Kati Johanna Piironen

Establishment of a Laboratory Workflow for Analysis of Exonic and Intronic Regions with Next-Generation Sequencing

Helsinki Metropolia University of Applied Sciences

Bachelor of Laboratory Sciences

Laboratory Sciences

Thesis

29.11.2013

Metropolia

Author(s) Title Number of Pages Date	Kati Johanna Piironen Establishment of a Laboratory Workflow for Analysis of Exonic and Intronic Regions with Next-Generation Sequencing 38 pages + 3 appendices 29 November 2013
Degree	Bachelor of Laboratory Sciences
Degree Programme	Laboratory Sciences
Instructor(s)	Anna Benet-Pagès, Head of Department (NGS) Tiina Soininen, Principal Lecturer

This thesis was carried out at Medizinisch Genetisches Zentrum München, department of Molecular Genetics in the section Next Generation Sequencing (NGS).

NGS is a fast developing research field in gene technology, it is also one of the most important fields in biological research. Next Generation Sequencing gains the ability to sequence multiple samples at the same time. This gives the opportunity for cheaper and faster research. Also NGS makes it possible for greater scalability and resolution.

Cancer is becoming one of the biggest diseases in the whole world, more than one million people per year manifests colorectal cancer and ~3 % of them have hereditary non-polyposis colorectal cancer (HNPCC) also known as Lynch syndrome. Four genes are known to be responsible for Lynch syndrome MLH1, MSH2, MSH6 and PMS2. These genes are responsible for the DNA mismatch repair in humans during DNA replication in the cell mitosis process. However, there are a number of patients who suffered of Lynch syndrome but the mutation was not found in their genome with Sanger Sequencing. For these patients it is interesting to sequence also their introns and promoter area since there is a high probability that the mutation causing the disease locates in the regulatory part of these genes.

For providing good sequencing results it is important to have as same size DNA fragments as possible. In this thesis two different fragmentation methods were compared, enzymatic and ultra-sonication. Fragmentation of the DNA had to be robust and little time consuming. Ultra-sonication was found to be the more robust method. The fragment size was narrowed even more. This was done with magnetic beads. When the desirable fragment size was achieved, the rest of the library preparation and enrichment was carried through and the libraries were sequenced with the Illumina MiSeq sequencer.

After the sequencing data-analysis was done. The results from ultra-sonication were found very good and robust. However, the mapped data had very uneven coverage which could not be explained with a mistake in the laboratory preparation.

Keywords	Next Generation Sequencing, Lynch syndrome, Colorectal cancer

Tekijä Otsikko Sivumäärä Aika	Kati Johanna Piironen Laboratorioprotokollan luominen intronisten ja eksonisten alueiden analysoimiseksi seuraavan sukupolven sekvensoinnilla 38 sivua + 3 liitettä 29.11.2013
Tutkinto	Laboratorioanalyytikko (AMK)
Koulutusohjelma	Laboratorioalan koulutusohjelma
Ohjaajat	Osastopäälikkö (NGS), Anna Benet-Pagès Lehtori, Tiina Soininen

Tämä opinnäytetyö toteutettiin Medizinisch Genetisches Zentrum Münchenissä, molekyyligenetiikan osastolla, seuraavan sukupolven sekvensoinnin (NGS) ryhmässä.

NGS on nopeasti kehittyvä tutkimuksen ala geenitekniikassa ja se on myös yksi tärkeimmistä tutkimuksen aloista biologisessa tutkimuksessa. NGS mahdollistaa monien näytteiden yhtäaikaisen sekvensoinnin. Tämä avaa mahdollisuuden halvempaan ja nopeampaan tutkimukseen. NGS mahdollistaa myös suuremman skaalattavuuden ja resoluution.

Syövästä on tulossa yksi maailman yleisimmistä sairauksista, yli miljoona ihmistä joka vuosi sairastuu paksusuolen syöpään ja heistä noin 3 % saa Lynchin syndrooman. Syndroomasta ovat vastuussa neljä geeniä: MLH1, MSH 2, MSH6 ja PMS2. Nämä geenit toimivat mismatch korjaajaentsyymikompleksin valmistajina. Toisinaan potilailta, joilla on Lynchin syndrooma, ei löydy geeni mutaatiota Sangerin Sekvensoinnilla. Kiinnostuksen kohteena on tehdä näille potilaille lisäksi tehdä myös sekvensointi, joka kattaa myös heidän introni ja promoottorialueensa.

Hyvien sekvensointitulosten takaamiseksi on tärkeää, että sekvensoitaessa DNA-fragmenttien koko olisi mahdollisimman sama. Tässä opinnäytetyössä verrattiin kahta eri fragmentointitapaa entsymaattista ja ultraäänikäsittelyä. Fragmentointitavan täytyi olla hyvin toistettava ja vähän aikaa vievä. Ultraäänikäsittelyn todettiin olevan toistettavampi. Tästä jatkettiin DNA-fragmenttien koon jakautumisen pienentämistä. Tämä toteutettiin magneettipartikkeleilla. Kun halutut fragmenttikoot oli eristetty, DNA-kirjaston valmistaminen saatettiin loppuun. Sitten kirjastot sekvensoitiin Illuminan Miseq-sekvensoitilaitteella.

Sekvensoinnin jälkeen tehtiin data-analyysi.Ultraäänimenetelmällä saatiin hyviä ja toistettavia tuloksia. Ultraääni-menetelmän fragmenttikoko optimoitiin. Kun sekvensointidata oli kohdistettu, tuloksena oli hyvin epätasainen kattavuus, jota ei voitu selittää laboratoriossa tehdyllä virheellä

Avainsanat	Seuraavan sukupolven sekvensointi, Lynchin syndrooma, paksusuolen syöpä

Contents

Abbreviations

MGZ	Medizinisch Genetische Zentrum
NGS	Next Generation Sequencing
HNPCC	Hereditary Non-Polyposis Colorectal Cancer
MLH1	MutL Homolog 1
MSH2	MutS Homolog 2
MSH6	MutS Homolog 6
PMS2	Postmeiotic Segregation increased 2
DNA	deoxyribonucleic acid
SNP	Single Nucleotide Polymorphism
ssDNA	single stranded DNA
dsDNA	double stranded DNA
bp	base pair
UV	Ultra Violet
PCR	Polymerase Chain Reaction
MMR	Mismatch Repair
CRC	Colorectal Cancer
MSI	Microsatellite Instability
HPLC	High Performance Liquid Chromatography
EDTA	Ethylenediaminetetraaceticacid

1 Introduction

Sequencing is a very important method in medical diagnostics. Sequencing has developed a lot during recent year's. Different methods and more efficient platforms have been created which allow sequencing of thousands of DNA fragments at the same time. Also the data-analysis plays a big part in the massively parallel sequencing, because the amount of data created in each run is colossal. The data of each run is filtered so that the sequences outside the regions of interest are excluded. Then the reads can be mapped and the possible mutations examined.

Sequencing helps especially in diagnosis of inherited diseases like Lynch syndrome. Lynch syndrome is caused by a mutation in the MMR genes and the mutation can be passed down generations. The six MMR genes, which are affected in Lynch syndrome, are responsible for repairing sequence mistakes during DNA replication and a mutation in any of these genes may cause malfunction of MMR. Lynch syndrome causes endometrial, stomach, breast, ovarian, small bowel, pancreatic, urinary tract, liver, kidney and bile duct cancers.

This thesis was carried out at the Medizinisch Genetisches Zentrum Münchens (MGZ), Next-Generation Department. The goal of this thesis was to create a laboratory protocol for Next-generation sequencing. To find a robust method, two different fragmentation methods were tested and compared. Tests were carried out with samples chosen from MGZ patients existing storages. The patients chosen for testing had undergone Sanger sequencing but resulted negative. These patients either manifested a Lynch syndrome or they were hot candidates to manifest the disease because of their familiar inheritance. The target regions were the six genes responsible for Lynch syndrome MLH1, MLH3, PMS1, PMS2, MSH2 and MSH6. From these genes, the promoter area, intronic and exonic regions were covered. Also the intergenic areas after the genes were covered.

2 Sequencing

2.1 Next Generation Sequencing (NGS)

In sequencing the nucleic acids A (adenine), T (thymine), G (guanine) and C (cytosine) order in DNA or RNA strand is determined. Sequencing has become extremely important in medical diagnostics and research. In diagnostic it is used in preventive medicine. Patients undergo germline mutations screening and can be treated in early stage of disease. So far sequencing has accomplished a lot in research. For example the whole human genome was sequenced in the year 2000. [1.]

Next Generation Sequencing is based on the Sanger sequencing technology, also known as First generation sequencing. Next generation sequencing started developing when there came a need for larger number of data. Development of Next Generation Sequencing has been quite fast. The possibility to produce large number of data cheaply has become the biggest advantage of NGS. Next generation sequencing has also few other advantages over Sanger sequencing, known also as first generation sequencing. One of them is that in NGS adapters are ligated in the end of blunt end fragments. These fragments can be selectively amplified by PCR and there is no need for to amplify the fragment on a bacterial intermediate. NGS is also faster. NGS platforms need between 8 hours to 10 days, to complete a run and the data output varies between a couple of hundreds of reads and tens of millions of reads. In contrast to Sanger sequencing which achieves only 700 bp of around hundred reads and run time between 3 hours to 8 days. [2.]

Commercially there are many different kinds of NGS platforms, but in this thesis Illumina bridge amplification platform is only covered. The Illumina sequencing platform is based on an oligo-derivatized surface in a flow cell. The oligos on the flow cells surface are adapters. The DNA has to be prepared with blunt ends and adapters. These adapter-DNA fragments bind to the oligos on the surface (cf. Figure 1).

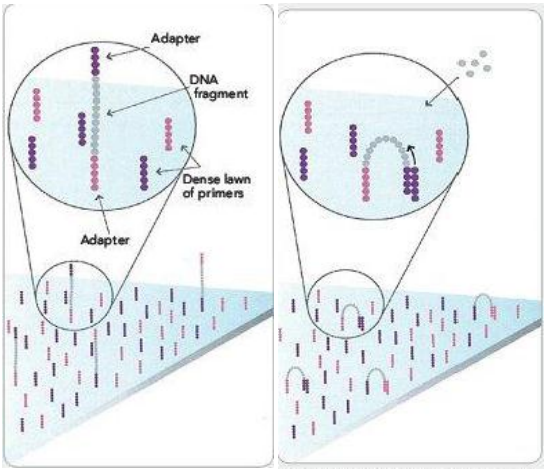


Figure 1 Bridge amplification at Illuminas sequencing platform [2]

When the DNA fragments are bound to the oligos, first a DNA polymerase amplifies the DNA strands and creates so called clusters (Figure 1). Amplification of the clusters is called bridge amplification. After creating the clusters by bridge amplification, labelled nucleotides are added with a polymerase, to make the reaction for sequencing. The nucleotides are labelled with base-unique fluorescent and in the label is a 3'-OH group, which inhibits the fluorescent (cf. Figure 2).

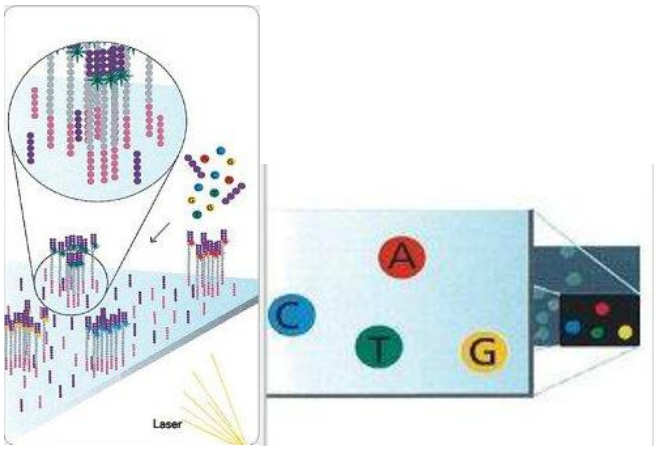


Figure 2 Imaging the DNA sequence [2]

The polymerase enzyme ligates the fluorescent labelled nucleotides in the clusters and the 3'-OH group detaches and the fluorescent reaction occurs (Figure 2). This reaction is detected with a laser. Every fluorescent labelled nucleotides has its own colour, due to these colours the sequencer is able to produce an image of the cluster. [2.]

2.2 On Target Sequencing

When the goal of research is not the whole genome but only a couple of genes, it is more time efficient and cost saving if the sequencing is done on target, because whole genome sequencing produces big amount of data. Data-analysis of big amounts of data also complicates the analysis. On target sequencing allows not only the investigation of a determinate number of genes, but also more patients in one run because the output of the run is not so big. Figure 3 shows the capture of target regions.

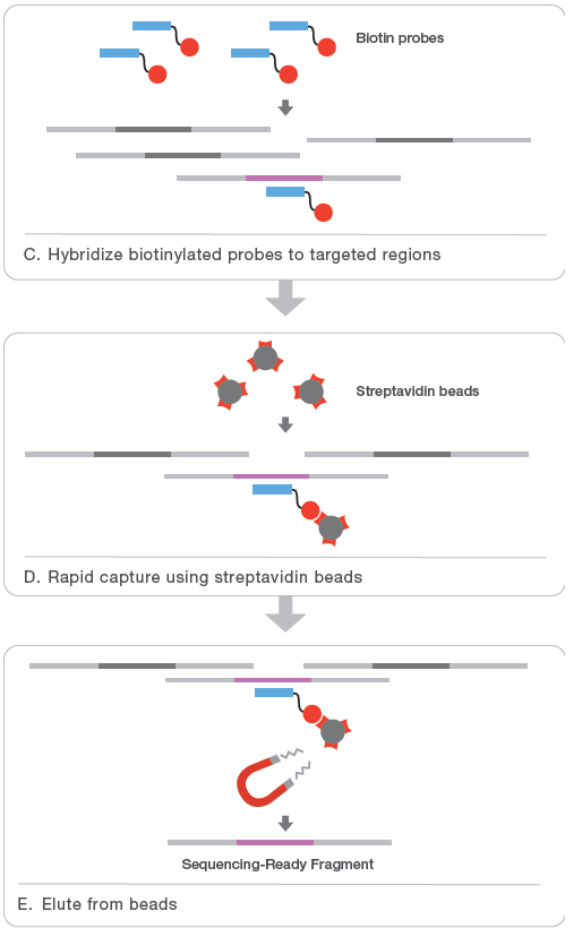


Figure 3 Capture of the target region by the streptavidin coated magnetic beads and biotin labelled probes, source: Illumina datasheet: Targeted Sequencing

For capturing the target region of a genome, probes, which are complementary to the wanted regions of the genome, are hybridized to the DNA (Figure 3). These probes are prepared with a biotin marker, which binds to the magnetic beads. The magnetic beads are used in-solution and are labelled with streptavidin. When the biotin marked DNA is mixed with the streptavidin labelled beads, the streptavidin and biotin makes a strong binding with each other, binding strength is around 1013 M-1. When the DNA binds to

the magnetic beads, the target regions can be isolated from the rest of the fragments with a magnet that immobilizes the DNA-capture beads complex and the unbound DNA can be washed away. Then, by breaking the DNA-bead binding, in the solution remains only DNA fragments with the wanted region of the genome. [3.]

2.3 Data-analysis

Data-analysis is the most difficult part in Next-generation sequencing because the amount of data produced in sequencing. During data-analysis sequence reads, produced during DNA sequencing, are aligned against the reference genome (GRCh37/hg19 human genome assembly). When the sequence data is mapped against the reference genome the coverage of the sequencing can be viewed. Coverage of a basepair is a term that refers to the number of sequence reads that can be mapped to the genomic position of that basepair (figure 4). The coverage is a result of the quality and quantity of the sequence. If the coverage is not even the SNP's, point mutations and structural variants can't be identified. Low quality reads will not mappe properly and the SNP's in the analysis can be thought as sequencing errors (cf. Figure 4).

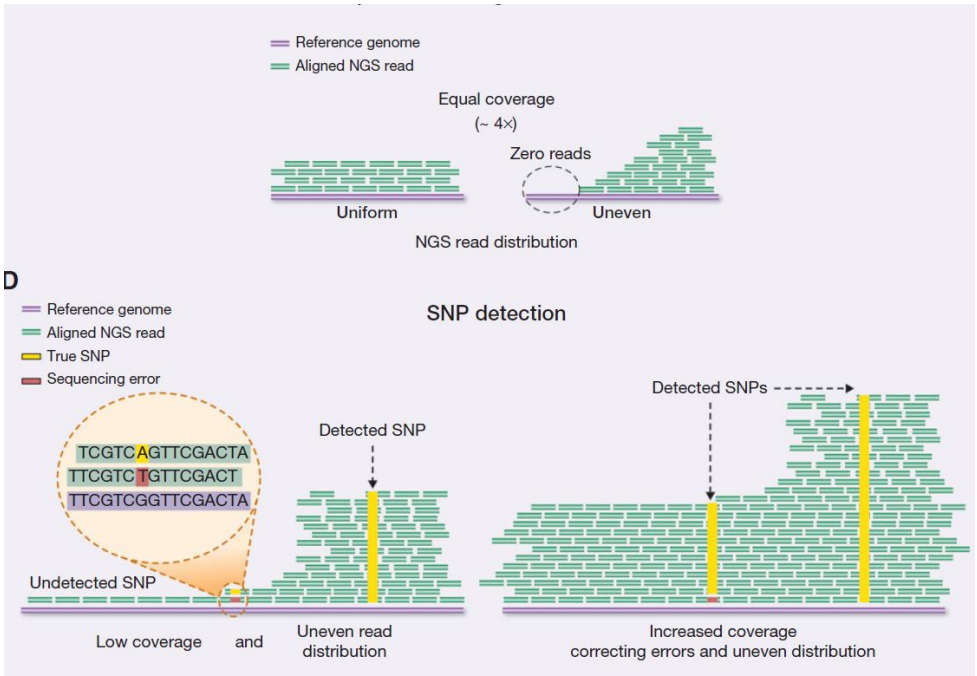


Figure 4 Coverage and discovery of the SNPs [4]

Since polymerases introduce also some errors (10^{-8} error rate) during sequencing, the sequences are never 100 % accurate and therefore it is important to have deep se-

quences and good coverage to prevent sequencing errors. The recommendations for coverage within a diagnostic setting are between 40 and 100 reads per basepair. The recommendations are depending on platform error rate and analytic sensitivity and wanted specificity. [4.]

3 DNA Library Preparation

3.1 Quality Management

Next Generation Sequencing and DNA library preparation are very sensitive methods for changes in the DNA concentration and fragment size. Because of this the DNA has to be controlled with different methods, to be sure that the quality and the quantity are in required limits. For measuring the quality and quantity Nanodrop device is used. Nanodrop is spectrophotometer, which uses UV/Vis absorbance for measuring the DNA. Also other measurements are used to get more accurate concentration and to examine the fragment size distribution.

3.1.1 PicoGreen

PicoGreen is a method to measure dsDNA in the solution. PicoGreen is based on fluorometric measurement. In the DNA solution is added a nucleic acid stain. The stain binds only the dsDNAs douple helix structure. When the stain is bound to DNA it emits light which can be detected as intensity of the light. The fluorescent intensity is comparable to the DNA concentration. The device used is called spectrofluorometer. To determine the unknown concentration from the sample, a standard curve with known concentrations has to be measured. The concentration of the samples can be calculated from the standard curves equation. The measurement is very reliable and it can measure up to 1 pg/µl of DNA. [6.]

3.1.2 Bioanalyzer

The Bioanalyzer is a device that measures the fragment size and molarity of a DNA sample. Bioanalyzer is specific for dsDNA and it can’t measure ssDNA. The Bioanalyzer uses a microfluidic platform. In this platform the flow of fluids is controlled in submil-

limeter dimensions. In the microfluid platform only 1 µl of the DNA solution is needed. The platform also uses electrophoresis. Electrophoresis is a technique where the DNA fragments are separated on a gel by to their size (bp). To make a more accurate definition of the fragment size, than just comparing the result to a ladder, each well has an upper marker and a lower marker. In the bioanalyzer the DNA fragments are stained with fluorometric stain and when the run is ready the molarity of the DNA solution can be determined by the fluorometric intensity. Bioanalyzer has different kits for different concentration of DNA. High Sensitivity is used when the concentration is low (<15 ng/µl) and when the concentration is higher (>20 ng/µl) 1000 DNA chip is used. In one High Sensitivity chip 11 samples can be run and in 1000 DNA chip 12 samples. The bioanalyzer gives as a result a graph were on x-axis is the basepair size of the fragment and on y-axis is the fluorescent unit (FU). [7.]

3.2 DNA Purification

Some times after DNA extraction or storing DNA long time, the DNA has not enough good quality. For improving the quality of the DNA, this is purified from the impurities that might have end up into the solution. By binding the DNA on a membrane the washing of impurities can be done (cf. Figure 5).

Figure 5 DNA purification proces [8]

During the purification process the DNA is bound to EDTA and then transferred into a column. The filtrate membrane in the column is silica, in which the DNA/EDTA complex binds. When only the DNA is bound to the silica membrane the other particles can be washed away with ethanol (Figure 5). Then the silica is dried. When all the ethanol has

been dried, the DNA is eluted in HPLC grade water. The method is really rapid and it results in high quality DNA. Purification kit is provided by Zymo Research. [8.]

3.3 DNA Fragmentation

Fragmenting DNA is one of the most crucial steps in library preparation. The read length is normally 300 bp or 500 bp. The read length can be chosen to what is closest to the DNA fragment size. The read is done paired end. Paired end reading means that the DNA strand is read once from up (150 bp or 250 bp) and once from down (150 bp or 250 bp) during sequencing. The smaller the distribution of the fragment size is the better sequencing results are reached. Too long fragments cause a gap between the paired end reads and too short fragments causes an over lapping of the reads. Both cases may cause challenges in data-analysis.

3.3.1 Ultrasonication

In ultra-sonication the DNA is cut randomly by ultrasonic waves. In sonication the ultrasonic waves are focused in a water bath to the sample vessel, as shown in Figure 6. The ultrasonic waves come in bursts to sample. Every bursts creates little air bubbles called cavitation bubbles.

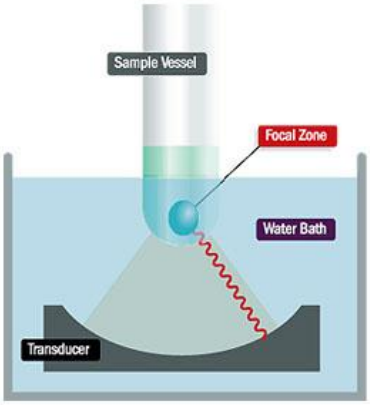


Figure 6 Ultra-sonication by Covaris [9]

These cavitation bubbles collapse in the end of every burst. When the bubble collapses, it creates little high velocity jets. The high velocity jets hits the DNA with such a power that it cuts the bonds in the DNA strand. The ultra-sonication treatment is continued until the DNA is fragmented to the wanted fragment size. The ultra-sonication leaves the DNA fragments in to overhangs so the ends have to be prepared for later

steps required in NGS. Due to the steps which have to be done after the fragmentation, ultra-sonication is more time consuming than the more traditional enzymatic fragmentation. Ultra-sonication is a good method because it cuts extremely randomly and it is also a very robust method. By variations of the run parameters different length fragments can be made (cf. Figure 7). [9.]

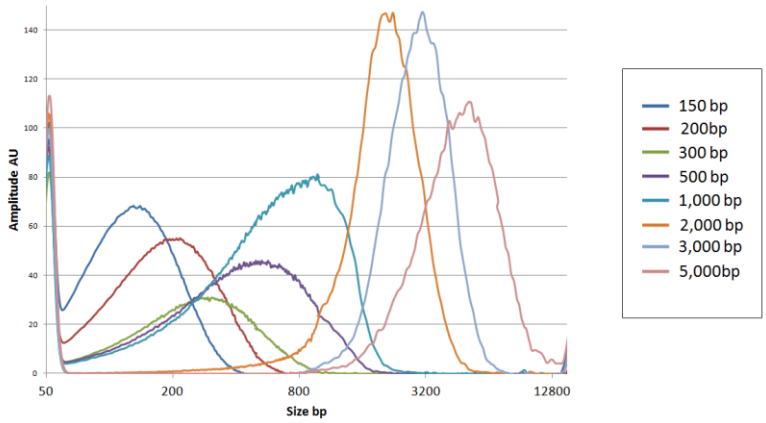


Figure 7 The scalability of ultra-sonication [9]. In the graph x-axis is the fragment size (bp) and on the y-axis is the amplitude unit (AU). The peaks width is the distribution of the fragment size.

3.3.2 Enzymatic fragmentation

In the enzymatic fragmentation an enzyme cuts the DNA. Normally a mixture of different enzymes is used, this is necessary because every enzyme has its specific cutting points on the DNA strand. Enzymes require very well adjusted conditions. If the conditions are not exactly the ones that the enzyme needs or even little bit varies from what they should be, it has a very big effect on the effectiveness of the enzyme. The enzyme fragmentation method is not robust and can bring easily different kind of results on each fragmentation time.

Tagmentation is called when a transposomes are used in the fragmention. Transposomes are enzymes that cut the DNA into 300 bp long fragments, if the reaction is successful and ligate a specific DNA sequence to the ends of the fragments (Figure 8). The sequences are called Read 1 Primer and Read 2 Primer, these sequences are always used when sample is prepared for the MiSeq-sequencer.

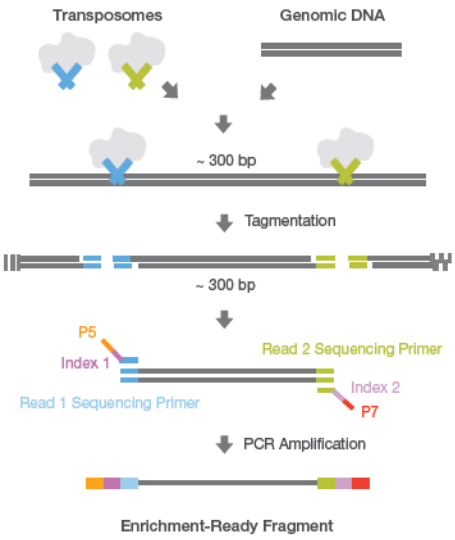


Figure 8 Transposome enzyme fragments and tags the DNA [10]

This proces is called tagmentation. Tagmentation is a very fast and easy proces because in one step the DNA is fragmented, the adaptors are ligated and the traspo-somes leaves blunt ends. [10.]

3.4 Size Selection with magnetic beads

Size selection helps to achieve a more specific fragment size. In the size selection with magnetic beads, the DNA fragments bind on a metallic particles surface. The DNA binds on the magnetic bead because on the beads surface is high positive charge and the highly negatively charged DNA strand is bound to the bead by the charge difference. On the magnetic beads the bigger fragments bind faster than the shorter ones because of their higher charge. With putting less or more magnetic beads in relation to the DNA fragments to the solution, the certain sized fragments can be selected from the solution. With use of a magnet the magnetic beads containing the DNA fragments can be pulled aside while the other fragments are discarded. Then the beads are washed with ethanol to be sure there won't be any unwanted fragments. [17.]

3.5 Indexing

In Next Generation Sequencing multiple samples are sequenced simultaneously. To identify and separate each patient sequences for analysis, sequencing indices are used. Indices work like barcodes. The indices are 6 or 8 nucleotides long, for example an index from New England BioLabs is ATCACG. These indices are ligated in each patients DNA samples fragments and each of indices are unique. When the sequencer reads the index, it can connect the sequencing result to the right patient. The use of multiple indices in one run is called multiplexing. The indices allow pooling of DNA of several patients simultaneously and therefore high throughput sequencing can be done. An index work also as a primer, the ligation of the indices happens during PCR when the wanted size of DNA fragments is amplified, in the annealing reaction. The indices are provided by Illumina or New England BioLabs, in the DNA library preparation kits and every kit has between 12 and 96 indices. [11.]

4 Lynch Syndrome

Every year over 1 million people in the world gets colorectal cancer (CRC), 3 % (~30700) of them have Lynch syndrome, also known as hereditary non-polyposis colorectal cancer (HNPCC). That means every year there are 28600 new cases of Lynch Syndrome diagnosed. Lynch syndrome is an inherited cancer of the digestive tract. Lynch syndrome causes endometrial, stomach, breast, ovarian, small bowel, pancreatic, urinary tract, liver, kidney and bile duct cancers. Most of the females with Lynch syndrome do manifest endometrial cancer. Lynch syndrome has an early onset age of about 45 years. [12.]

4.1 Diagnosis

Lynch syndrome is a genetic condition, which means that testing of germline mutations if used for diagnosis. Because Lynch syndrome lacks specific phenotypic features not everybody with colorectal cancer are tested for Lynch syndrome. Patients' family history is looked when thought of genetic testing. If in patients family history is found other CRC cases and the patient is fairly young person diagnosed with CRC, the patient usually goes to MSI testing.

Microsatellite instability (MSI) is a method where the numbers of nucleotides in the DNA strand are measured (cf. Figure 10). The DNA strands are marked with five different markers (BAT25, BAT26, D2S123, D5S346, and D17S250).

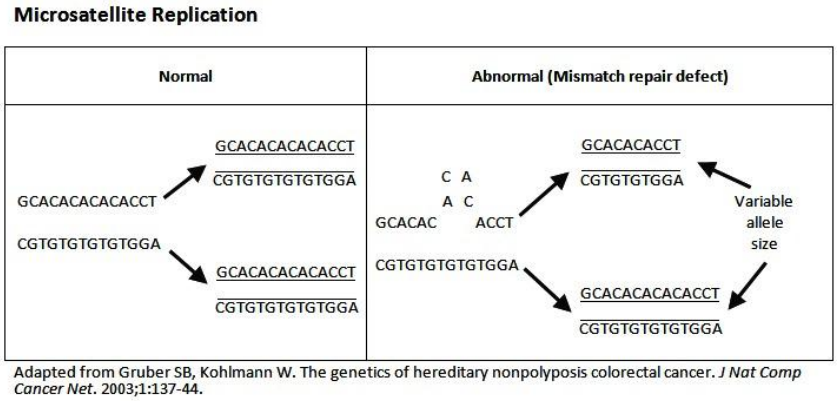


Figure 9 The MSI labelling [13]

In this test, tumor tissue and normal tissue are compared (Figure 10). The number of microsatellite nucleotide repeats is calculated from the tissues. The MSI is high when more than 30 % of markers show instability. When less than 30 % of markers show instability the MSI is low and when 0 % the MSI is stable.

To help the diagnosis criteria called the Amsterdam criteria I was developed. Amsterdam criteria I includes that 3 relatives have been diagnosed with CRC and one of them first degree relative, 2 generations of family line are affected, 1 relative have had a diagnosis before age 50 and familial adenomas polyposis is not an option. If the Amsterdam criteria's are filled and tumour is MSI-high, it is recommended to take part in a germline mutation testing. If mutations are found, the tested patient should attend endometrial screening annually from age 30 on. If the patient filled all the Amsterdam criteria 1 conditions, the patient has 82 % lifetime risk to develop colon cancer. [12.]

4.2 Mismatch repair system (MMR)

Mismatch repair (MMR) is a system in cells to prevent mutations to occur in the daughter cells, after parent cells division. The mutations springs from insertions and deletions that take place in DNA replication, because the DNA polymerase enzyme is not 100 %

accurate. MMR is a mechanism to prevent damage in the cell's DNA. Six genes are responsible of the enzyme complex which builds the MMR. The enzyme complex moves along the DNA strand after DNA replication. When it detects mismatches, it removes the wrong nucleotide and leaves a gap in the strand. Then the MMR fills the gap with a right nucleotide and ligates it in its place (cf. Figure 11). [11.]

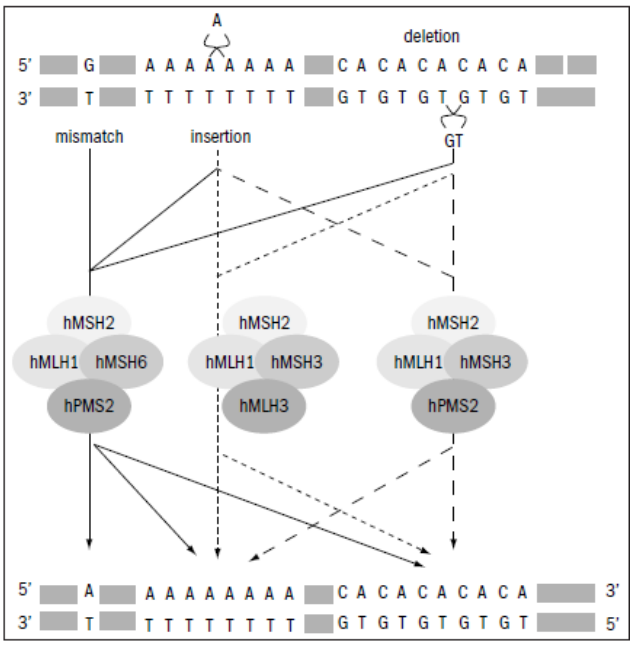


Figure 10 The MMR proces [13]

Mutations in any of the genes responsible the enzyme complex can cause malfunction of MMR. [12.]

There are six genes responsible of mismatch repair MSH2, MSH6, MLH1, PMS2, PMS1 and MLH3. The four first genes make the enzyme complex. These four genes can be divided in two smaller complexes MSH2-MSH6 and MLH1-PMS2. MSH2-MSH6 heterodimer is responsible for the detection of 1 or 2 unpaired nucleotides. It can also detect larger insertions and deletions. Sizes of the genes are in the table 1.

Table 1 MMR genes size table

Gene	Coding exons	Amino Acids	Size
MLH1	19	756	57 497 bp
MSH2	16	934	80 162 bp
MSH6	10	1360	23 872 bp
PMS2	15	862	35 868 bp

Most pathogenic mutations are found in MLH1 (50 %) and MSH2 (40 %), the rest of the genes are responsible for only 10 % of the mutations. The heterodimer MSH2-MLH1 is responsible for 64 % of all germline mutations. [11.]

5 Mutations

Inherited diseases like Lynch syndrome are caused by mutations in the genome of the patient. The mutations can be inherited or the mutations can occur *de novo* in the child's genome. Mutations are nucleotide changes in the DNA strand. Mutations happen when the DNA replicates. Replication stands for a process where one DNA molecule is copied into two identical DNA molecules with the help of a polymerase enzyme. The mutations are mistakes of the polymerase enzyme. There is a bright spectrum of mutations that can occur during DNA replication but the most common ones caused by the DNA polymerase are point mutations and insertions or deletions of single nucleotides.

The DNA codes for proteins. Every three nucleotides of the DNA strand within a gene form a codon. Each codon is translated in one amino acid and a raw of amino acids form a protein. There are different mechanisms how a mutation at the DNA level can lead to a misfunctional protein and cause a disease state. A deletion or an insertion can change the codons so that it results in a stop codon and a shortened protein. Nonsense mutations are when the change in the genetic code changes a codon into stop codon and not to another amino acid. A missense mutation occurs when the change in the sequence codes a different amino acid. Silent mutations mean a change in the genetic code that does not change the amino acid. Some mutations, called splice site mutations, occur within the bases that narrow the coding exons by changing the splicing process of the DNA into RNA, which means that the protein sequence changes.

6 Materials and methods

6.1 Materials

All the reagents and kits used in the experiments are listed below:

Table 2 Used kits

Kit
NEBNext Ultra DNA Library Prep Kit
TruSight Rapid Capture
TruSeq Enrichment
NEBNext Adaptor for Illumina
MiSeq v2 reagent kit

Table 3 Reagents not provided in the kit

Reagent	Supplier
Axygen magnetic beads	Axygen
AMPure XP magnetic beads	Beckman Coulter
PicoGreen	Life Technologies

All the reagent kits (Table 2) are available commercially.

6.2 Samples

DNA samples used in the tests, were from MGZ storages. All the samples were 1-15 years old. They were distracted from blood with commercial kit FlexiGene DNA. All the samples were stored at 8 ˚C. Patients were chosen, from patient who had suffered of Lynch syndrome but had negative results in Sanger sequencing. These samples were considered as the hot candidates for sequencing the introns and promoter area. Patient cohort was 47 where 17 MSH2, 20 MSH6, 5 MLH1 and 5 PMS2. To compare the two fragmentation methods 4 samples were done with ultra-sonication and tagmentation.

Some of the patient samples had to be purified, because they had bad results at absorbance 260/280 and 260/230 which indicated a poor DNA quality. For purification a commercial kit Clean and Concentrator-5^{TM} from Zymo Research was used.

To determine the accurate concentration of the DNA, PicoGreen was used. For standard curve 8 samples were prepared: 100 ng, 50 ng, 25 ng, 12,5 ng, 6,25 ng, 3,18 ng, 1,7 ng, 0 ng. The samples result placed to the standard curves equation (1), to calculate the sample concentration:

(1)

$$C_{sample} = mx + c$$

Where C_{sample} = Concentration of the sample, m = slope, x = independent variable, c = y intercept on the line.

6.3 Target Selection and Design

Normally in the enrichment kit is ready designed oligos for the target selection but in this project only the four major genes from Lynch syndromes were taken under examination and the oligos had to be designed separately in software provided by Illumina. The regions of these genes were taken from the human genome assembly GCRh37/hg19 by using the UCSC (University of California, Santa Cruz) Genome Browser (cf. Figure 12).

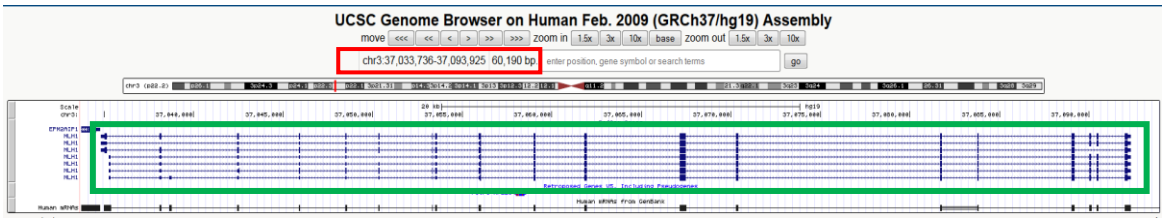


Figure 11 The MLH1 gene with chromosomal position (marked with red), intronic and exonic regions (marked with green). Exonic regions are the thick blue lines. Every line, in the area marked with green, represents an isoform of the MLH1 gene. [http://genome-euro.ucsc.edu/cgi-bin/hgGateway?redirect=auto&source=genome.ucsc.edu]

From UCSC the chromosomal position (figure 12) and the nucleic region of the genes was calculated and then the chosen regions were given into the Illumina's design studio-software where the capture oligos could be designed. In the design studio the target area (chromosome, start and end nucleotide) is given and then the design studio calculates the area that is covered and places small DNA oligos that will be used for capturing the genomic regions of interest. If the result does not fulfil the request of the user, it is always possible to add more oligos to some part. These oligos are called custom selected oligos.

6.4 Workflow

In the workflow (cf. Figure 13) the most important steps that were under survey, were fragmentation and size selection.

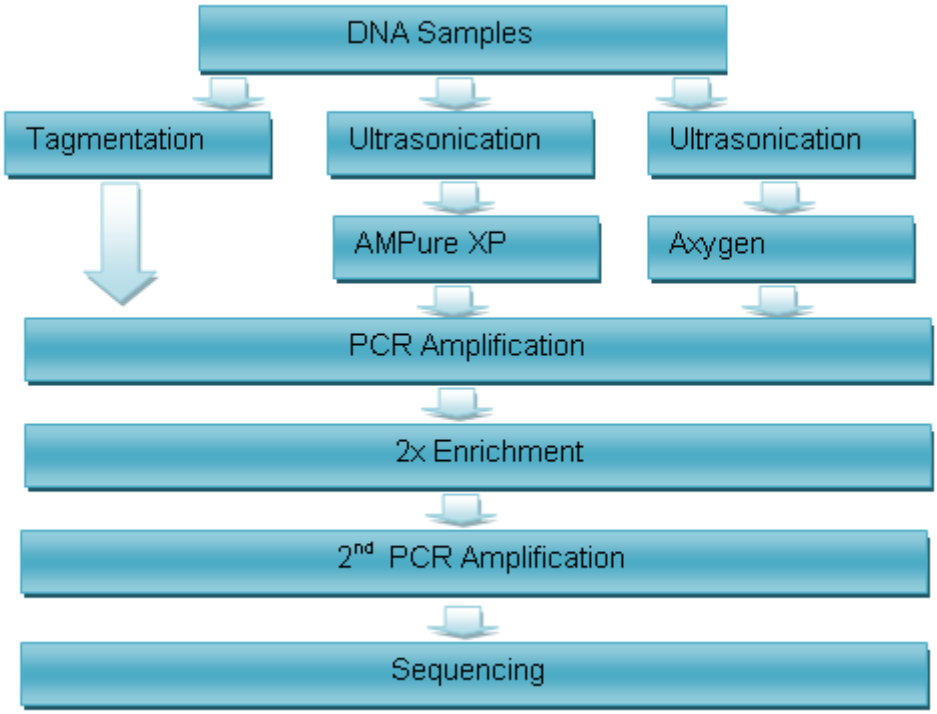


Figure 12 Workflow

6.5 First Test Tagmentation

Before the tagmentation the DNA concentration and quality of the DNA was measured with Nanodrop and then diluted. After dilution was measured with PicoGreen which is a method, that measures double stranded DNAs concentration in 1xTE solution. After PicoGreen the concentration was adjusted to 2,5 ng/µl.

Fragmentation was done by tagmentation enzyme. Tagmentation enzyme is from Illuminas commercial kit TruSight Rapid capture. Tagmentation enzyme fragments the DNA and also ligates adapter in the fragments. Conditions for the tagmentation: 10 minutes in 58 ˚C, with heated lid 100 ˚C. The cycler used was Biorad DNA engine Tetrad 2.

Undertagmented DNA was observed at the Bioanalyzer control check, so the tagmentation concentration was lowered to half 1,25 ng/µl. The time and temperature were kept at the same.

The results were examined with Bioanalyzer 2100 High sensitivity chip. The next step was amplification of tagmented DNA short fragments by PCR, where also the indices (cf. Table 4) were ligated to the fragments. Into the tagmented DNA was pipetted 20 µl of Nextera Library Amplification Mix and 5 µl of index E502 and 5 µl of one index N.

Table 4 Indices from Illumina

Illumina	
Index	**Sequence**
E502	CTCTCTAT
N701	TCGCCTTA
N702	CTAGTACG
N703	TTCTGCCT
N704	GCTCAGGA
N705	AGGAGTCC
N706	CATGCCTA
N707	GTAGAGAG
N708	CCTCTCTG
N709	AGCGTAGC
N710	CAGCCTCG
N711	TGCCTCTT
N712	TCCTCTAC

For PCR program a three step PCR from the protocol TruSight rapid capture was used. The PCR program:

Table 5 PCR program for Illumina TruSight Rapid capture kit PCR

Temperature	**Time**
72 ˚C	3 minutes
98 ˚C	30 seconds
98 ˚C	10 seconds
60˚C	30 seconds
72˚C	30 seconds
Cycle from step 3 10 times	
10 ˚C	forever

After the PCR the result was examined with Bioanalyzer and PicoGreen.

6.6 Second Test Ultra-sonication

DNA was measured with Nanodrop to examine the DNAs quality and concentration. A dilution was made based on Nanodrop result and then measured with PicoGreen. Then the DNA was diluted to its final working concentration of 20 ng/µl and volume 50 µl.

Fragmentation was done with an ultrasonicator, Covaris M220. When fragmenting with Covaris it is important that the water bath temperature stays as stable as possible. The water temperatures changes may have some small effect in the sharing resulting in uneven DNA fragment lengths. Parameters used for sharing were taken from Covaris M220 protocol with target fragment size 400 bp. Following parameters were used:

Table 6 Covaris M220 parameters for target fragment size 400 bp

Parameter	Value
Peak Incident Power (W)	50
Duty Factor	20 %
Cycles per Burst	200
Treatment time	50
Temperature (˚C)	20
Sample volume (ml)	50

For the treatment DNA was pipeted in microTUBE AFA Fiber with screw-cap. After the treatment with Covaris, the samples were examined with Bioanalyzer 1000 Chip.

6.6.1 End Preparation and Size Selection

After ultrasonication the DNA fragment ends are overhangs. So that the sequencing adapters can be ligated the ends have to be prepared in to blunt ends. For end preparation a commercial kit from New England Biolabs (NEB) was used. The kits name was NEBNext ultra Library Prep Kit for Illumina.

The blunt ends were cut with an enzyme that cuts only single stranded DNA. Once the DNA has blunt ends, A-tailing is done by using an enzyme that ligates A-tale to the both ends of the fragments. After this step the Illumina adapters can be ligated. In the fragmented DNA 3,0 µl of End Prep Enzyme Mix and 6,5 µl of End Repair Reaction buffer were added. The solution was incubated 30 minutes at 20 ˚C and 30 minutes at

65 ˚C. Then 15,0 µl of Blunt/TA Ligase Master Mix, 2,5 µl of NEBNext Adaptor for Illumina and 1,0 µl of Ligation Enhancer was added in to the mixture.

After the DNA end preparation, the DNA fragment size required for sequencing is selected. For size selection AMPure XP beads were used and the size selection was done like in the NEB protocol recommended with a target size of 320 bp. In the first step adding 55,0 µl AMPure XP beads. Then incubating the solution at room temperature 5 minutes and then incubating at the magnetic rack for 2 minutes. The supernatant was transferred into fresh tube. Then 25 µl of the beads were added and again incubated 5 minutes in room temperature and 2 minutes on the magnetic rack. The supernatant was discarded and then washed with 80 % EtOH twice. The beads were dried for 15 minutes and then eluted in 10 mM pH=8,0 Tris-HCl.

After the size selection, the samples were amplified with PCR. The protocol used for PCR was from NEBNext. To the size selected DNA 25 µl High fidelity PCR Master Mix and 1 µl of each primer (index and universal PCR primer) were added (cf. Table 7).

Table 7 Indices from New England BioLabs kit

New England BioLabs	
Index	**Sequence**
Index 1	ATCACG
Index 2	CGATGT
Index 3	TTAGGC
Index 4	TGACCA
Index 5	ACAGTG
Index 6	GCCAAT
Index 7	CAGATC
Index 8	ACTTGA
Index 9	GATCAG
Index 10	TAGCTT
Index 11	GGCTAC
Index 12	CTTGTA

The number of cycles chosen was 10 because the protocol recommended that when using 1 µg of starting amount of DNA. The used PCR program:

Table 8 PCR program for New England BioLabs kits PCR

Temperature	**Time**

98 ˚C	30 seconds
98 ˚C	10 seconds
65 ˚C	30 seconds
72 ˚C	30 seconds
Cycled from step 2 10 times	
72 ˚C	5 minutes
4 ˚C	forever

After the PCR the fragment size was ~320 bp.

6.7 Third Test Size Selection

The sample preparation, fragmentation and end preparation were done the same way as in 6.6.

After the end preparation purification was made with AMPure XP Beads. Adding 86,5 µl of beads and incubating for 5 minutes. Elution from the beads was done with 50 µl HPLC grade water. For size selection different type of magnetic beads was used, Axygen size select magnetic beads. For using the Axygen beads the size selection had to be optimized from the Axygens official protocol. The Axygen beads size selection works in two steps. First step removes the big fragments and second step the small fragments. The first steps ratio is the more important, so different ratios were tested: 0,2x, 0,4x, 0,6x, 0,8x, 1,0x and 1,3x (cf. Figure 14).

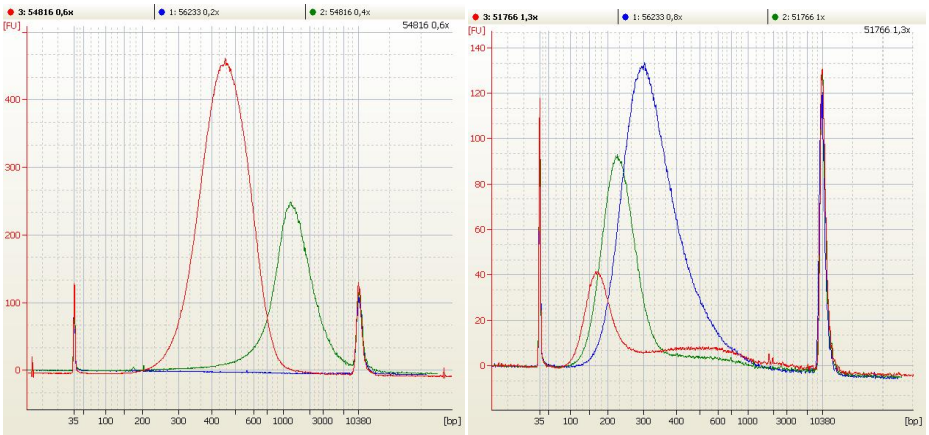


Figure 13 Optimization of the Axygen beads. Left side graph: Red line 0,6x, green line 0,4x, blue line 0,2x. Right side graph: red line 1,3x, green line 1,0x, blue line 0,8x. From the graph the fragment size (bp) distribution can be observed. The wanted result was right side graph blue line ~320 bp.

Results were examined with Bioanalyzer and then 0,7 x was decided because the wanted fragment size 320 bp, was between the results from 0,6x (cf. Figure 14 right side graph red line) and 0,8x (cf. Figure 14 left side graph blue line). All the samples were done with the ratio 0,7 x of Axygen beads. The fragment size was ~ 350 bp. For the elution 28 µl of 10 mM 8,0 pH Tris-HCl was used.

After size selection, the samples were amplified with same primers and PCR program as in part 6.6.2. For the PCR the number of cycles was increased to 12 to make sure enough product would amplify. After the PCR the fragment size was ~410 bp. The same primers were used as in 6.6.2.

The next batches samples were also size selected with Axygen beads but the ratio was optimized so that the fragment size would go down to ~280 bp. The first ratio was changed to 0,8x. To go as close as possible to the preferred 300 bp fragment size after the PCR and index ligation. The number of cycles was lowered this time to 11, to minimise the amplification of big unwanted fragments, which were observed with 12 cycles. Otherwise the PCR was done the same way as in 6.6.2. For the PCR Eppendorf Nexus thermocycler was used.

The PCR results was examined with Bioanalyzer High sensitivity chip with dilution 1:10. Also the concentrations were measured with PicoGreen.

6.8 Pooling

The DNA concentration was determined with PicoGreen measurement. From the results it was calculated 500 ng each sample and 9 - 12 samples were pooled. All the pools that had bigger volume than 40 µl, were concentrated with vacuum concentrator. Vacuum concentrator works at rate 10 µl/15 min, lowering the volume.

6.9 Enrichment

In the pool were pipetted 10 µl of custom selected oligos, which were designed for the enrichment, and 50 µl capture target buffer. Then the mixture was placed in thermal cycler for 20 hour incubation in 58 ˚C with heated lid at 100 ˚C. For enrichment step

Illuminas TruSeq Enrichment kit was used. After the incubation the oligos bound to the fragments containing the regions of interest were captured with streptavidin beads. When the fragments were bound to the steptavidin beads a wash was preformed. The wash was done in accordance of a protocol from Illumina: TruSeq DNA enrichment.

After the hybridization steps the fragments, containing the region of interest, were amplified with PCR. 25 µl of PCR master mix and 5 µl of PCR Primer cocktail, from TruSeq Enrichment kit, were pipetted into the pool. The PCR program used was from Illuminas TruSeq Enrichment protocol. The PCR program was altered with 1 extra cycle for the last two pools because after the PCR lots of primers were left in the sample. The PCR program:

Table 9 PCR program for Illuminas TruSeq Enricment

Temperature	Time
98 ˚C	30 seconds
98 ˚C	10 seconds
60 ˚C	30 seconds
72 ˚C	30 seconds
Cycle from the step 2 13 times	
72 ˚C	5 minutes
10 ˚C	forever

The primers were shown in the Bioanalyzer result, next to the lower marker peak.

6.10 Sequencing

For the sequencing the libraries were denaturated, diluted to 10 pM and PhiX is added, as an internal DNA control. Phix and used reagents, like HT1, are provided in Illuminas MiSeq v2 reagent kit. First the PhiX was denaturated and diluted with 0,2N NaOH, and 0.5 µl of 0,4 nM PhiX library was mixed with 5 µl of 0,2 N NaOH. Incubation of 5 minutes was necessary to denaturate the DNA strands. The 10 µl of PhiX was diluted in 990 µl pre-chilled HT1. The DNA library was diluted to 2 nM in Tris-HCl. 10 µl of diluted DNA library was then mixed in 10 µl 0,2 N NaÒH. The Mixture was incubated for 5 minutes to denaturate the DNA. The denaturated library was diluted into 20 pM by adding 980 µl pre-chilled HT1.

Prepared PhiX and DNA library were combined. 495 µl of DNA library and 5 µl PhiX control. Then the concentration was diluted to 10 pM and DNA libraries were sequenced in an Illumina MiSeq system.

7 Results and discussion

7.1 Comparison of Two Fragmentation Methods

The first test with Illuminas tagmentation enzyme was not working at all like wanted. It resulted with too big average DNA fragment size. The wanted fragmentation size was 300 bp and the result was closer to 1000 bp indicating the undertagmented DNA (cf. Figure 15, graph A). When the concentration was lowered to 1,25 ng/µl, the tagmentation was still not working consistently, but produced acceptable results with an average fragment size of 300 bp in the best case (cf. Figure 15, graph B) and in the worse result the fragment size ranged between 200 bp and 1000 bp (cf. Figure 15, graph C). In general the fragments were bigger than wanted, but the optimization of the method was not possible.

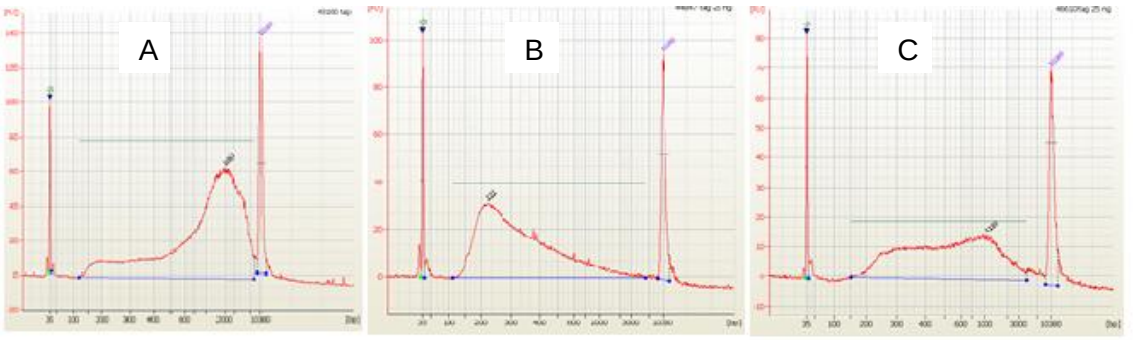


Figure 14 Tagmentation results: Graph A basepair peak ~1000 bp, graph B peak ~ 300 bp, graph C fragment size distribution is too big 200 – 1000 bp. The result is from Bioanalyzer High Sensitivity Chip.

The PCR, after tagmentation, worked well but the problem was the big fragments from the poorly tagmented samples which were also amplified during PCR (Figure 16 B). The PCR ended up with big variation between the samples as seen in Figure 16. If the figure results are compared the sift in the size can be easily seen. This is an indicator for how the PCR worked. From the sift towards the bigger fragments can be seen, that the indices had ligated properly. The PCR produced a lot of product.

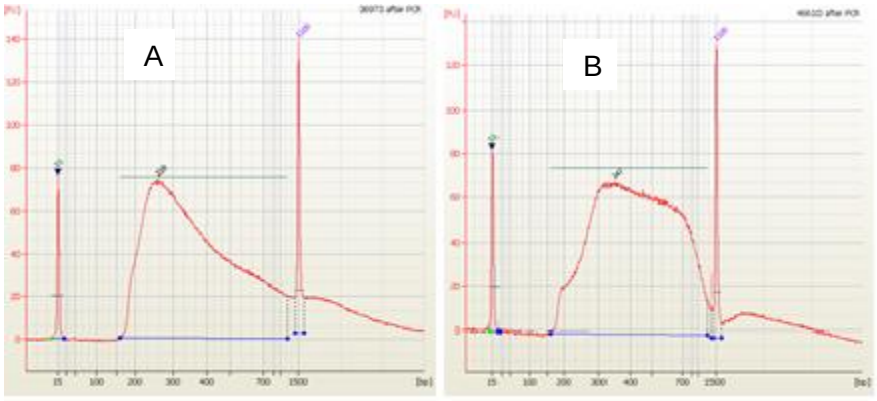


Figure 15 PCR after tagmentation. Graph A the PCR has amplified the fragments ~300 bp and in graph B the PCR has amplified a lot all the fragment sizes. The amplification of DNA fragments can be observed from the graph y-axis, FU is equal to concentration. The result is from a Bioanalyzer 1000 DNA chip.

When fragmented with Covaris the protocol worked without any alternations and the method was very consistent (cf. Figure 17). The results were reproducible for all of the 47 samples. The result was fragments between 100 bp and 1000 bp. The average was around 450 bp.

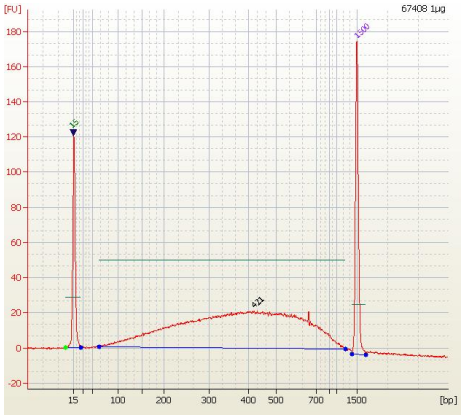


Figure 16 Exmaple of a fragmentation result, from the cohort of 47 patients, after Covaris treatment. The fragment size distribution is 100-900 bp. The result is from a Bioanalyzer 1000 DNA chip.

End preparation and size selection worked also without alternations to the protocol. The fragment size resulted in around 300 bp as was wanted. After the main DNA peak, a minor peak of bigger DNA fragments between 800 bp and 1000 bp (cf. Figure 18) was observed.

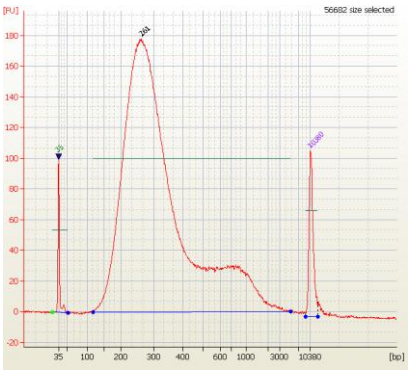


Figure 17 Size selection example result after AMPure XP beads. After the size selection the peak of fragment size at ~300 bp. The result is from a Bioanalyzer High Sensitivity chip.

After the PCR the peak sifted with 50 bp forward which indicates that index ligation at the end of the DNA fragments by PCR was successful. In Figure 19 can be seen how the unwanted big fragments had amplified more together with the rest of the DNA sample, because of the slight peak (marked with blue in Figure 19) after the main peak. The end amount of the DNA was not as high as expected but enough for the next enrichment step.

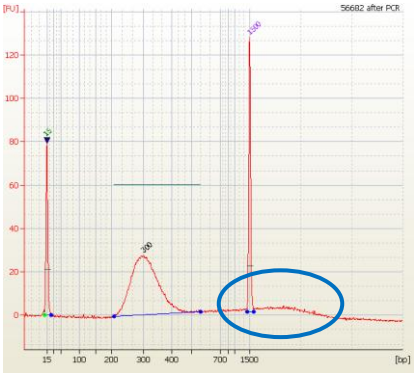


Figure 18 Example of a PCR result after the size selection with AMPure XP beads. The main peak is at ~300 bp. After the main peak is a low peak ~1500 bp. The low peak indicates of big fragments are present in the sample. The result is from a Bioanalyzer 1000 DNA chip.

After pooling and enrichment of the samples the results of the two methods differed from each other. The samples that were enzymatically fragmented showed a fragment size between 200 bp and 10 000 bp range. In the method with Covaris the size of the fragments was around 320 bp, although after the main peak is a second peak indicating the unwanted fragments was observed. Also from the results (cf. Figure 20) can be noticed two peaks next to the lower marker (Figure 20 A marked with blue). These peaks show that in the solution is still lot of primers left after the second PCR.

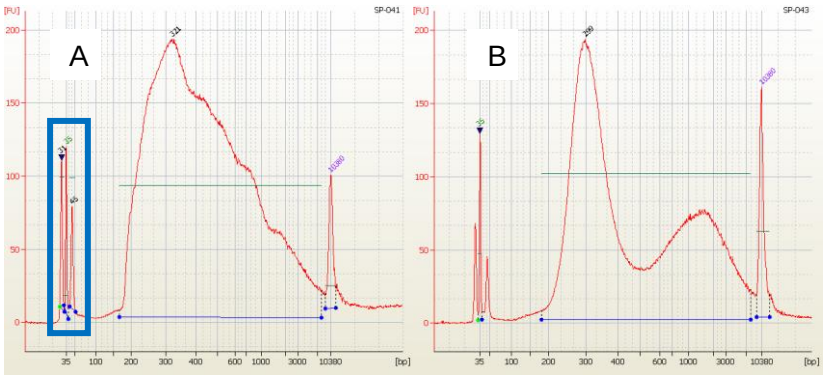


Figure 19 Sequencing ready pools 1 and 2. The pools were done with two different methods. The result A is result from the entzymatically fragmented and enriched DNA library. The result B is result with ultra-sonication fragmented and enriched library. The fragment size distribution is smaller than in result A, but still lot of big fragments present. The result is from a Bioanalyzer High Sensitivity chip.

After sequencing, coverage of on target reads results show big differences (Table 10). Despite the over amplification of Covaris tagmented samples gives an average of ~16 % more reads on target than the method preparation with the Nextera tagmentation enzyme (Table 10).

Table 10 Reads on target results from the 4 comparison patients

Patient	Nextera on target	Covaris on target	Difference
64921	20,63%	36,80%	16,17%
67408	20,77%	37,05%	16,28%
68929	20,52%	38,03%	17,51%
66012	21,87%	38,71%	16,84%

7.2 Optimizing the Size Selection

For size selection beads, were changed from AMPure XP beads to Axygen beads. The Axygen beads were optimized from the protocol. The Axygen beads selected the size of the DNA with a very clean peak and discarded all DNA fragments that were too big for further oligo-hybridization and capture. In the graph no big unwanted fragments are observed after Axygen bead size selection. The average fragments size was ~350 bp (cf. Figure 21). The fragment size grew to ~400 bp after PCR indicating correct ligation of indices by PCR method.

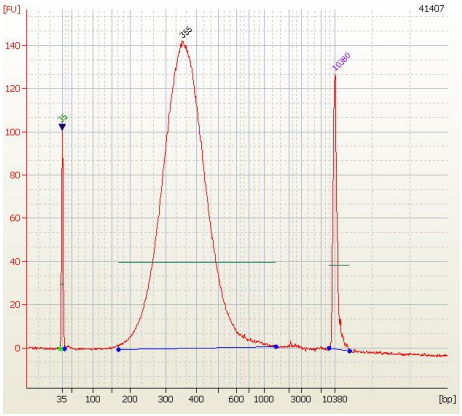


Figure 20 Axygen beads size selection target size 350 bp. The size selection very successful because only one peak present. The result is from a Bioanalyzer High Sensitivity chip.

For the next pool, the fragment size was even more optimized so that after the PCR the fragment size would be ~300 bp. Results (cf. Figure 22, graph A) showed a perfect optimatizion to ~280 bp DNA fragment size by using a bead/DNA ratio 0,8x. After the fragment size was lowered also the amount of DNA was less, by doing a PCR amplification with less DNA the production of unwanted big fragments as previously shown in figure 18 pool 2, was automatically and successfully reduced. After the PCR amplification the DNA peak was good and the amount of unwanted big fragments was very low (cf. Figure 22, graph B).

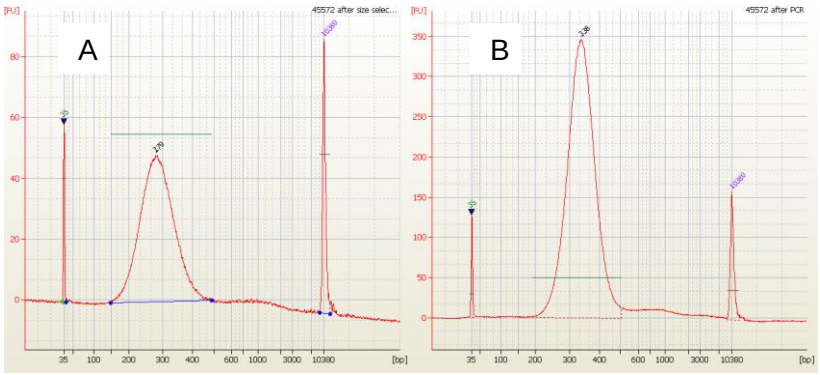


Figure 21 Axygen size selection target size 280 bp result graph A and the result from PCR graph B. The result A is from a Bioanalyzer High Sensitivity chip and result B 1000 DNA chip

When the samples were pooled an enriched the result of the optimized pools very similar between each other (cf. Figure 23). Despite small different average fragment size, the amount of the big DNA fragments produced by PCR artefacts was low and also the primer peaks were smaller after increasing the number of PCR cycles with one indicating a better performance of the PCR.

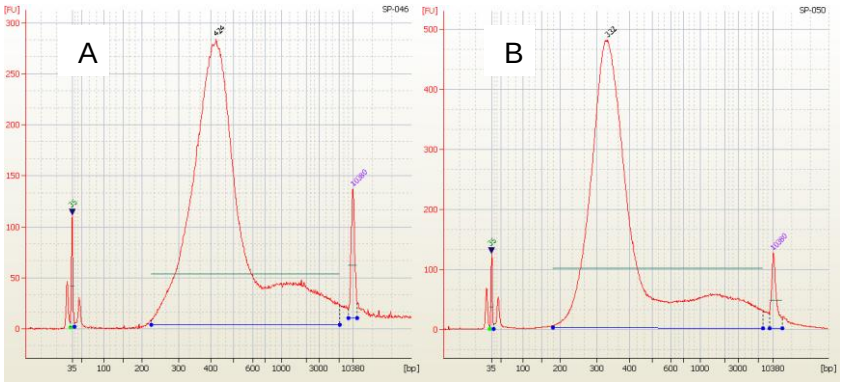


Figure 22 Sequencing ready pools 3 (graph A) and 4 (graph B). The results differ from each other in the fragment size. Result A is ~400 bp and result B ~300 bp. Results from Bioanalyzer High Sensitivity chip.

7.3 Sequencing Quality

After sequencing, the sequencing results can be viewed in the Illumina Sequencing Analysis Viewer. In the first page the quality of the runs are shown. The one of the most important quality parameter is QScore Distribution indicates the probability of errors during sequencing. QScore is cumulative for current cycle and previous cycles. The QScore is calculated from the samples that pass the quality filter.

Q10 = 10 % chance of wrong base call
Q20 = 1 % chance of wrong base call
Q30 = 0,1 % chance of wrong base call
Q40 = 0,01 % chance of wrong base call

The amount of data produced by sequencing is also calculated. The amount of data should be between 4-6G. If the generated data is more than 6 G it has an effect on the percentage of bases with Q30. The run results are collected in Table 11. All the runs were done the same way, but had different patient groups.

Table 11 Data quality of all runs

	1st Run	2nd Run	3rd Run	4th Run
Total yield	4,4 G	5,2 G	4,6 G	5,3 G
reads with >=Q30	91,8 %	95,3 %	95,4 %	92,5 %

All the runs reached the Q30 for more than 90 % of the reads which means that all the runs were good in quality and sequenced reads presented very few sequencing mistakes.

7.4 Differences in the Methods

The two tested methods worked very differently. The enzymatic fragmentation was found to be very inconsistent. The results were not reproducible and every tagmentation brought different fragment size distribution. Overall the DNA fragment size was too big. It is described that big DNA fragments tend to hybridize more unspecific than short fragments and bring bias to the capture enrichment results by hybridizing non-wanted regions of the genome. On the other hand the ultra-sonication was working extremely consistently and the results were reproducible. The end preparation after the fragmentation worked also as wanted. The size selection was easily scalable and it was only possible to obtain a fragmentation result of 300 bp DNA fragments which is the size recommended for further hybridization, capture and sequencing.

Tagmentation was very time efficient because the adaptor ligation is combined with the fragmentation in one step. On the other hand the ultra-sonication was not time consuming but the steps following to perform adaptor ligation took some of time. However result of the pool prepared with ultra-sonication as fragmentation method, brought better results. For this reason the ultra-sonication was the chosen method to prepare the rest samples for next-generation sequencing in this project Fragment size distribution was smaller in the ultra-sonication method than in the enzymatic. With ultra-sonic fragmentation and magnetic bead size selection the wanted narrow DNA fragment size distribution was reached. The ultra-sonic DNA fragmentation method complemented with the magnetic beads size selection method was found better because its scalability and reproducibility.

The effect of different DNA fragment size because of the methods in the hybridization of a genomic region of interest can be seen in the reads on target result tables (table 7). The on target reads percentage grows when the DNA fragment size distribution is narrower. This happens because the longer the fragments are, the worse the hybrization of the regions on target is. With the narrow size distribution from Covaris method the target selection has been better because there were more the probes for right size

fragments. When target selection is better the sequenced reads on the target region is higher.

7.5 Reads

When the patients are pooled the calculation and pipetting are not 100 % accurate, which has an affect then on the amount of sequenced of sequenced reads per patient (cf. Figure 24). The differences were not bigger than 5 %, indicating good pipetting and pooling of the prepared libraries. This small difference has not a significant effect on the reads on target.

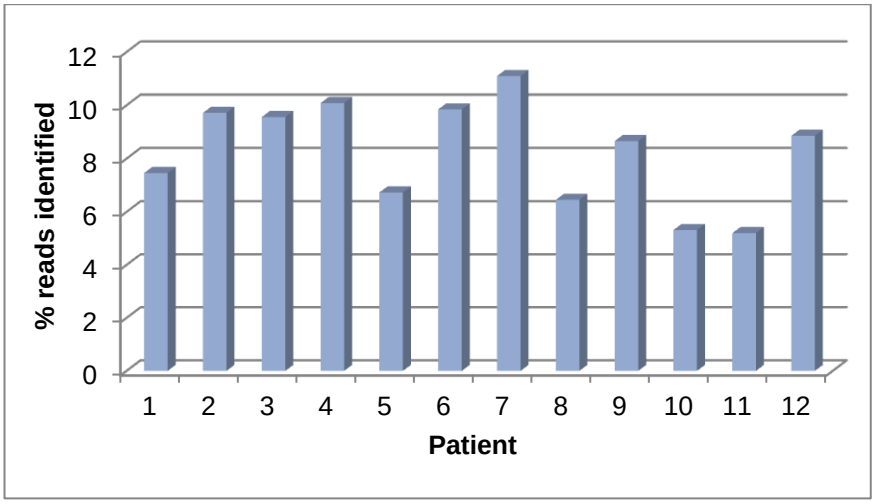


Figure 23 Identified reads per patient 2nd run. Good result is max 6 % difference.

The next step in the data-analysis is mapping the reads against a human genome reference. During mapping, the sequenced reads have to be filtered for wrong insert size, wrong orientation, one mate unmapped, both mates unmapped and for multiple hits in the genome. From the all sequenced reads and reads after filtering is calculated the percentage of reads on target. The read results from each run were collected in a table such as Table 12. The tables of the other batches results are shown in Appendices 1, 2 and 3. All the runs had different patients except the in the first and second runs which had 4 patient, for comparing the results.

Table 12 Amount of reads from the second run during the filtering steps in the data analysis

Patient	Sequenced Reads	Mapped Reads	Filtered reads	Reads on target
4813	2636338	2634434	840675	36,26%
24626	3433302	3429838	1074853	35,78%
30405	3378744	3374854	973017	34,82%
30669	3563116	3559830	1110776	35,93%
48594	2377624	2375702	36,39	36,39%
50725	3479012	3475010	1130404	36,60%
56682	3921686	3918148	1270104	36,55%
64921	2280354	2277766	763902	36,80%
67408	3129428	3126944	1044799	37,05%
73276	3058906	3055592	1021359	37,40%
68929	1878274	1876808	647664	38,03%
66012	1837624	1835602	646096	38,71%

The 1st batch average of reads on target was 20,72 %, 3rd batch 30,54 %, 4th batch 30,74 %.This result show that improving of the protocol improved also the number of reads on target by reducing the amount of background reads.

In the sequencing the most important factor for getting good results are the quality and amount of sequence reads with the next generation sequencer. When multiple samples are run simultaneously the distribution of the reads should be even between patients. If read distribution is not even, some of the samples might not obtain enough data for good mapping as well as for good detection of the mutations. The amount of reads that map the target region is very important. In this thesis the on target region included intronic regions which are harder to capture. There doesn't exist many studies in which would have been next-generation sequencing to sequence intronic regions. The results can not really be compared to any other data. When only exons are sequenced the expectation is to sequence 50-60 % of the reads on target. But when the intronic regions are also included the expectation of the percentage goes down, to around 30 %. The results from the sequencing were really good over expectations since between 30 % to 40 % of the reads being on target also in the intronic regions was obtained.

7.6 Coverage

After mapping the sequencing data against the human genome reference, the sequencing reads and coverage of the target regions was possible to be viewed at an igv-

software, which is software for viewing NGS read data. The coverage was good but it showed a repetitive pattern of well covered regions against reduced covered region along all the genomic regions of interest. Some places of the mapped reads showed gaps and the coverage is not distributed evenly to the whole region. The coverage forms a pattern which looks very systematic (cf. Figure 25 marked with green).

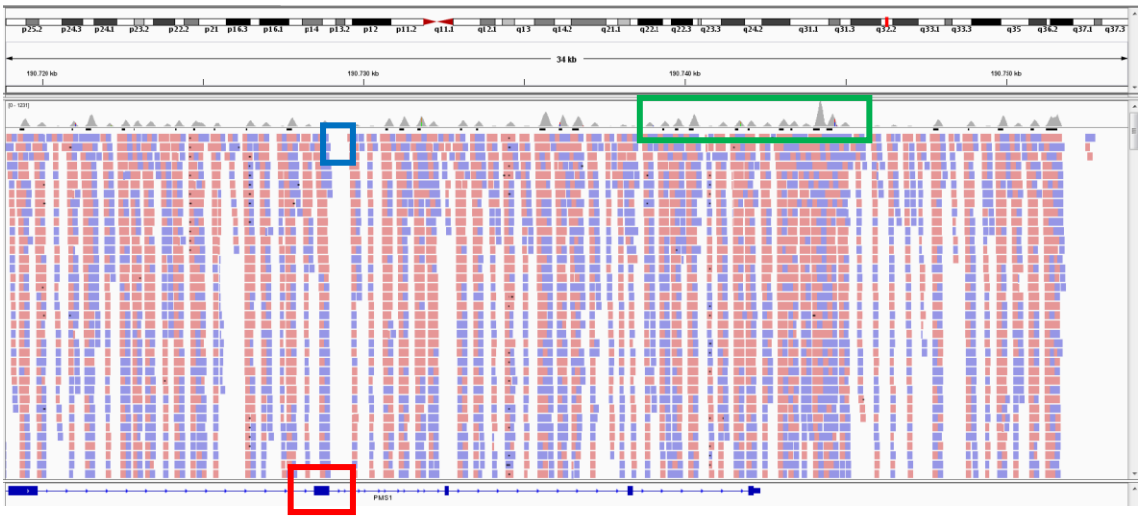


Figure 24 Coverage example from one of the 2nd batch patients. The gene the reads are aligned against is PMS1

In Figure 25 the gene which aligned data can be seen in under the mapped data. The thick blue bars are the coding exons and the area between them is intronic area (cf. Figure 25 marked with red). The oligos were selected so that they also cover after the genes to the intergenic region.

The coverage was surprisingly uneven (cf. Figure 25). The coverage was expected to be even, with an equal distribution along the intronic and exonic region of the genes. Coverage peaks show a very systematic pattern. The pattern continues through all the sequenced regions. The peaks are too systematic to be a mistake made in the laboratory, in the sample preparation. More likely it is a problem in custom selected oligos. It seems that in the Illumina software design for the oligos does not make a difference when designing oligos to capture exons or introns. It is known that oligo design for intronic regions should be calculated with a different method that the one for exons to improve intronic capture, since introns are more difficult to capture than exons. The intronic regions are very difficult to sequence because in the genome is similar parts as the target intron. This causes that some unwanted areas are also captured which lowers the coverage in some introns and explains the gaps. However, overall coverage

results showed enough good quality data to go on with analysis of the nucleotide variants and the SNPs could be examined.

7.7 SNPs

The next step in data analysis was the variant calling. For variant calling a software was used to find differences between the patient DNA nucleotide strand and the reference DNA strand. From each patient was first found a lot of SNP but this number was reduced after some filter steps to avoid false positive variants because of errors induced by the sequencing process. The SNPs were first filtered (Table 13 colum 3) with quality measures of at least 10−fold coverage per base, 10 % Frequency, 5 % forward and reverse read frequency.

Table SNP results from 2nd run

Patient	All SNPs	Filtered SNPs	Common mutation	Rare mutations	Silent mutations	Missense
4813	1474	956	819	219	1212	17
24626	1703	1151	1034	259	1521	12
30405	1093	642	475	208	817	10
30669	1724	1177	1031	287	1558	16
48594	1495	968	863	257	1301	10
50725	1409	903	758	279	1184	9
56682	1523	973	860	268	1286	11
64921	1613	1029	939	258	1375	13
67408	1735	1184	1079	299	1624	15
73276	1599	1014	913	249	1368	11
68929	1485	986	882	254	1364	13
66012	1796	1199	1127	278	1653	15

Then the SNPs were calculated for known SNOs with a rsNumber, the rsNumber is an ID number for UCSC for known SNOs found in the genome and for common SNPs present in the healthy normal population and rare SNPs not so often present in the healthy normal population. The common SNPs were defined by a frequency over 0,1 in 1000 genomes control population. If the SNP had a rsNumber, which indicates that the SNP has been previously described, but no frequency or it had no rsNumber they were calculated as a rare SNP variants. These SNP candidates will be considered as a causative for the HNPCC disease phenotype. The effect on the protein of the candidate

was calculated. First silent mutations were discarded from the disease-causing candidate list of SNPs, since silent mutations are mutations were the change of the nucleotide doesn't effect on the amino acid sequence. These mutations are not looked into detail because they are not likely the cause of a disease. The missense mutations are the mutations which are the ones that change the amino acid sequence and from these it is possible to find pathogenic mutations. Each runs SNP results were set in a table such as Table 9. The tables of the other batches results are in Appendices 1, 2 and 3.

7.8 SNPs Possible Pathogenic Mutations

When a SNP causes a missense mutation it changes the amino acid sequence. The change in the sequence can result in a stop codon and too short protein or in a malfunction of the protein. After filtering for known and common SNPs with a rsNumber and a high 1000Genomes healthy population frequency a list of putative mutations was defined per each patient. The missense mutations found from the patients will be looked into detail. These missense mutations which don't have an rsNumber are mutations that are unknown.

The mutations which are thought of being hot candidates for unknown pathogenic mutations will go into further testing. For the testing RNA sequencing can be used. In RNA sequencing the RNAs are sequenced and looked if the expression has changed. The RNA is aligned with slice junctions and then possible isoforms, novel transcriptions and gene fusions can be examined. Also the genes can be tested with mice. In mice the function of a mutation can be examined.

8 Conclusion

The purpose of the thesis was to find a good reproducible laboratory protocol to sequence intronic and exonic regions with next-generation sequencing. As fragmentation is one of the most crucial steps in the laboratory sample preparation two different methods were tested. Ultra-sonication method together with a DNA size selection previous to addition of adaptors and indices during sample preparation was found to be more scalable and the results were more reproducible. Also this method gave more reads on target and less background reads from sequencing. Following this mehod

enough good quality data was produced to test patient samples for disease causing mutations. The missense mutations which were found will need further testing and examination before they can be determined as pathogenic mutations causing Lynch-syndrome.

References

[1] Michael L. Metzker 2010. Sequencing technologies - the next generation. Nature Reviews: 11: 31-44

[2] Elaine R. Mardis 2008. Next-Generation DNA sequencing Methods. Annual Rev. Genomics Hum. Genet. 9: 387-402

[3] Comparison of Commercially Available Target Enrichment Methods for Next Generation Sequencing: Journal of Biomolecular Techniques: 2013: 24: 73-86

[4] J.M. Rizzo and M.J. Buck 2012. Key Principles and Clinical Applications of “Next-Generation” DNA Sequencing. Cancer Prev Res: 5: 887-900

[5] http://www.nanodrop.com/Library/T009-NanoDrop%201000-&-NanoDrop%208000-Nucleic-Acid-Purity-Ratios.pdf

[6] Quant-iT^{TM} PicoGreen dsDNA Reagent and Kits 2008

[7] http://biomedicalgenomics.org/How_does_Agilent_2100_Bioanalyzer_work.html

[8] http://www.zymoresearch.com/dna-purification/genomic-dna/genomic-dna-clean-up/genomic-dna-clean-concentrator

[9] http://covarisinc.com/technology/afa-vs-sonicators/

[10] Illumina datasheet: NexteraTM DNA Sample Preparation Kits

[11] Illumina datasheet: Multiplexed Sequencing with the Illumina Genome Analyzer System

[11] Silva FCC 2009. Mismatch repair genes in Lynch syndrome: a review. Sao Paulo Med J.: 127: 26-51

[12] Henry T. Lynch 2009. Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. Clin Genet: 76: 1-18

[13] Wendy Kohlmann, Stephen B Gruber, 2004. Lynch syndrome. GeneReviews Available: http://www.ncbi.nlm.nih.gov/books/NBK1211/#hnpcc.REF.hendriks.2006.312

[14] Illumina protocol. TruSight Rapid capture Sample Preparation Guide

[15] Illumina protocol. TruSeq DNA enrichment

[16] New England Biolabs. NEBNext Ultra DNA Library Prep Kit for Illumina

[17] Axygen Biosciences. AxyPrep™ Mag FregmentSelect-I Protoco

1st runs results

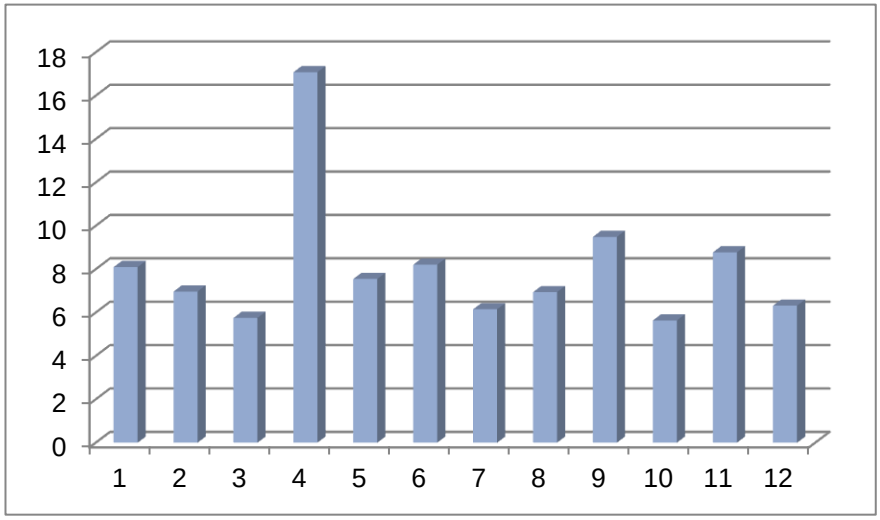
Reads

Patient	Sequenced Reads	Mapped Reads	After filtering	Reads on target
36973	2458392	2453364	453889	20,58%
39199	2116802	2101882	382201	20,88%
41696	1746690	1743860	342311	21,25%
44847	5186640	5178290	891991	19,87%
46610	2292512	2287786	429498	20,51%
46614	2491692	2476016	444006	20,61%
47232	1867594	1855666	337976	20,86%
49160	2107294	2098404	376212	20,25%
64921	2882214	2877810	532611	20,63%
67408	1711264	1710280	324415	20,77%
68929	2664728	2662478	493972	20,52%
66012	1918322	1917252	382080	21,87%

SNPs

Patient	All SNPs	Filtered SNPs	Common mutation	Rare mutations	Silent mutations	Missense
36973	1214	737	619	197	968	13
39199	1293	746	568	196	910	14
41696	1671	1162	1031	251	1489	12
44847	1884	1370	1224	310	1795	18
46610	1605	1143	986	286	1474	12
46614	1323	799	617	211	978	10
47232	1511	853	652	210	1030	8
49160	1517	887	690	232	1100	13
64921	1709	1087	946	293	581	15
67408	1660	984	861	244	1304	13
68929	1751	1191	1023	320	1623	14
66012	1968	1294	1165	310	1742	15

Read division



3rd runs results

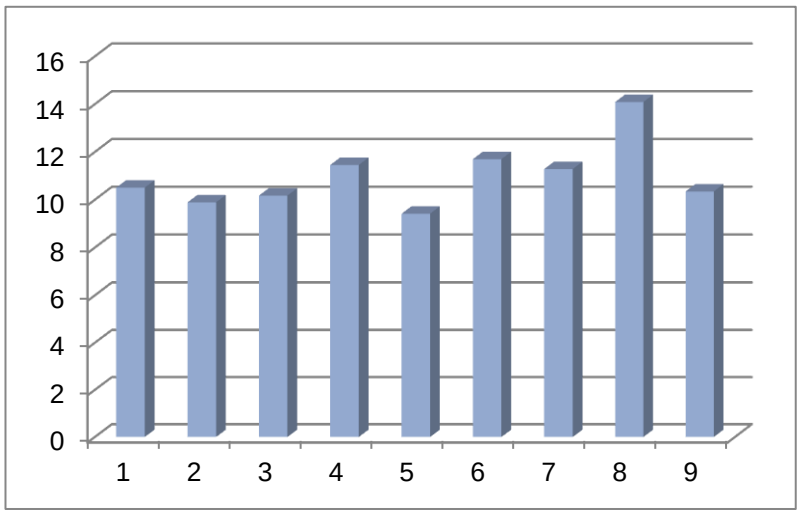
Reads

Patient	Sequenced Reads	Mapped Reads	After filtering	Reads on target
46612	3266438	3263070	853018	30,27%
51766	3076126	3073302	784147	30,30%
54816	3162670	3159534	828345	31,08%
56233	3561468	3557146	919109	30,86%
57277	2927526	2924912	781692	30,77%
24292	3637636	3633366	893988	30,14%
39331	3509258	3504724	912585	30,93%
41407	4380344	4375558	1073891	29,85%
56770	3216722	3213870	830508	30,66%

SNPs

Patient	All SNPs	Filtered SNPs	Common mutation	Rare mutations	Silent mutations	Missense
46612	1783	1313	1193	314	1769	12
51766	1444	1022	902	277	1391	13
54816	1780	1344	1243	298	1836	18
56233	1715	1309	1128	344	1715	13
57277	1851	1419	1236	353	1837	16
24292	1436	1088	928	289	1463	14
39331	1532	1165	1017	309	1524	14
41407	1790	1342	1212	343	1829	16
56770	1653	1237	1119	302	1680	12

Reads division



4th runs results

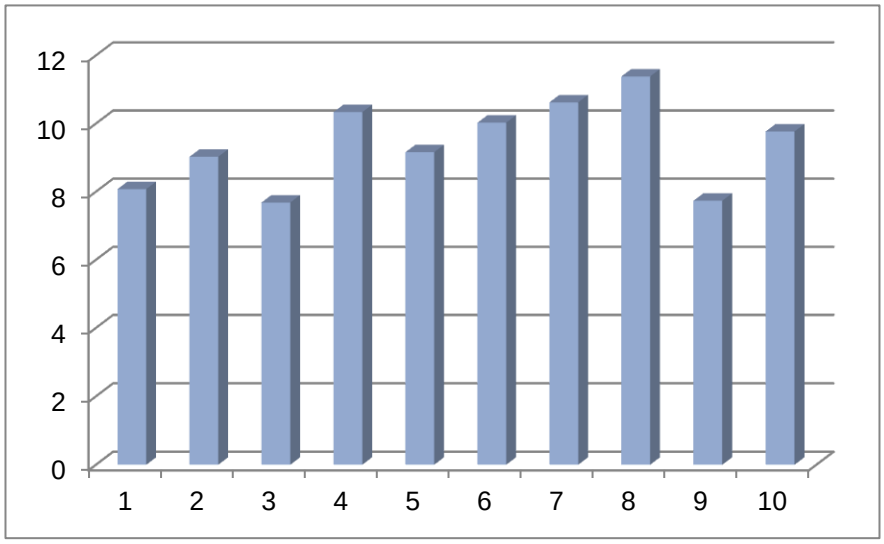
Reads

Patient	Sequenced Reads	Mapped Reads	After filtering	Reads on target
43338	3009272	3007210	753940	31,22%
45421	3362848	3360962	752380	30,09%
49014	2862316	2860536	728129	31,79%
54317	3850630	3847926	1020166	32,49%
68003	3412246	3410614	858088	31,67%
40759	3735246	3732436	803461	29,34%
52356	3957976	3955278	846082	29,13%
52406	4241750	4239064	997158	30,73%
52589	2881084	2879288	709560	31,44%
67162	3639478	3636632	787586	29,51%

SNPs

Patient	All SNPs	Filtered SNPs	Common mutation	Rare mutations	Silent mutations	Missense
43338	1738	1185	1114	285	1634	12
45421	1932	1288	1205	301	1785	14
49014	1644	1075	978	278	1440	12
54317	1504	1027	876	284	1357	13
68003	1293	858	747	271	1161	9
40759	1699	1116	1009	305	1465	12
52356	1493	1003	897	269	1387	12
52406	1686	1182	1104	267	1630	17
52589	1279	853	744	245	1191	7
67162	1483	1018	894	293	1348	12

Reads division


12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10