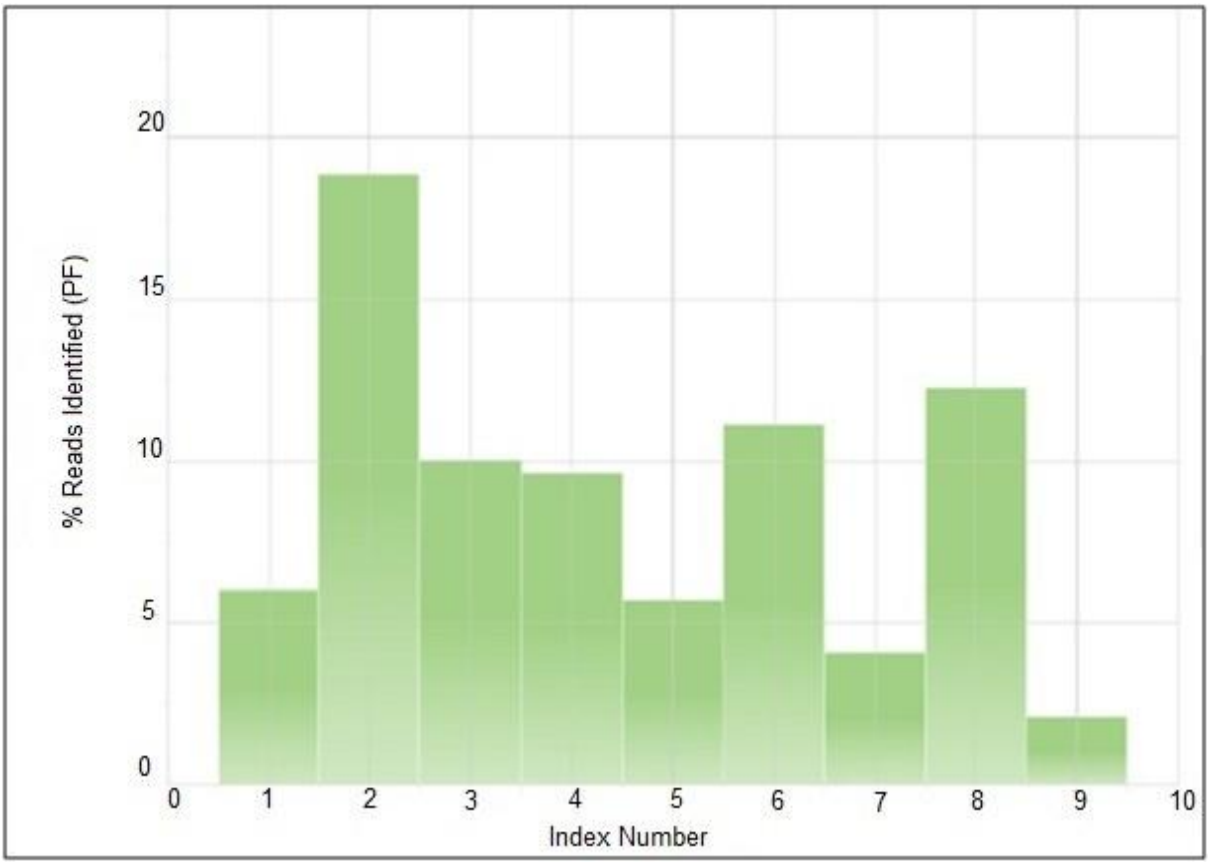

Figure 1. Reads Mapped to Index ID. The graph shows the distribution of the identified reads among the different samples sequenced.

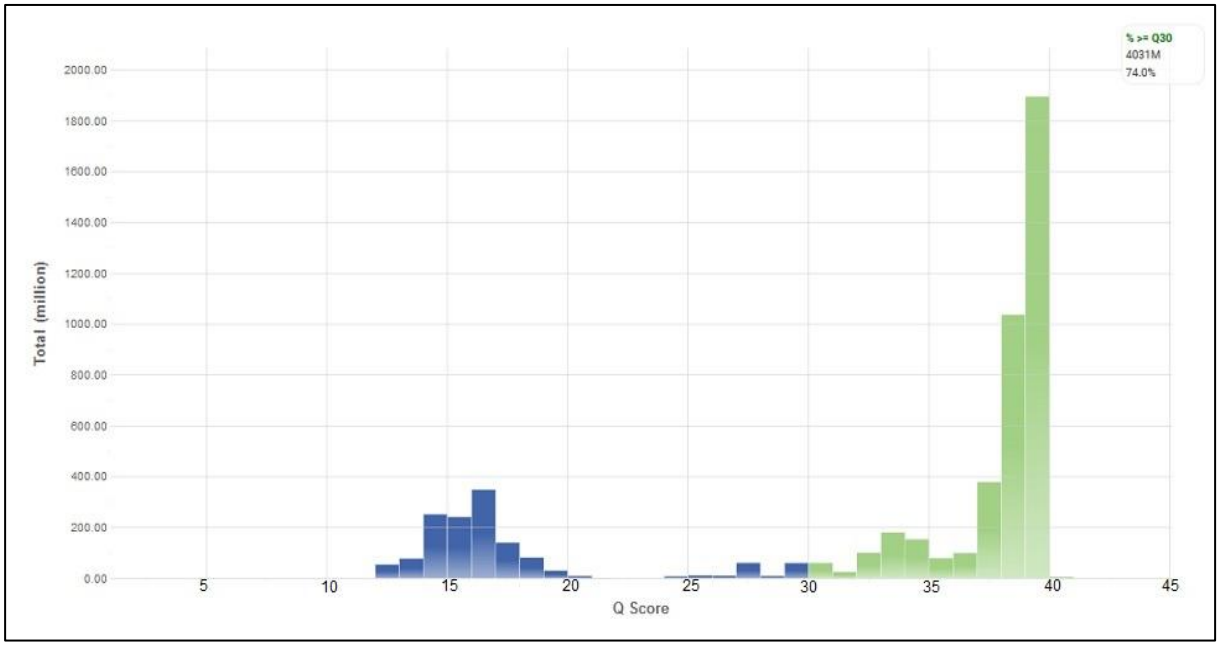

Figure 2. QScore Distribution. The plot shows the distribution of the quality score of the bases. 74.0% of all of the bases (green) have a quality score of over Q30. Bases with a quality score of less than Q30 are in blue. 26.0% of the bases have a quality score ≤ Q30.

Calculating sequencing coverage

Coverage was calculated with the following equation:

Coverage = N x L/G

N = number of reads identified for samples

L = average read length

G = length of targeted region

N = 15 305 716 reads

L = 100 bp

G = 94 607 bp x 9 samples = 851 463 bp

C = $\frac{100\ bp\ x\ 15\ 305\ 716}{851\ 463\ bp}$ = 1 797

Appendix 14. Sequencing and qPCR Results

Sequencing

Figure 1. shows the detected SNVs within the selected cell lines from DNA Pool 1.

		Sample 1	Sample 10	Sample 11	Sample 4	Sample 6	Sample 13	Sample 14	Sample 9	Sample 16
LNCaP	ABL1									
	AR	1101	245	142/2326 =6 %	737/4105 =18 %	88/206 = 43 %	227/356= 64 %	564/705= 80 %	500/502= 100 %	165
	NOTCH1									
	PIK3R1		308	63/1518= 4 %	329/2804 =12 %	106/362= 29 %	129/409= 32 %	332/965= 34 %	546/1164 =47 %	221
	SMO									
MDA-MB-415	ALK									
	BRAF									
	MAP2K4	3164/3169 =100 %	3044/3053 =100 %	4454/4957 =90 %	8668/1263 4 = 69 %	1921/6962 =28 %	524/3015 =17 %	1043/1169 2 = 9%	7544	839
	CSF1R	1659/3888 =43 %	1496/3604 =42 %	2404/6027 =40 %	4334/1407 9 = 31 %	1213/7394 =16 %	229/2914 =8 %	265/8675 =3 %	4785	1291
	ERBB4									
		0 % LNCaP 100 % MDA		10 % LNCaP 90 % MDA	30 % LNCaP 70 % MDA	70 % LNCaP 30 % MDA	80 % LNCaP 20 % MDA	90 % LNCaP 10 % MDA	100 % LNCaP 0 % MDA	Agilent control

Key: 0 reads; SNV, with % of SNV; No mutation, with # of reads

Figure 1. Sequencing call map. All the sequenced samples with detected SNVs are shown according to cell line. Only four SNVs were detected. The % of SNV = number of reads with SNV/total number of reads at target region. To make comparison of the fractions of DNA easier, samples are in order of increasing amounts of LNCaP DNA and decreasing amounts of MDA-MB-415 DNA.

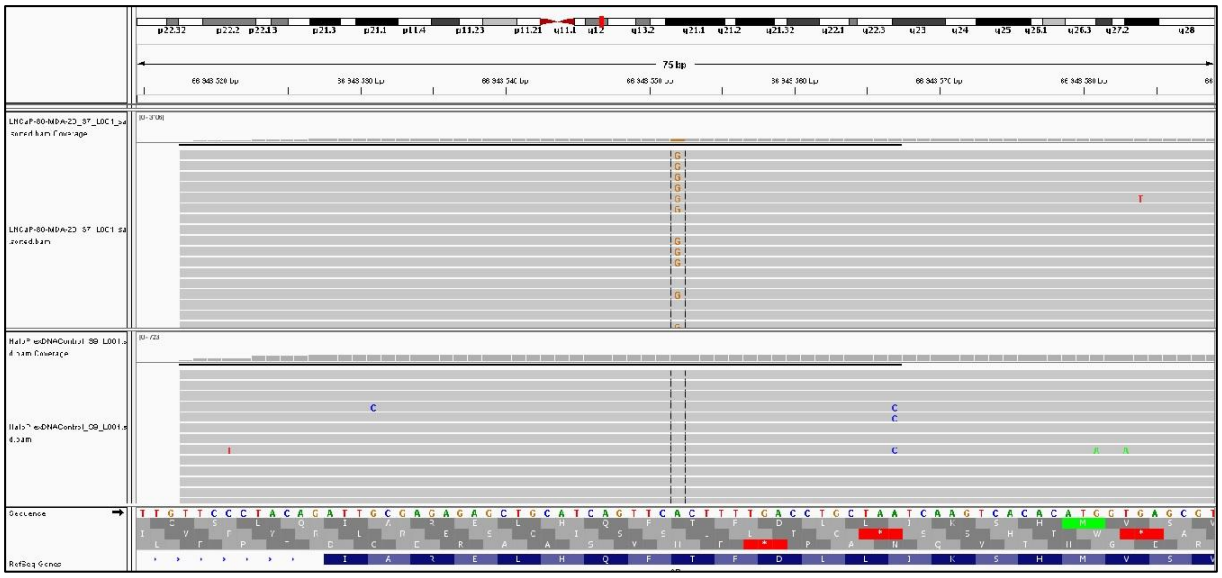

Figure 2. Example of viewing with IGV. A screenshot of IGV with sample 13 being analyzed. There are two tracks being viewed. The top track is sample 13 and the lower track is sample 16 (Agilent DNA control). Both tracks have numerous reads, but only sample 13 has an A>G substitution at a particular coordinate, seen in the vertical path of brackets in the center of the screen. Directly below the last track, the reference genome can be seen.

Quantitative Real-Time PCR

The call map for qPCR genotyped samples is seen in **Figure 3**.

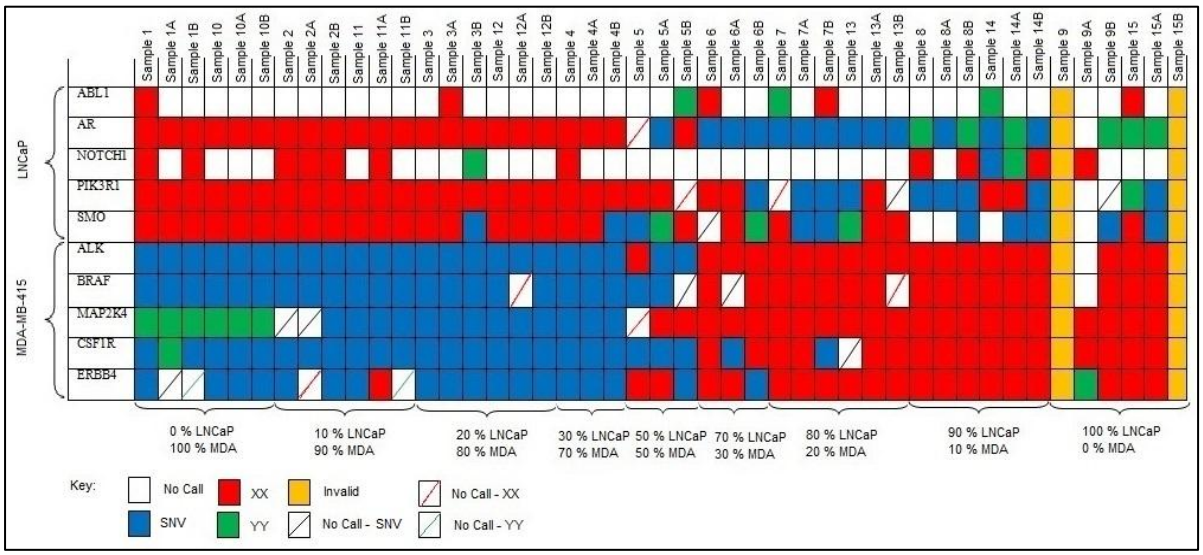

Figure 3. qPCR call map. Shown are all the samples that were genotyped by qPCR. All samples are shown as triplicates. To make comparison of the DNA fractions easier, samples are in order of increasing amounts of LNCaP DNA and decreasing amounts of MDA-MB-415 DNA. Some No Call results changed to SNV, XX, or YY when the confidence threshold was decreased from 65 to 50.