School of Biology

Dissertation Title: Hybridization enrichment of subgenomic targets for next generation sequencing

Submitted by: Ioannis Ladas

Supervised by: Professor John A.L. Armour

Programme of Study: MRes in Human Molecular Genetics

(2009/2010 Academic Year)

Acknowledgments

First of all I am very thankful to my project supervisor Professor John A.L. Armour for his precious contribution and guidance. He responded to all my enquiries and showed great patience and understanding.

I would like also to express my appreciation to Dr. Jess Tyson for her help and encouragement throughout my master degree.

Finally, I would like to thank the member of C10 laboratory Raquel Pala, Sugandha Dhar, Somwang Janyakhantikul, Fayeza Khan, Suhaili Abu Bakar, Danielle Carpenter and Tamsin Majerus for their cooperation and precious assistance.

Table of Contents

List of Abbreviations

Abbreviation	Meaning
ATP	Adenosine-5'-Triphosphate
BSA	Bovine Serum Albumin
dATP	Deoxyadenosine Triphosphate
dCTP	Deoxycytidine Triphosphate
dGTP	Deoxyguanosine Triphosphate
dTTP	Deoxythymidine Triphosphate
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic Acid
EtOH	Ethanol
HMFM	Hogness Modified Freezing Medium
HNPCC	Hereditary Non-Polyposis Colorectal Cancer
LB	Luria Broth
$MgCl_2$	Magnesium Chloride
NaH_2PO_4	Sodium Phosphate

NaOH	Sodium Hydroxide
PCR	Polymerase Chain Reaction
SDS	Sodium Dodecyl Sulfate
SOC	Super Optimal broth with Catabolite Repression
SSC	Saline-Sodium Citrate
TBE	Tris/Borate/EDTA
TE	Tris/EDTA
T4 PNK	T4 Polynucleotide Kinase
UCSC	University of California, Santa Cruz
X GAL	5-bromo-4-chloro-3-indolyl-b-D-galactopyranoside

Abstract

The aim of the project was the development of a technique by which a subgenomic set of target sequences could be captured in order to be analysed by next-generation sequencing technology.

The technique involved two rounds of filter hybridization enrichment by which genomic DNA fragments of interest were captured, followed by MAPH (Multiplex Amplifiable Probe Hybridization) for evaluation of the enrichment. Enriched genomic DNA was cloned and sequenced, for the level of enrichment to be estimated.

After two rounds of enrichment for human MSH2 exons, approximately 90% of the total cloned sequences were found to contain sequences of interest, equal to an enrichment of 600,000 times.

The new technique was therefore shown to be efficient with high specificity and could be used as a potential clinical diagnostic tool.

Chapter 1

Introduction

1.1 Human Genetic Variation

With the development of DNA sequencing techniques and of technology in molecular biology it has been determined that two randomly -selected human genomes are approximately 99.9% identical (Feuk *et al*. 2006). That 0.1% difference is responsible for the different phenotypes of humans, and there is evidence which suggests that genome variability might be responsible for a number of genetic diseases, such as cancer, diabetes, muscular dystrophies and others (Shaikh *et al.* 2009). Copy number variations involve about 5-12% of the human genome (Kato *et al*. 2009, McCarroll et al. 2008 and Redon *et al.* 2006). Genome variation occurs in variable types, such as variable number tandem repeats, single nucleotide polymorphisms and others as can be seen in Figure 1. In the current research project a novel method for capturing specific

regions, for next generation sequencing analysis was developed.

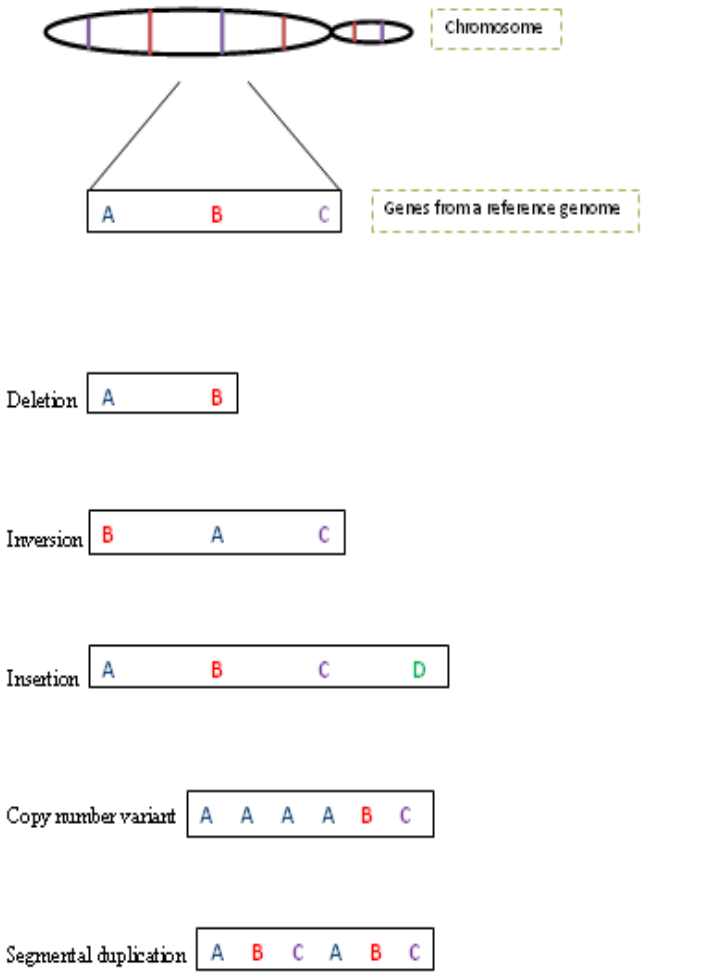


Figure 1. A summary of the variations which are present in the human genome.

1.2 High- Throughput sequencing

1.2.1 Next generation Sequencing

The analysis of the human genome was always challenging and a great ambition for scientists. The development of DNA sequencing technologies in the past few years has allowed DNA sequencing to be done with low cost and high efficiency. That high-throughput technology is generally known as next generation sequencing technology (Shendure and Ji 2008).

DNA sequencing can provide valuable information about essential biological processes such as protein- DNA interactions and chromosome conformation. In addition to that, with the analysis of the whole human genome sequence, polymorphisms and mutations can be discovered and with the development of ultra- deep sequencing it would in principle be possible to develop personal medical treatment and diagnostics (Ansorge 2009).

1.2.2 Sanger Sequencing

Since 1977 the main technique for DNA sequencing has been the Sanger method. That method is based on sequencing by synthesis, where termination is used to identify the nucleotides. The DNA strands are denatured and a primer binds next to sequence of interest in such manner that the 3' end of the primer is attached right at the beginning of the sequence of interest. Sanger sequencing requires four distinct reactions in its original form as seen in Figure 2. Each reaction consists of the sequence of interest with its primer, DNA polymerase, four dNTPs and ddNTPs at about 1% of the concentration of dNTPs. The major difference between dNTPs and ddNTPs is that ddNTPs contain a 3' H instead of 3'OH. That small difference is vital for sequencing, because in each cycle the primer extension is terminated by the use of fluorescent labelled dideoxynucleotides (ddNTPs) (Russel 2002, Sanger *et al*. 1977). Once the dideoxynucleotide (ddNTP) is incorporated, the growing chain is fluorescently labelled and because it lacks a hydroxyl group at the 3' end the chain cannot be further elongated and therefore is terminated. The sequence is then

identified by high- resolution polyacrylamide gel electrophoresis. The fragments are distinguished according to their lengths and their fluorescent last nucleotide can be identified by laser excitation (Shendure and Ji 2008, Primose and Twyman 2003 and Russel 2002).

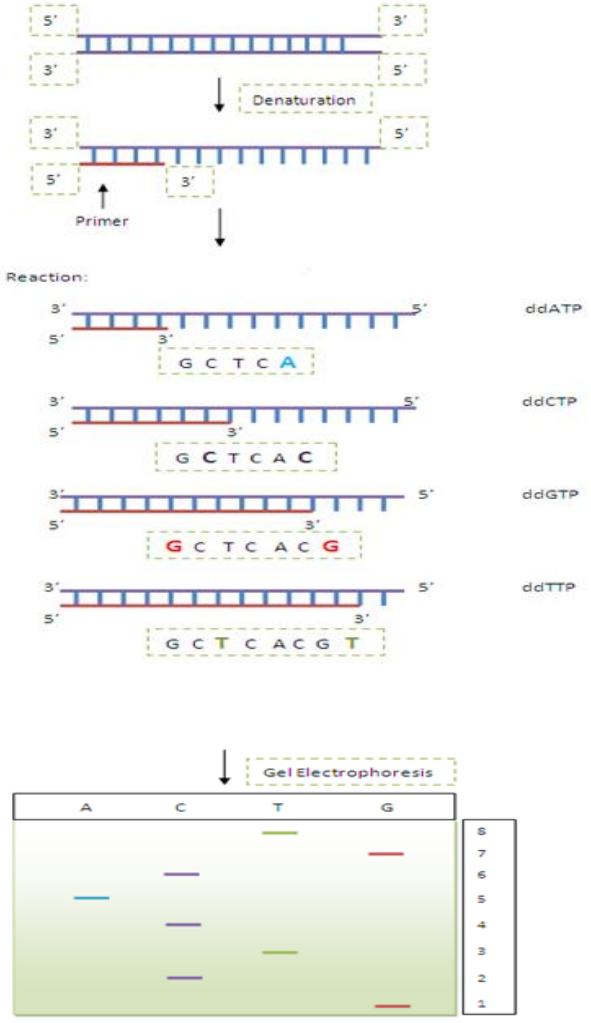


Figure 2. The Sanger method, the first DNA sequencing technique which can succeed in reading lengths up to 1000bp, includes four types of reaction each of which contains a different labelled ddNTP.

Once the reaction started and a ddNTP is incorporated the elongation is terminated. The identification of the different ddNTPs which were incorporated is achieved by gel electrophoresis, where the different chains are distinguished according to their length (Shendure and Ji 2008).

1.2.3 Roche/454 FLX pyrosequencing

The Sanger method was the main method with which fragments of DNA could be sequenced until 1985, when the Roche/454 FLX pyrosequencing basic principle was first described (Nyren and Lundin 1985). The 454 pyrosequencing was the first next generation sequencing technique described. In that technique, microscopic beads which are covered with DNA primers are mixed with DNA fragments which allow the binding of the DNA fragments to the one end. Every bead contains product from one DNA fragment. The DNA binding is followed by emulsion PCR and after a appropriate number of cycles every bead contains thousands of copies of the same DNA fragment (Zhou et al. 2010). The beads eventually are placed on a picotiter plate array and each hole is expected to

contain only one bead. Also, even smaller beads which contain a number of enzymes that are required for pyrosequencing are placed on the plate (Mardis 2008). During pyrosequencing, deoxyribonucleotide triphosphates are added sequentially, and incorporation is followed by release of pyrophosphate in equal amount to the amount of the incorporated nucleotide as seen in Figure 3. Pyrophosphate is then converted by ATP sulfurylase to ATP in the presence of adenosine 5’ phosphorylase. The ATP which is generated converts luciferin to oxyluciferin which produces visible light. The light is then detected by a charge-coupled device and peaks are generated which indicate the number of nucleotides incorporated.(Shendure and Ji 2008, Ansorge 2009).

The most important drawback of this technology is in sequencing repetitions of the same base, because every base is identified by its light signal, and with a long sequence consisting of the same bases, their accurate number is in general difficult to discriminate (Shendure and Ji 2008).

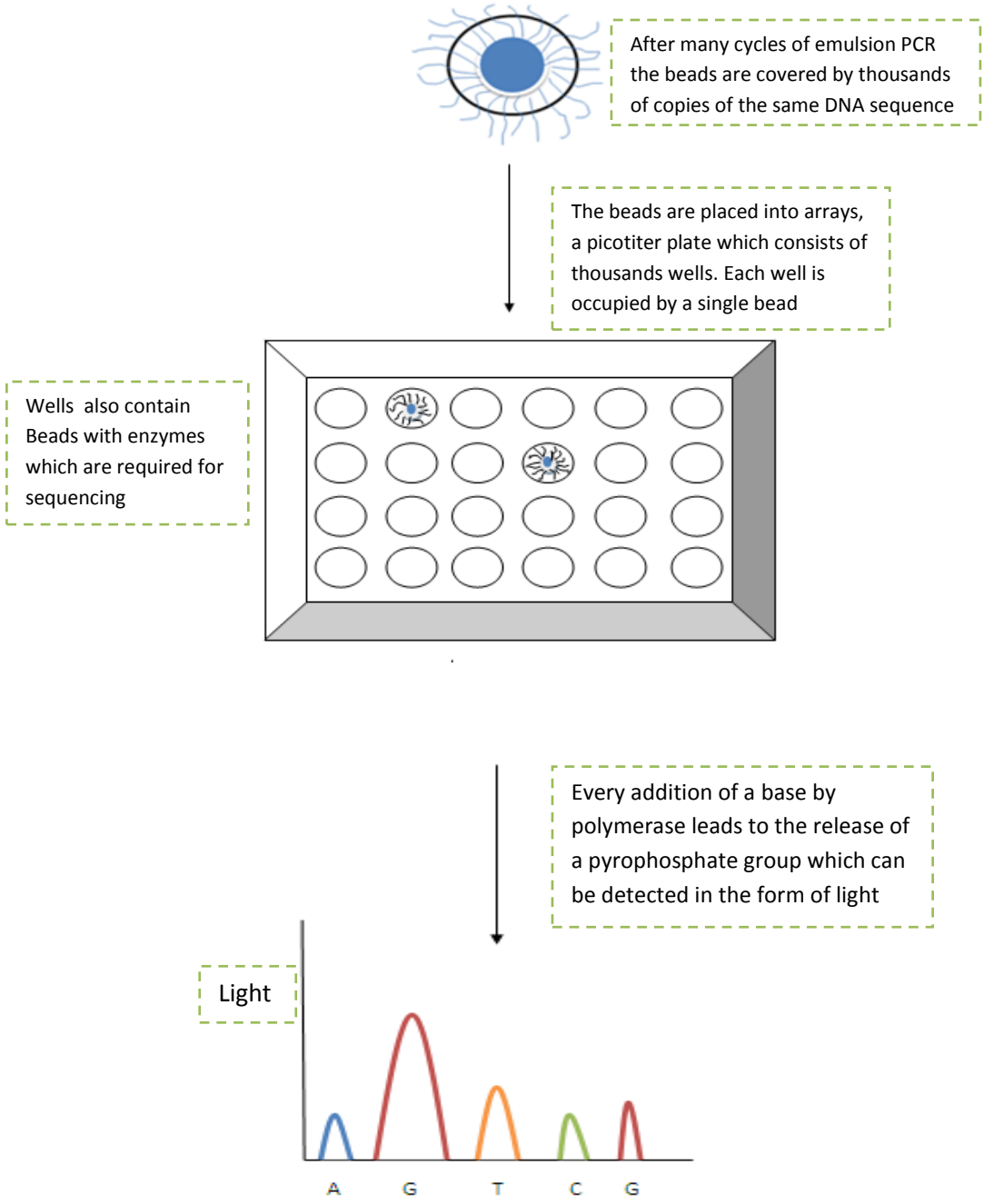


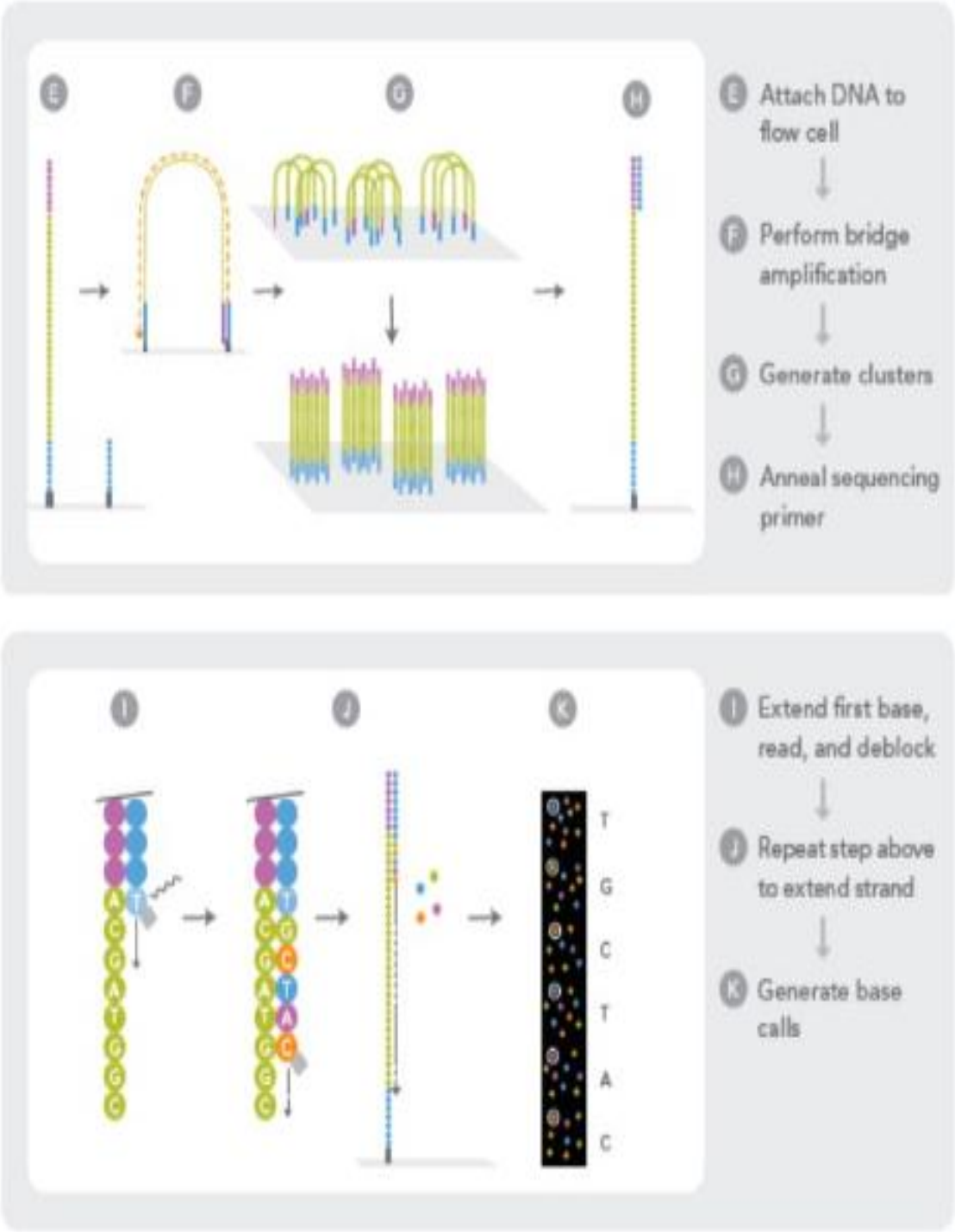
Figure 3. 454 pyrosequencing is the first next -generation sequence technique, and uses the generation of visible light for the identification of DNA sequence. The beads with the DNA fragments on are placed on a picotiter plate which consists of thousands of

microscopic wells organized in an array, so every well contains only one bead. Every well is connected with a charge coupled device. Therefore, approximately 400,000 reads of 200 to 300 bp per run can be achieved and despite the fact is relatively more expensive than the other next generation sequencing techniques, this is ideal in experiments where long reads are required (Shendure and Ji 2008 and Zhou et al.2010).

1.2.4 Illumina Solexa sequencing

The Illumina Solexa is another technique for next generation sequencing analysis, which was introduced in early 2007. Illumina like other DNA sequencing methods is based on sequencing by synthesis. In that method DNA fragments are ligated at both ends with adapters and fixed on a solid support which has on its surface a great number of the complementary adapter. The DNA fragments form bridges with the free complementary adapters. Then PCR occurs for several and cycles and clusters of 1000 copies of single stranded DNA are generated on the surface. In order for the sequencing process to begin a primer which is required for sequencing binds to a universal sequence and the sequencing is initiated (Zhou et al.

2010). The clusters of the DNA sequences are present on the surface of a glass flow cell. Every lane contains primers which are complementary to those on the DNA library (Mardis 2008). The identification of the DNA sequences is accomplished with the use of reversible terminator nucleotides which are labelled with different fluorescent dyes. The labelled terminator nucleotide can be detected after every incorporation, by a charge coupled device. The terminator group is then removed as well as the fluorescent label and the synthesis cycle is repeated as seen in Figure 4. It is powerful technique which can analyse more than 40 million polonies in parallel. On the other hand, the major limitation of the technique is the average size of read length which is approximately 40bp and is relatively short, and the technique is susceptible to errors which accumulate as the DNA strands are elongated (Shendure and Ji 2008, Ansorge 2009 and Zhou et al. 2010).



(Ansorge 2009)

Figure 4. Illumina sequencing is a relatively new technique where reads up to 36 bp can be easily sequenced which makes Illumina

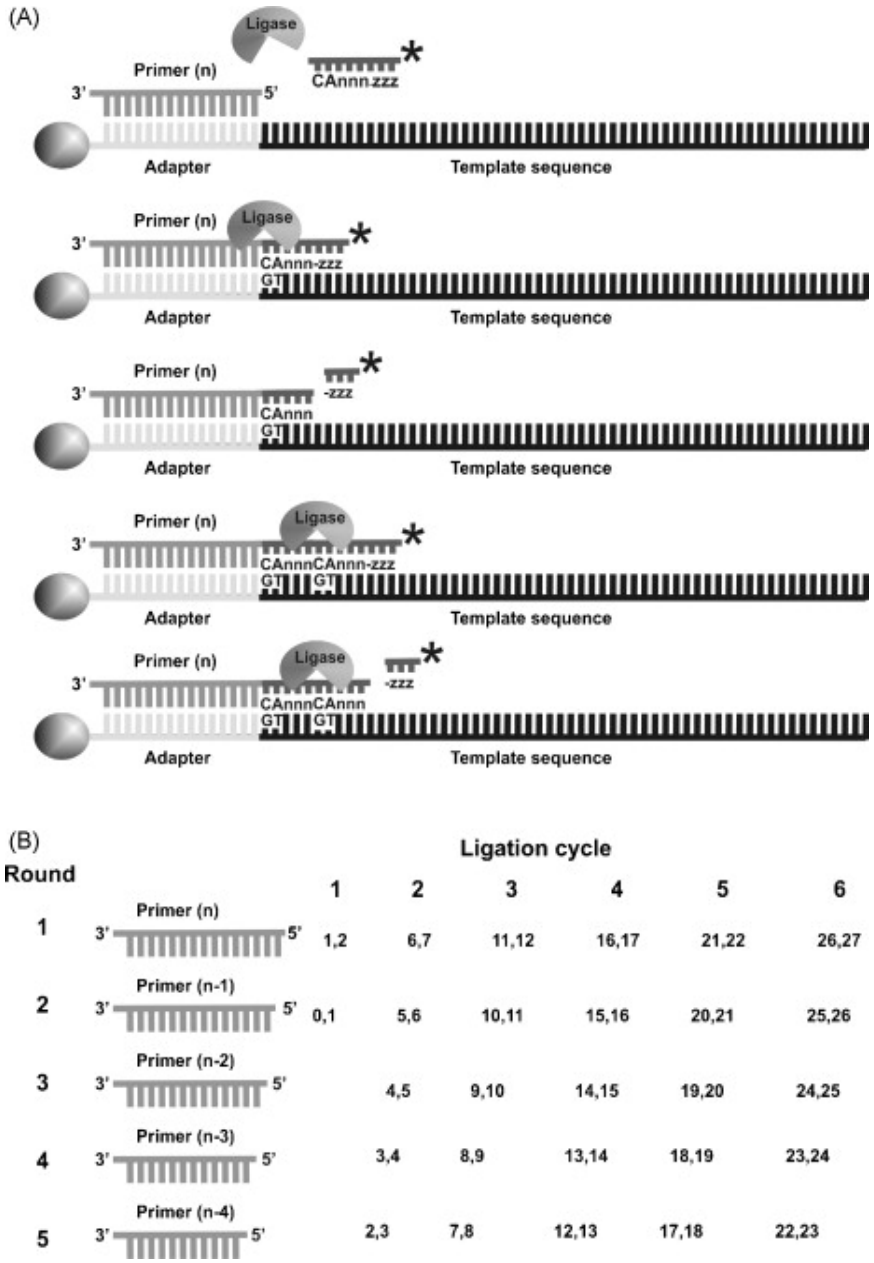
ideal for experiments which require sequencing of short reads (Ansorge 2009).

1.2.5 SOLID Sequencing

In addition to Illumina, a more recent technique is the ABI SOLiD sequencing system. In this method adapters are ligated to DNA fragments and then bound to beads. Then the DNA fragments are amplified by emulsion PCR. The beads with the fragments are then placed onto a glass surface where primers are hybridized to adapters in addition to oligonucleotide octamers which also hybridize to the DNA sequence. The fourth and fifth nucleotide of the octamer are determined once the fluorescent label, at the four nucleotides at the end of the octamer is detected. After ligation the nucleotides after the fifth one are cleaved off, removing the fluorescent label and the cycles are repeated. In order for the nucleotides in between to be identified, shorter primers are used and the same process is repeated as can be seen in Figure 5. SOLiD sequencing is a powerful method which can analyse up to 10 Gigabases per run with an average read length of approximately 30bp

(Shendure and Ji 2008, Ansorge 2009, Mardis 2008 and Zhou et al. 2010).

The major advantage of the technique is accuracy, as every based is measured twice because of the use of different size of primers in order the full length of the sequence to be covered. On the other hand, the read length in comparison with the other sequencing techniques is relatively short and that is indicated as the main drawback of SOLID sequencing (Zhou et al. 2010).



(Ansorge W. J. 2009)

Figure 5. SOLiD sequencing is a low cost sequencing technique compared to the other next generation sequencing methods, with very low error rate. However, emulsion PCR is technically challenging which makes the application of the technique more difficult. (Shendure and Ji 2008)

1.2.6 Heliscope Sequencing

Highly parallel sequencing by synthesis was a technological revolution and therefore is the basic method of sequencing of the majority of next generation sequencing technologies. However, has several drawbacks related to the inaccuracy of the DNA synthesis. Errors can be introduced during the amplification and potential loss of synchronicity in synthesis can lead to accumulation of errors. Therefore, single-molecule sequencing (SMS) was developed, a technique which has overcome those limitations (Gupta 2008). Heliscope single-molecule sequencing is a novel technique for next-generation sequencing which uses the same general principle as Illumina sequencing but with the difference that no amplification is required. Primers are hybridized to the DNA sequence which is on a glass support with labelled nucleotides and DNA polymerase. The labelled nucleotides are incorporated as the primers are extended, as Illumina sequencing, every incorporation produces visible light which can be detected by a charge coupled device (Shendure and Ji 2008, Ansorge 2009).

Technology	Approximate single-read length	Paired end (size)
Roche 454	450	3 kb, 8 kb, 20 kb
Illumina GAIIx	100	200 b to 1.5 kb
Life Technologies SOLiD	50	500 b to 10 kb
Helicos	35	NA
Complete Genomics	35	200–300 bp

(Snyder M. et al. 2010)

Table1. The properties of next generation sequencing techniques

On the other hand, the most important limitation of next generation sequencing technologies is the high cost per run. Next-generation technologies have improved the accuracy and read length of sequencing, but the cost still remains high. It is estimated that for sequencing of 100 genes from 100 samples, where every gene contains 10 exons, the cost can vary from $300,000 to approximately $1,000,000 (Shendure and Ji 2008).

1.2.7 Personal Genomics

Next-generation sequencing of the human genome can identify variants of 1% frequency as well as somatic mutations. It has been shown that the genetic basis of the majority of common

diseases is polygenic and much remains to be discovered about their aetiology of these diseases. The development of sequencing technologies allows the personal sequencing of many individuals, aiming to provide valuable information related to gene regulation, chromosome structure and the genetic basis of diseases (Snyder et al. 2010). A characteristic example is the sequencing of the genomic DNA of tumour and healthy skin cells from an individual with M1 subtype of acute myeloid leukaemia (AML) (Ley *et al*. 2008). The study identified ten non-synonymous somatic mutations. Two of them were an internal tandem duplication of FT3 and a four-base insertion in exon12 of the NPM1 gene. Those mutations were common in approximately 30% of AML tumours and were considered to be responsible for the progression of the disease. The eight remaining mutations were all single base changes and were considered to be strongly associated with cancer pathogenesis. It was important that they had not been detected in an AML genome before (Ley *et al*. 2008). Therefore personal sequencing can provide valuable information about the genetic mechanism of the diseases. In addition to that, personal sequencing has also been used for the construction of a

detailed catalogue for somatic mutations from human cancer genomes. That catalogue was constructed from mutations that were recorded from COLO-829 cell line (Table 2).

Type of change	Count	Percentage
Substitutions	33,345	100.0
Coding	292	0.9
Silent	105	0.3
Missense	172	0.5
Truncating	15	<0.1
Non-coding	319	1.0
UTR	205	0.6
ncRNA	113	0.3
miRNA	1	<0.1
Intronic	9,543	28.6
Splice	7	<0.1
Other intronic	9,536	28.6
Intergenic	23,191	69.6
Small insertions and deletions	66	100.0
Coding	0	0.0
UTR	2	3.0
Intronic	27	40.9
Intergenic	37	56.1
Rearrangements	37	100.0
Breakpoints	74	
Coding	1	1.4
UTR	0	0.0
Intronic	36	48.6
Intergenic	37	50.0
Classes	37	100.0
Intrachromosomal	34	91.9
Deletions	25	67.6
Inversions	6	16.2
Duplications	2	5.4
Other	1	2.7
Interchromosomal	3	8.1

miRNA, microRNA; ncRNA, non-coding RNA; UTR, untranslated region.

(Pleasance E.D. et al. 2010)

Table 2. Catalogue with the somatic mutations which were found in COLO-829 cell line

The sequencing of individuals in the future will give the opportunity for the creation of many catalogues which will provide with valuable information related to DNA damage,

mutation and repair of DNA that are associated in human cancer cases (Pleasance *et al*. 2010).

Moreover, current drug treatment does not work with all individuals due to genetic variability; the same drugs can be beneficial for some individuals but the same time can be harmful for others (Ginsburg and McCarthy 2001). Therefore human genome sequencing may enable better understanding about the cause of the diseases and is a new opportunity for the generation of more effective treatment (Guttmacher et al. 2010). A significant example is the genome sequencing of a 40 year old individual, the family of whom had history of early sudden death and vascular diseases. The genome sequencing identified, apart from the variants and SNPs which were related to sudden cardiac death and cardiopathies, 63 pharmacogenomic variants and SNPs which were associated with drug response. The individual was found to have variations associated with low maintenance dose of warfarin. Therefore, for this individual a future potential drug treatment with warfarin should be dose modified with lower expected doses (Ashley et al. 2010). Recent sequencing and comparison the genome of two different individuals where the first was healthy and the

second was diagnosed with acute myeloid leukaemia is a significant example of the potential benefits the personal sequencing comparison may have (Ley *et al.* 2008). In addition to that, the sequencing of a heavy smoker Chinese individual identified variant alleles that were risk factors in tobacco addiction (Stevens *et al*. 2008). Also, physical mapping can be achieved by DNA sequencing and enables the recording and identification of the different loci which are responsible for the different physical characteristics and can be useful for animal and plant breeding programmes. Human genome sequencing can use DNA and RNA information for better understanding of biological systems. A significant example is the sequence of the genome of a male Yoruba from Nigeria, where approximately 4 million SNPs were identified. In addition has been shown that this genome from Yoruba individual shows higher frequency of polymorphism than a genome from an individual with European origin. Also 153 premature termination codon SNPs were considered to affect protein function (Bentley *et al.* 2008).

1.2.8 1000 Genome Project

Finally, as new DNA sequencing techniques are developed they are expected to reduce the cost of sequencing and increase its efficiency even more. In addition to that a personal genomics project which aims to sequence 1000 human genomes is already in progress which promises to reveal many secrets related to diseases and human evolution (Ansorge 2009). The 1000 genomes project is aiming to identify the majority of genetic variations that appear to be present at more than 1% frequency in the human populations. The identification of the potential variation will be a crucial step for the discovery of disease causing genes and will provide valuable information related to the evolution of the human genome, as the project is an international collaboration with genomes from all the continents. In the future, that knowledge could become beneficial in clinical medicine, for prediction of disease susceptibility and discovery and response. The first results from the 1000 genome project were so far from the pilot 1 analysis. The outcome was so far encouraging, with more than 9 million

SNPs, numerous new indels and few large structural variants identified (Via *et al*. 2010).

However, the genomic information can be psychologically harmful for the individuals examined. In general, patients tend to exaggerate genetic predictions, which might cause panic and fear for the rest of their lives. (Ransohoff and Khoury 2009 and McGuire *et al*. 2008)

1.2.9 Exome Sequencing

As described above, genetic variants could be vital for the understanding of rare and common diseases. Therefore, the sequencing of only the protein-coding genes has been shown to be ideal for the identification of rare variants with high specificity and reduced cost which makes the technique a powerful clinical diagnostic tool. The exome which consists of the protein coding genes represents 1% of the human genome. Also, it has been considered that about 85% of the disease-causing mutations are located in the coding regions or the canonical splice sites (Ng *et al*. 2010, Ng.*et al*. 2009, Choi M. *et al*. 2009).

Exome sequences can be captured with the use of a human exome oligonucleotide array and sequenced with Illumina next-generation sequencing technology. The genomic DNA is fragmented with sonication and adapters which contain primer site sequences are ligated to the DNA fragments. Those fragments are then separated according to their size by gel electrophoresis and the desired size DNA is extracted from the gel. The fragments are then PCR amplified and hybridized to the exome array which can capture approximately 180,000 coding exons with the use of 2.1 million probe array. The array can capture with high specificity of approximately 90% (Hedges *et al.* 2009). The captured fragments are then sequenced using Illumina sequencing technology. The process is relatively simple and less expensive as the fragments are only the exome and not the whole genome and also because Illumina sequencing is relatively cheaper than the other sequencing techniques (Choi *et a*l. 2009).

As mentioned above, exomes are captured by an array based technique, where genomic DNA is randomly fragmented and ligated to specific linkers in order to become amplifiable are hybridized to exon tiling arrays. The captured fragments are

then ligated with specific Illumina 1G linkers and then are amplified. The amplified product is placed into a flow cell of eight lanes and an in situ amplification occurs in order for clusters to be generated. As in other next-generation sequencing techniques, after every incorporation of fluorescent nucleotide light signal is obtained and the nucleotide is identified (Hodges *et al*.2007).

Exome sequencing has been used in published studies as a clinical diagnostic tool for the identification of rare diseases. Patients who were suspected to have Bartter syndrome had been examined with exome sequencing which found a homozygous missense D652N mutation at a position in SLC26A3 which is an extremely conserved position among all invertebrates and vertebrates and also, is the known congenital chloride diarrhoea locus. The unexpected genetic diagnosis in that study of a patient with an undiagnosed illness illustrates the clinical implications of exome sequencing as a powerful diagnostic tool (Choi *et al*. 2009).

Additional evidence about the efficiency of the method was illustrated by studies which proved that with exome sequencing candidate genes for Mendelian disorders can be identified by sequencing a small number of unrelated, affected individuals as illustrated in Figure 6. That study included 12 individuals, eight Hapmap individuals and four with Freeman- Sheldon syndrome, which is a rare dominant disorder. The comparison of exome sequencing with previous genome sequences revealed that the technique can be as accurate and efficient as whole genome sequencing. Also it has been shown that the combination of exome sequencing with the use of bioinformatic filters can achieve the direct identification of the gene of interest. The exomes from the HapMap individuals were used as filters and in addition to the removal of common variants, they ended up with a single gene MYH3 which has been shown from previous studies to be the causative gene for Freeman-Sheldon syndrome (Ng *et al*. 2009).

Figure 6. The stepwise filtering process for the identification of the candidate gene

A similar method has been used for the discovery of the gene which is responsible for the Miller Syndrome, a rare Mendelian disorder. Four affected individuals were examined, in three independent kindreds. A filtering approach was used, in order

to exclude a vast number of variants which were unlikely to be causative and to focus on a small number of candidate genes. They identified eight candidate genes under a dominant model and only one (DHODH) under a recessive model. Additional screening of three unrelated people with the disorder and a sibling of the second revealed a total of 11 mutations in 6 kindreds with Miller syndrome. Those results were solid evidence for the identification of DHODH as causative gene of Miller syndrome. That was additional evidence for the potential of exome sequencing not only as a diagnostic tool but as a powerful technique for the discovery of disease causative genes (Ng et al. 2010). The technique of exome sequencing was also used in cancer where new somatic mutations have been identified. (Choi *et al*. 2009)

Exome sequencing might be efficient in identifying genes for rare mutations which are located in protein coding regions, but cannot identify any structural variants and noncoding variants which can be found by whole genome sequencing. Finally, it has been illustrated above that has been already been used as successful diagnostic tool and with the improvement of the

technique event further it could have implications in medical care.

1.3 Assessing copy number

1.3.1 Multiplex Amplifiable Probe Hybridization

In order for specific DNA sequences to be analysed by next generation sequencing, they have to be identified and captured and methods for measuring copy number can be used to assess the effectiveness of enrichment. One of the most efficient PCR based methods for the identification of copy number variants is the Multiple Amplifiable Probe Hybridization (MAPH). This is a PCR based technique which relies on the principle of the DNA complementarity as illustrated in Figure 7. In MAPH the genomic DNA is fixed on a nylon filter. MAPH probes for any gene are designed to be of different sizes, in order to be able to be distinguished after gel electrophoresis and are flanked by the same set of primers so they can all be amplified together. MAPH probes hybridize with genomic DNA,

binding to the complementary sequences on the filter and the unbound DNA is washed with stringent solution. The hybridized DNA is then denatured with the use of either high temperature or addition of NaOH. The hybridized strands are thus released into a buffer solution. The released strands are amplified and with the use of gel electrophoresis the hybridized sequences can be identified and analyzed for the detection of deletions or duplications (den Dunnen and White 2006, Hollox *et al*. 2002, Patsalis *et al*. 2005, Sellner and Taylor 2004, Tyson *et al.* 2009 and XueiMei and HuaSheng 2009).

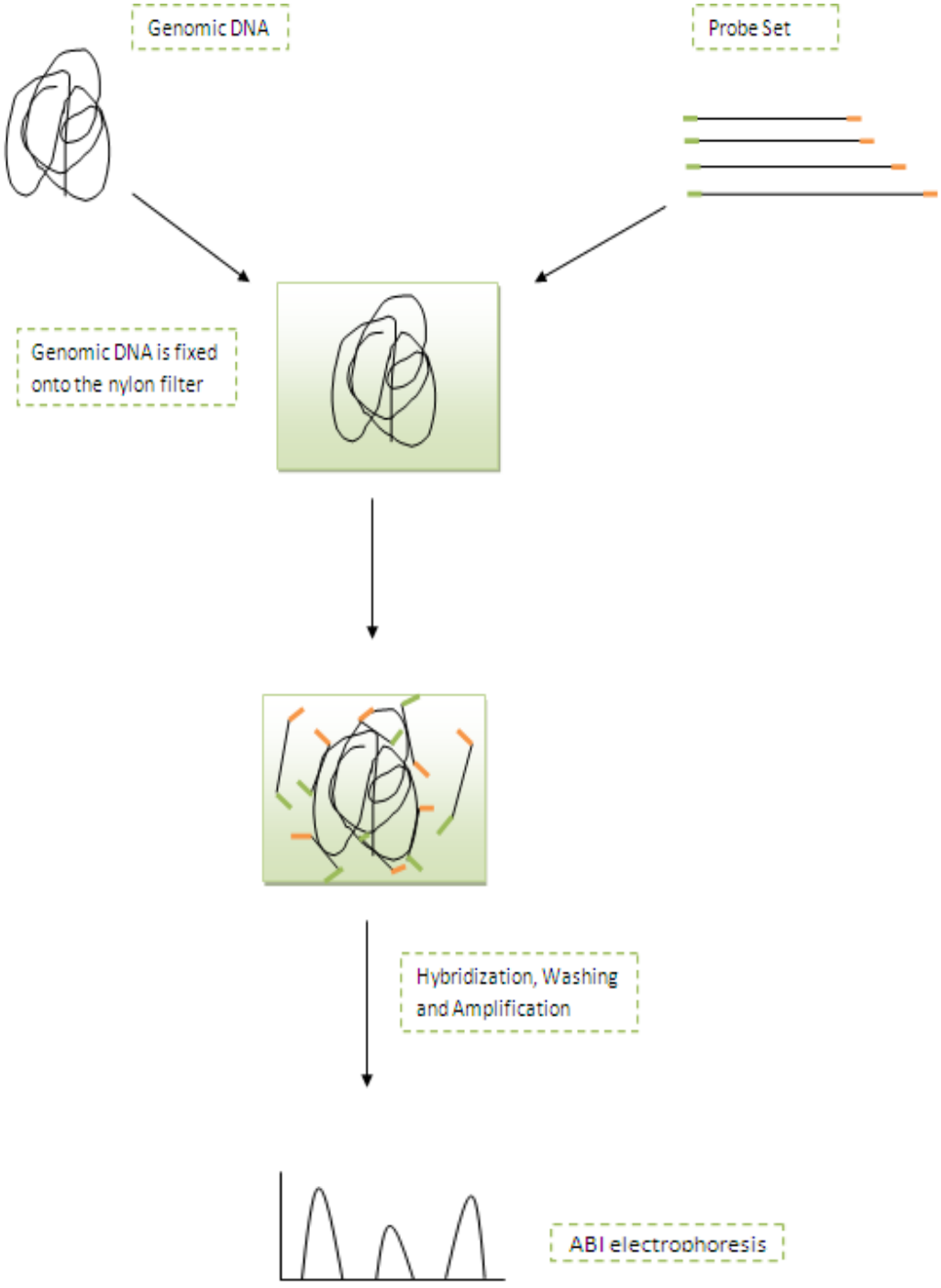


Figure 7. In multiplex amplifiable probe hybridization, genomic DNA is fixed onto nylon filters and hybridized with probes which contain the sequence of interest. The washing step is vital for the removal of background DNA and capture only of the hybridized probes. The

captured sequences are then analyzed by ABI electrophoresis for the identification of potential deletions or duplications.

For the identification of the possible deletions or duplications ABI signals are used for the estimation of the normalised ratios for every probe. Controls which were unaffected had a normalized mean of 1.0. On the other hand ratios below 0.6 and over 1.5 were considered as deletions and duplications respectively (Hollox *et al*. 2002).

1.3.2 MAPH as a clinical diagnostic tool

MAPH can be used as useful diagnostic tool for genetic diseases where the only mutation is copy number change. Significant examples are the deletions or duplications in DMD and BRCA1 genes which have been shown to be responsible for causing Duchenne muscular dystrophy and breast cancer respectively. Also, deletions or duplications of subtelomeric DNA are highly associated with congenital abnormalities and mental retardation (Hollox *et al.* 2002). Hereditary non-polyposis colorectal cancer (HNPCC) has been found to be

caused by mutation in four mismatch repair genes (MLH1, MSH2, MSH6, PMS2). However, MLH1 and MSH2 contain approximately 90% of the mutations that have been detected in HNPCC patients (Kohlmann and Gruber 2006). Therefore, MAPH was evaluated in its efficiency to identify and discover potential deletions or duplications in HNPCC patients. The study included 73 control samples which were healthy individuals and had no known deletions or duplications and 50 samples which derived from hereditary non polyposis colorectal cancer (HNPCC) patients which had no known point mutations according to previous tests with other methods. The study identified exonic deletion in 10 of the samples. Three independent mutations and a deletion of exon 3 of the MSH2 gene have also been found which have been identified also in previous studies. Moreover, a deletion has been detected of exon 13 of MLH1 gene which has been also found in three individuals which were relatives of the patient. Three of the patients were discovered to have a deletion in exon 8 of MSH2 which is considered to be a frequent deletion among HNPCC patients as it has been found in other HNPCC patients in previous studies. This study certifies the quality and the value

of the results using MAPH technique and proves that MAPH is capable for the identification of potential deletions and duplications in comparison with the other modern methods available (Akrami *et al.* 2005).It has been noted that early diagnosis is vital for the treatment of cancer, and HNPCC gene tests are an important tool for early diagnosis, therefore MAPH can contribute so the gene test can be more specific and to include the maximum number of potential mutations for the best diagnosis (The Johns Hopkins University 1995).

MAPH offers a variety of advantages compared to other modern techniques. It is a relatively simple method which does not require the use of complex technology; it is cheap and is fast with only few days required for a single experiment (Patsalis *et al*.2005 and Hollox *et al*. 2002). The major advantage of MAPH is the great number of loci that can be analyzed in a single experiment. An amplification with many probes flanked by different primer pairs results in low quality or non-amplification. That problem is overcome by generating a set with different probes flanked by the same set of primers (Armour et al. 2000 and White et al. 2002). Furthermore, MAPH probes have been shown to be particularly not susceptible to

substitutional polymorphisms, which make them ideal for copy number variation identification (Sellner and Taylor 2004).

1.3.3 Disadvantages

MAPH is a convenient technique for copy number detection, but has also several limitations. The major drawback of the technique is the limited number of probes that can be analyzed in a single experiment due to gel electrophoresis. A number of 100 probes distinguished by 5bp interval have been previously analysed simultaneously (Patsalis *et al*. 2005).

1.3.4 Other Methods for copy number determination

Alternative methods include microarray based techniques such as comparative genome hybridization (array-CGH) and representational oligonucleotide microarray analysis (ROMA) and PCR based techniques such as multiplex ligation dependent probe amplification (MLPA) and quantitative multiplex PCR of short fragments (QMPSF).

1.3.5 Array-based comparative genomic hybridization (aCGH)

In array CGH, the reference and test genomic DNA which contain the sequences of interest are fluorescently labelled with different fluorophores and are then hybridized to arrays which contain DNA fragment from various loci in a competitive manner. Potential deletions or duplications are then identified from resulting hybridization signal intensity of the arrays (Kousoulidou *et al*. 2008 and Feuk *et al*. 2006). The arrays are usually constructed by oligonucleotides from 60 to 100 bp or by large genomic clones such as BACs with the size of approximately 150 kb (XueiMei and HuaSheng 2009 and De Lellis et al. 2007). The major advantage of the technique is its ability to assess a great number of loci in a single experiment (Vissers *et al*. 2003). On the other hand the large size of BAC probes leads to decreased resolution and in combination with its high cost makes the technique inconvenient as a diagnostic tool (XueiMei and HuaSheng 2009 and De Lellis *et al*. 2007).

1.3.6 Representational oligonucleotide microarray analysis (ROMA)

The principle of representational oligonucleotide microarray analysis (ROMA) is similar with the array CGH technique, the major advantage in this technique is that the samples are simplified before the hybridization process. The genomic DNA is first digested with the use of restriction enzymes. The fragments are then ligated with specific adaptors which make them amplifiable. The amplification step is vital for the process as is the stage where the hybridized DNA is simplified. In that process due to polymerase enzyme, only fragments with maximum length of 1.2 kb are amplified successfully, therefore the DNA is simplified by excluded sequences with higher number of bases. The test and reference genomic DNA is fluorescently labelled and hybridized in arrays which contain loci specific sequences. ROMA is sensitive technique which can detect copy number changes of even 30kb in the human genome. On the other hand, one of the drawbacks of any array technology is that it cannot identify balanced rearrangements in the human genome (Kousoulidou et al. 2008, XueiMei and HuaSheng 2009 and Feuk *et al*. 2006).

1.3.7 Multiplex ligation dependent probe amplification (MLPA)

Other PCR-based techniques include multiplex ligation dependent probe amplification (MLPA) and quantitative multiplex PCR (QMPSF). In MLPA, genomic DNA is hybridised in solution containing sets of probes. These probes are not amplifiable at early stage but are becoming amplifiable later stages. Every set consists of two halves. Each half comprises a 20 to 30 nucleotide long target specific sequence which is flanked by universal primer sequence. In the middle of the two halves there is a fragment of random length which is used to differentiate the size of the probes so they can be used for electrophoresis. The probes bind adjacent to the target DNA and then are joined together by a ligase. Once the probes are adjacent to each other and ligated, are becoming amplifiable too. The unbound probe halves are unable to be amplified; therefore no washing step is required. With the use of MLPA, copy number differences up to 50 regions can be identified (Van Eijk *et al*. 2010). The basic advantages of MLPA over MAPH are the relatively small amount of DNA that is required which can be only 20ng compared with the 1μg of DNA in

MAPH, the lack of washing step which makes the technique simpler and less susceptible to contaminations and fact that MLPA probes become amplifiable only after the ligation step which decreases the contamination susceptibility even more. It is a low cost and easy in application technique and the kits that which are provided by the companies reduce the time that is required for the completion of the process to minimum (De Lellis *et al*. 2007). Also, it is capable in detecting not only large copy number changes but small mutations too (Bunyan *et al*. 2004). However, the generation of MLPA probes is much more complicated than for MAPH probes (Sellner and Taylor 2004 and Feuk *et al*. 2006).

1.3.8 Quantitative multiplex PCR of short fragments (QMPSF)

The quantitative multiplex PCR of short fragments (QMPSF) is another PCR- based technique for the identification of deletions or duplications. In that technique, labelled primers are used to flank short locus specific genomic DNA sequences. The primers are mixed with genomic DNA and quantitative PCR

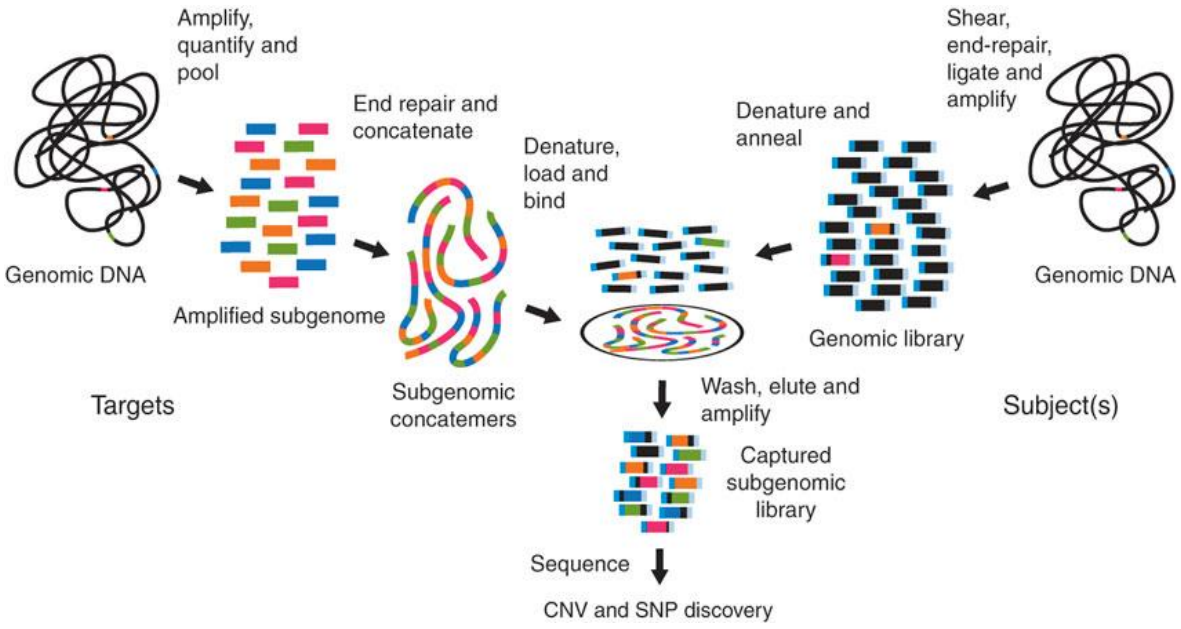
occurs, which is stopped in log phase and gel electrophoresis can identify possible deletion or duplications for each target region (Feuk *et al.* 2006 and De Lellis 2007). With the use of short fragments the amplification accuracy is increased and therefore better comparison can be achieved. Also, the multiple PCR where the fragments are amplified simultaneously contributes to the better comparison of the resulted peaks. (Charbonnier *et al*. 2000).

1.4 Project Overview

This project was based on studies which have used filter-based hybridization enrichment technique for subgenomic targets to be captured. In Herman *et al*. (2009), filter based hybridization was used for the detection of potential single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) in genes associated with hypertrophic cardiomyopathy (HCM) and high density lipoprotein (HDL) cholesterol.

The principle of filter based hybridization was for the target DNA sequence which was to be analysed to be fixed on a nylon

filter, so the complementary amplifiable sequences from genomic libraries after hybridization and stringent washing could be trapped and further analysed for CNV detection as seen in Figure 8.



(Herman *et al*. 2009)

Figure 8. Filter based hybridization:

Target DNA which was in form of concatemers was fixed on a nylon filter. Then genomic DNA is sonicated and cut in pieces approximately 150 to 250bp length. The small fragments were then blunt ended, phosphorylated and ligated to each other containing adaptors in between with 2bp identifying barcodes in order to become amplifiable. The ligated amplimers are then denatured and

loaded to the appropriate filters. The filters are washed with stringent solution for the removal of unbound material. Once the bound sequences are selected, they amplified and sequenced for the identification of possible SNPs and CNVs.

The results identified 289 SNPs, 238 of which were identified in previous studies. Approximately 20 kb of the subject HCM1 had been previously sequenced and illustrated complete similarity with the sequences analyzed. The results illustrated that the technique is powerful in detecting CNVs and more specifically deletions over 32bp and insertions over 64bp with sensitivity of more than 95%. Also, from the fragments captured it has been found that ~58%of the total sequences captured for HCM were sequences of HCM genes and ~67% from the fragments captured for HDL were sequences of HDL genes, which was a very good level of enrichment (Herman *et al*. 2009).

In this project, a filter based hybridization technique was used followed by MAPH analysis in order to assess the ability of the technique to enrich subgenomic targets. That has been achieved with the application of two rounds of enrichment. The

level of enrichment has been estimated and with the use of site-specific primers followed by various PCR cycles, the quantity of the enriched product has been estimated too. The results illustrate that high a level of enrichment can be achieved with the use only of few pg of enriched product and that with the combination of those methods, specific targets can be captured from genomic DNA with accuracy and sensitivity.

Chapter 2

Materials and Methods

2. 1 Generation of Amplifiable linkered genomic DNA

Genomic DNA from two individuals (TT297 and AF103, a male and a female respectively) was sonicated by Covaris sonicator. In the sonication process, genomic DNA was sheared to random size fragments between 100 to 900bp. The sonication process used a, 5% duty cycle, intensity 3 and cycle/burst intensity of 200 at 4 °C, two times for a cycle of 45 seconds. Once the two genomes were sonicated, 100 µl of each DNA sample were added into a Covaris microtube (6x16 mm AFA fiber with snap cap) with the addition of 1µl of 100 times TE buffer(1M Tris-Cl, pH 8 and 100 mM EDTA).

2.1.1 Size estimation

A 2% agarose gel was prepared in order for the size range of the sonicated fragments to be estimated. For the agarose gel generation, 100ml of 0.5 x TBE buffer (250ml of 10x TBE, 500µl of 5mg/ml Ethidium Bromide, 5L of deionised water) were mixed with 2g of agar. For the gel electrophoresis for each sample, 5µl were mixed with 2µl of loading dye (0.025% bromophenol blue dye, 40% sucrose, 25x TBE buffer). The 2% agarose gel was electrophoresed at 100V for approximately 1 hour.

2.1.2 End Repair

Sonication sheared DNA fragments randomly; therefore not all blunt ends were generated. In order for the ends of the fragments to be repaired, phosphorylated and blunted, 20µl of 10x NEB buffer 2(New England Biolabs), 8µl of 10µM dNTPs (New England Biolabs), 2µl of 100µM ATP, 40 units of T4 DNA polymerase (New England Biolabs) and 80 units of T4 PNK (New England Biolabs) were mixed with 48.7µl of deionised

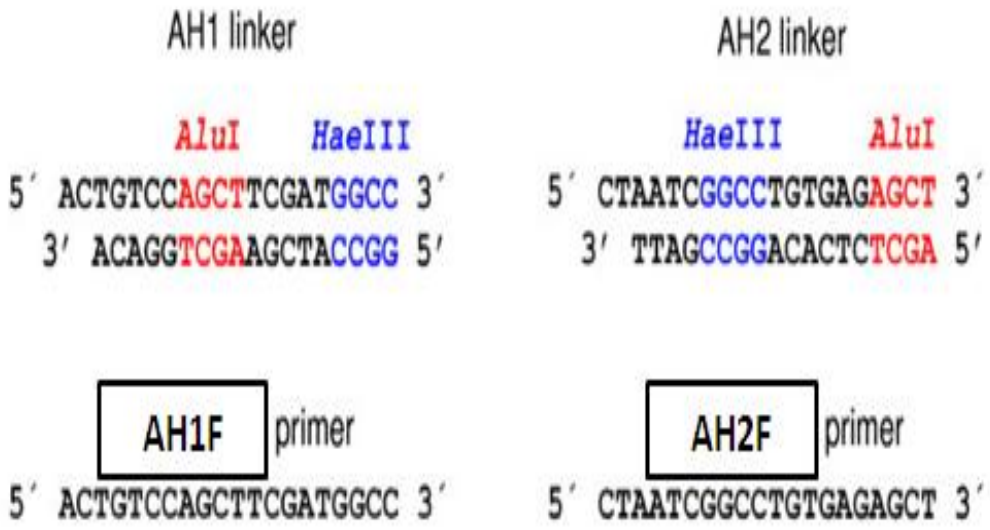
water in total volume of 100µl. The samples were incubated at 37 °C for 30 minutes.

2.1.3 PCR purification

The two genomic DNA samples (100µl) were purified by QIAquick PCR purification kit (Qiagen). Therefore, each 100µl of genomic DNA was mixed with 500µl of buffer PB. The mix was then applied onto a QIA quick spin column and centrifuged for 1 minute at 13,000rpm (maximum speed). The flow- through was discarded and then 750µl of buffer PE were added and centrifuged for 1 minute. The flow through was discarded and the tube was centrifuged again for 1 minute with the remaining flow through to be discarded again. The QIAquick column was then placed on a new eppendorf tube and 30µl of elution buffer EB were loaded and the tube was centrifuged for 1 minute. Once the PCR purification was completed, a gel electrophoresis using 2% of agarose gel was repeated in order for the efficiency of the purification process to be checked.

2.1.4 Ligation

The purified genomic DNA samples were then ligated with linkers in order to become amplifiable as indicated in Figure 9. From the gel electrophoresis it was estimated that the mix consisted of approximately 6μg of genomic DNA. Therefore for the ligation, 10μl of purified solution (containing 2μg of genomic DNA) were mixed with 2.5μl of 10x ligation buffer (NEB recipe), 2.5μl of 10mM ATP, 0.5μl of 500μM Linker AH1, 0.5μl of 500μl Linker AH2, 2μl (2.5units) of T4 DNA ligase (New England Biolabs) and 7μl of deionised water in a total volume of 25μl. Three mixes were prepared, one for TT297 genomic DNA, one for AF103 and one with only deionised water as control sample. The three samples were left at room temperature for an hour and then placed at 4°C for overnight incubation.



Tyson et al. 2009

Figure 9. The figure illustrates the sequences of the AH1 and AH2 linkers with the restriction enzymes sites indicated. The sequences of interest were ligated with AH1 and AH2 linkers in order to become amplifiable and AH1F and AH2F primers were used for the PCR amplification of the fragments.

2.1.5 Ethanol Precipitation

The samples from ligation were ethanol precipitated. To each 25µl of ligated sample 0.5µl of 10mg/ml tRNA were mixed to act as a carrier, followed by the addition of 75µl of 100% ethanol. The mixes were placed on ice for approximately 10 minutes and were then centrifuged for 10 minutes at 13,000

rpm. The supernatant was discarded and the tubes were centrifuged for an additional minute. The remaining supernatant was removed again and then 150μl of 80% ethanol was added. The tubes were centrifuged for 10 minutes at maximum speed and the supernatant was discarded. The remaining pellet was resuspended with the addition of 10μl of deionised water.

2.1.6 Amplification

In the next stage, the precipitated material which consisted of 2μg of ligated genomic DNA was diluted. Specifically, 1μl of the precipitated material was mixed with 10μl of deionised water to result in 20ng/μl of genomic DNA. Three samples were prepared for the amplification, TT297 genomic DNA, AF103 genomic DNA and a third sample which was control and contained deionised water. The samples were amplified using AH1F (ACTGTCCAGCTTCGATGGCC) and AH2F (CTAATCGGCCTGTGAGAGCT) primers. For each sample a total mix of 19μl was prepared consisting of 2μl of 10x AB gene PCR buffer IV, 0.8μl of 25mM $MgCl_2$ (AB gene) , 0.16μl of 25mM dNTPs, 1μl of 10μM AH1F primer, 1μl of 10μM AH2F

primer, 0.2µl of 5U/µl Taq polymerase (New England Biolabs) and 13.84µl of deionised water. For the amplification of the samples 25 PCR cycles were used. The samples were incubated before the amplification at 70 °C for 1 minute for the polymerase enzyme to fill in the ends of the fragments, followed by a 95°C step for 1minute, 60°C for 1 minute and 70°C for 1minute for 25 cycles. The PCR product was then tested by gel electrophoresis using 2% agarose gel.

2.1.7 PCR Efficiency

Once the PCR at 25 cycles was checked and it was ensured that 25 cycles were sufficient for the generation of significant amount of product, different conditions were applied by repeating the same PCR in order to improve the yield. The PCR tests were of 4 different kinds, with AB x10 buffer IV (AB gene), with 10x PCR buffer (500mM Tris HCl pH 8.8, 120mM Ammonium sulphate, 50mM $MgCl_2$, 74mM 2-mercaptoethanol, 11mM dATP, 11mM dCTP, 11mM dGTP and 11mM dTTP and 1.25mg/ml BSA), 10 x PCR buffer as above but with 0.4µl of HiDi formamide and the last was 10x PCR buffer with 2µl of

glycerol. The test was repeated for both genomic DNA samples (TT298 and AF103) at 25 cycles. Once that test was completed the same process was repeated but during the PCR the 95°C step instead of 1 minute it was used 30 seconds. Also, series of tests were prepared using the same PCR conditions at different PCR cycles.

2.1.8 Quantification

In order for the yield of the product from the PCR cycle tests to be estimated, herring sperm DNA was used. Herring Sperm DNA was first cut by *Hae*III restriction enzyme in order to generate a smear after gel electrophoresis. For the restriction-cut DNA three different sets were prepared where, 2µl of 10x NEB buffer2, 2µl/5µl/10µl of 10mg/ml Herring sperm DNA, 0.5µl of 5 units of *Hae*III restriction enzyme and 12.5µl/7µl/3µl of deionised water were mixed and incubated at 37°C for 30 minutes. Once the restriction digest was completed, 0.2µl of Herring sperm DNA was diluted into 20µl of deionised water then 0.2µl, 0.5µl and 1µl which are (equal to 200ng, 500ng and 1µg of DNA) were loaded on a 2% agarose gel with the

genomic DNA from the previous PCR cycle tests in order for their product to be quantified.

2.2 Multiplex amplifiable probe hybridization (MAPH)

Genomic DNA or enriched samples from PCR were purified by QIAquick PCR purification Kit (Qiagen) and then ethanol precipitated. Once the precipitation was completed, 1µl of 1M NaOH was added to each sample.

2.2.1 Filter preparation

For the hybridization, 10 ~2x4mm nylon filters were cut and labelled. The genomic DNA which was previously precipitated was loaded onto the filters (~1µg). The DNA was then bound to the filter after exposure to 50mJ of UV radiation (Armour *et al*. 2000).

2.2.2 Filter pre-hybridization

Filters were prehybridised together in a solution which contained 1ml of prehybridization solution (0.5M sodium phosphate pH7.2, 7% SDS, 0.1 mg/ml alkali-denatured herring sperm DNA). The filters incubated at 65 °C for 2 hours. That solution was then replaced by 200ml of prehybridization solution which contained 10µg/ml of denaturated human Cot-1 DNA (Armour *et al*. 2000).

2.2.3 Hybridization

The filters were incubated further at 65°C for approximately 40 minutes and then were mixed with 1µl of 10mg/ml human Cot-1DNA (Invitrogen), 1µl of 10mg/ml of Herring Sperm, 2µl of 250µg/ml of ΦX174/HaeIII (New England Biolabs), 1µl of end blocking primers PZA and PZB and 1µl of probe mix HNPCC which was first denatured by the addition of 2µl of1M NaOH followed by the addition of the mix to the filters. After denaturation, the probes were incubated for 1 minute at 37°C and then were placed straight on ice followed by addition of 3µl

of 1M Na_2PO_4. The probes were the filters then incubated overnight at 65°C (Armour *et al*. 2000).

2.2.4 Washing

The filters were washed with the use of two different hybridization solutions. The first washing step was with 500ml of 1x SSC/1% SDS, followed by washing with 500ml of 0.1x SSC/0.1% SDS. Once the washing was completed, each filter was transferred into 50µl of 1x buffer IV (ABgene) (Armour et al. 2000).The tubes were then incubated for 5 minutes at 95°C to release bound probes for amplification. For the amplification, for each sample, 2µl of 10xPCR mix, 0.8µl of 10µM HexPZA (Hex-AGTAACGGCCGCCAGTGTGCTG) primer, 2µl of 10µM PZB (CGAGCGGCCGCCAGTGTGATG) primer, 0.4µl of 5000U/ml Taq polymerase (New England Biolabs) and 14µl of deionised water were mixed. The PCR consisted of 3 steps; the first was 30 seconds at 95°C, the second 1 minute at 60°C, and the third was 1 minute at 70°C at 25 cycles. Once the PCR was completed the samples were incubated at 70°C for 40 minutes.

2.2.5 ABI Electrophoresis

In order for the captured fragment representation to be evaluated, it was prepared for ABI electrophoresis. A mix of 2µl of ROX500 marker (Applied Biosystems) and 170µl of HiDi formamide per 16 samples was prepared and for each sample 1.5 µl of the PCR product was added to 10µl HiDi/Rox 500 marker. The plate was centrifuged and denatured at 96°C for 3 minutes. The samples were then electrophoresed on an ABI 3100 36cm capillary, with the use of POP-4 polymer and injection of 45 seconds at 1kV.

2.3 Selection Methods

In this experiment, MLH1 and MSH2 exon probes were amplified and 1µg of exonic DNA was loaded onto filters in order to be hybridized with linkered genomic DNA. The captured DNA fragments were then amplified and analysed by MAPH.

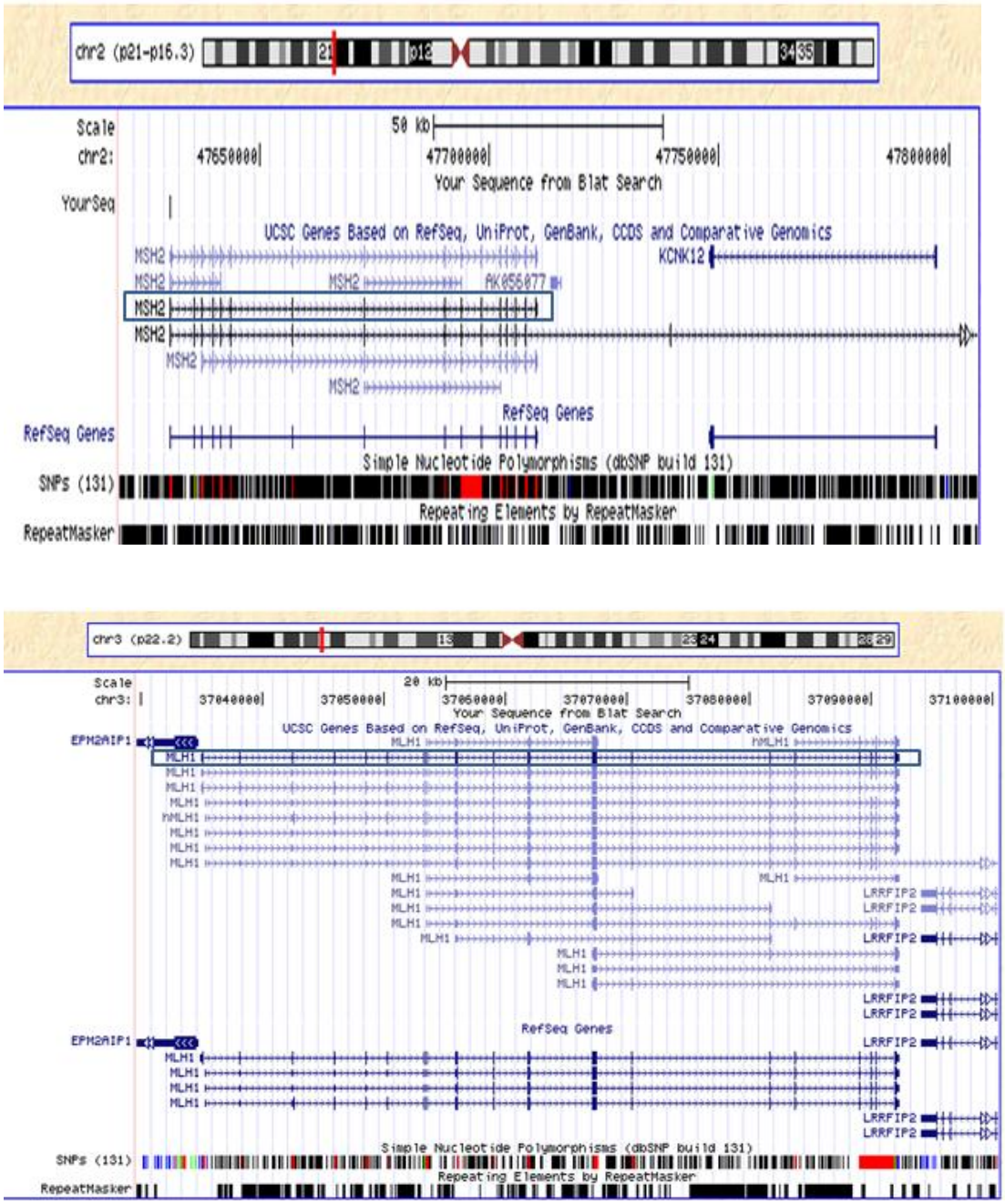


Figure 10. Distribution of exons across the MSH2 and MLH1genes. The exons are illustrated as vertical small lines; adapted from UCSC genome browser.

2.3.1 First Round of Enrichment

The HNPCC MAPH probe set consists of 16 MSH2 probes, 19 MLH1 probes and 1 X and 1 Y linked probes. In this experiment MSH2 and MLH1 probes were amplified individually and mixed into three different groups. For the amplification, for each sample, 1µl of probe DNA, 2µl of 10x PCR mix, 1µl of 10µM PZA primer, 1µl of 10µM PZB primer, 0.2µl of 5000U/ml Taq polymerase and 14µl of deionised water were mixed. After individual amplification all the MSH2 exon probes were mixed together and all the MLH1 together. Also, an additional mix was prepared which contained only the S11 and S13 exonic probes. The three different mixes were then purified by QIAquick PCR purification Kit (Qiagen) and the DNA was loaded onto ~2x4mm nylon filters. For the first step of the experiment, 11 filters were prepared, 3 for MSH2 probes, 3 for MLH1 probes 3 for S11/S13 probes. Once the DNA was denatured loaded to the appropriate filters and crosslinked by 50mJ UV radiation, the filters were placed into 3 different tubes which contained 1ml of prehybridization solution (See section 2.2.2). Each tube contained 1 filter with MSH2 probes, 1 with MLH1 probes and 1

with S11/S13 probes. The first tube was hybridized with TT298 linkered genomic DNA, the second tube with AF103 linkered genomic DNA and the third was a control with no input DNA. The filters were incubated overnight at 65°C and then were washed with the same washing protocol as for MAPH (See section 2.2.4). The filters were placed into 50µl of 1x buffer solution and incubated at 95°C for 5 minutes to release hybridising fragments. The hybridizing DNA was then amplified at 25 PCR cycles using AH1F/AH2F primers. The yield of PCR was checked by gel electrophoresis and then the PCR product was purified by QIAquick PCR purification Kit (Qiagen).

2.3.2 MAPH analysis of enriched DNA

The enriched DNA after purification was denatured with the addition of 2µl of 1M NaOH and was loaded onto the filters. The DNA was then fixed onto the filters by exposure to 50mJ of UV radiation. The samples were prehybridized and then were hybridized with the HNPCC probe set (2ng each probe) (See sections 2.2.2 and 2.2.3). The filters were then stringently washed, placed into 50µl of 1x buffer solution and incubated at

95°C for 5 minutes. The DNA samples were then amplified at 25 cycles with Hex PZA and PZB primers, followed by incubation for 40 minutes at 72°C. The PCR product was then analyzed by ABI electrophoresis at 1kV for 45 seconds. The DNA from first round of enrichment filters was amplified for 35 cycles. The samples were purified by QIAquick PCR purification Kit (Qiagen). With the use of 2% agarose gel electrophoresis the product was quantified. Then 10x and 100x dilutions were made which were equal to 100ng and 10ng in 3µl of dilution respectively. The filters were prehybridized and hybridized with HNPCC probe set and then stringently washed, placed in 50µl of 1x buffer solution and incubated at 95°C for 5 minutes; the released probes were amplified at 25 cycles followed by 40 minutes incubation at 72°C and prepared for ABI electrophoresis.

2.3.3 Second Round of Enrichment

In the next stage MSH2 and MLH1 probes were prepared exactly as for first round of enrichment in order to be loaded onto filters. For the hybridization, 11 filters were prepared, 3 for

MSH2 probes, 3 for MLH1 probes, 2 for S11/S13 probes and 3 blank filters. In this experiment 1 filter with MLH1 probes was placed into the tube with MSH2 filters and 1filter with MSH2 probes was placed into the tube with MLH1 filters to act as controls. It was estimated that a total of about 1µg of DNA was loaded onto the filters. The DNA was then crosslinked by exposure under 50mJ of UV light. Each tube with filters was prehybridized and then hybridized overnight at 65°C with captured genomic DNA from first round of enrichment. The filters were then stringently washed, placed into 50µl of 1x buffer IV (AB gene) solution and incubated at 95°C for 5 minutes. The captured DNA was then amplified at 30 cycles using AH1F and AH2F primers and purified by QIAquick PCR purification Kit (Qiagen). The PCR product was quantified by 2% agarose gel in gel electrophoresis. The DNA samples were then prepared for MAPH. Filters were generated with the MSH2, MLH1 and S11/S13 probes where only 1ng of DNA was loaded but for the control filters 1µg of the appropriate DNA was loaded. The DNA was crosslinked onto the filters under 50mJ of UV light. The filters were prehybridized and then hybridized with HNPCC probe set (See sections 2.2.2 and

2.2.3). Hybridization was followed by stringent washing of filters, placement in 50μl of 1x buffer and incubation in 95°C for 5 minutes. The captured DNA was then amplified for 25 cycles followed by incubation at 72°C for 40 minutes and prepared for ABI electrophoresis analysis at 1kV for 45 seconds.

2.3.4 Single probe amplification assay

In this experiment, new primers were designed as shown in Table 7, with the use of UCSC Genome Browser and Primer3 software in order for specific regions of captured DNA sequences to be amplified, and the amount of the captured DNA to be estimated and the hybridization efficiency to be determined. For each PCR experiment 2μl of 10x PCR, 1μl of 10μM of the appropriate left primer, 1μl of 10μM of the appropriate right primer, 0.2μl of 5000U/ml Taq polymerase and 14μl of deionised water were mixed.

Primer	Sequence
S12 Left	TTCCTGTGTACATTTTCTGTTTTT
S12 Right	TATACGTCATTAGGAATAAATGAA
S13 Left	AGAAGTTTAAAATCTTGCTTTCTGA
S13 Right	TTCCAACATTTCAGCCATGA
S12-S13 Left	GATGGAGAAAATTCCCAGTTCTT
S12-S13 Right	CATGAGCCTATATGCAAGGCTA
L6 Left	GCCCCAGTCAGTGCTTAGAA
L6 Right	GGTCCTCCACCTGAACAGAA
L16 Left	GATGCTCCGTTAAAGCTTGC
L16Right	GGTCCTCCACCTGAACAGAA

Table 7. The primers which were used for the amplification and capture evaluation of specific regions at MSH2 and MLH1 exons which were used at the selection experiments.

2.4 Nanodrop Spectrophotometry

For the accurate measurement of the MSH2 captured DNA from second round of enrichment, Nanodrop spectrophotometry was applied. For the generation of sufficient DNA product 4 different amplifications were prepared. Each sample contained 2µl of 10x PCR mix, 1µl of AH1F primer, 1µl of AH2F primer, 0.2µl of Taq polymerase, 14µl of deionised water and 1µl of second round of enrichment DNA. The mixes were amplified at 30 cycles followed by 40 minutes incubation at 72°C. The samples were then purified by QIAquick PCR purification Kit (Qiagen). Nanodrop spectrophotometer was then zeroed with the use of 1µl of elution buffer and then 1µl of the purified DNA sample was used for the measurement.

2.5 Cloning of enriched DNA

2.5.1 Ligation

Approximately 31ng of enriched DNA was estimated to be ligated with 50ng of pGEM-T (Promega). In total three ligations

were prepared, the first was with MSH2 enriched DNA the second was with control insert DNA 4ng/µl (Promega) and the third was with deionised water. For the ligation reaction, 1µl of input DNA, 2µl of 10x ligase buffer (500mM Tris-HCl, 100mM $MgCl_2$, 10mM ATP, 100mM Dithiothreitol, pH 7.5), 2µl of 10mM ATP, 1µl of 50ng pGEM-T (Promega), 1µl of 400,000U/ml T4 DNA ligase (New England Biolabs) and 13.5µl of deionised water were mixed for each ligation mix. The samples were incubated at room temperature for 1 hour and then were placed at 4°C for overnight incubation.

2.5.2 Ethanol Precipitation

To each ligated sample 0.5µl of 10mg/ml tRNA(Sigma) was added followed by the addition of 60µl of EtOH. The samples were incubated on ice for 10 minutes and then were centrifuged at 13,000rpm for 10 minutes. The supernatant was then removed from the pellet and the tubes were spun again for 1 minute. The last drops of supernatant were removed and 120µl of 80% EtOH were added to each tube followed by centrifugation at 13,000rpm for 10 minutes. The supernatant

was then removed and the tubes were spun for an additional minute. The last drops of supernatant were then removed and the tubes were left for 5 minutes to air dry before the pellet was redissolved in 15µl of elution buffer.

2.5.3 Transformation

For the transformation E*. coli* TOP 10 cells were defrosted from -80°C. Cells were then washed in 10% glycerol. The supernatant was removed and then the cells were centrifuged again for 7 minutes at 7K at 4° C. The washed cells (40µl) were then electroporated with the addition of 2µl of each precipitated ligation sample at 12kV/cm (100Ω, 25µl). With the electroporation samples were also included a sample with no DNA input, and one with 100pg of plasmid ms3 CE10 as electroporation control. Once the electroporation was completed 500µl of SOC medium was added into each sample and incubated at 37°C for 1 hour. During that time 6 plates were prepared from 300ml of LB agar with 300µl of 50mg/ml ampicillin. To each plate before the addition of the electroporated samples 25µl of 25mg/ml XGal were added in

order for possible lac- recombinants to be identified. To the first two plates 400µl and 100µl of electroporated MSH2 DNA was added respectively, one with control insert DNA, one with water only sample, one with the TOP 10 cells only and one with 100pg of ms3 ce10 as electroporation control sample. Approximately 500µl of the appropriate electroporated samples were added to each plate followed by overnight incubation at 37°C.Once colonies were grown 50ml of SOC were mixed with 100µl of Ampicilin and 5ml of 10x HMFM and from that mix 100µl were placed into each well of a 96-well plate. White colonies which indicated recombinant clones were picked and placed in each well followed by overnight incubation at 37°C.

2.5.4 Identification of cloned inserts

The grown colonies were PCR amplified directly from cultures in the plate. Therefore, as shown in Figure 14, for each sample 2µl of 10x of PCR mix, 1µl of 10µM 1277primer (TGGCGAAAGGGGGATGTGCTG), 1µl of 10µM PGB primer (AGGCGGCCGCACTAGTGAT), 0.2µl of 5000U/ml Taq polymerase 14µl of deionised water and about 0.5µl of bacterial

culture were mixed and amplified at 30 cycles in a three step PCR (95°C for 30 seconds, 60°C for 1 minute and 70°C for 1 minute). The PCR product was checked on a 2% agarose gel, with 10µl of the product were mixed with 5µl of deionised water and 2µl of loading dye. The remainly 10µl of PCR products were purified with the use of AMPure (Agencourt) and eluted in 40µl of deionised water.

Figure 14. Illustration of the position of 1277 and PGB primers in correspondence to an inserted sequence. The PGB region is located just next to the inserted region, whereas the Primer 1277 which is indicated with pink colour is located approximately 140b before the insert.

2.5.5 Sequencing

Approximately 30ng of each purified sample was sequenced with the use of 0.5μl of BigDye Terminator v3.1 mix, 3.5μl of Big Dye sequencing buffer (Applied Biosystems), 0.5μl of 10μM of 1277 primer and 4.5μl of deionised water. The samples were amplified for 25 cycles of 95°C for 30 seconds, 50° C for 15 seconds and 60 ° for 4 minutes. The sequenced samples were then purified by CleanSEQ (Agencourt).

Sequencing was followed by bioinformatic analysis where the captured sequences were aligned with the human genome with the use of UCSC genome browser and BLAST search in order the level of enrichment to be estimated. Sequences had to fulfil certain criteria in order to be considered as on target sequences. The sequences had to be in between AH1F and AH2F primers and to contain the sequences of both primers. The sequences once were aligned were recorded for their position the possible mismatches and the percentage of alignment with the sequence of interest.

Chapter 3

Results

Filter based hybridization was used for the capture of specific subgenomic targets to evaluate its applicability to next-generation sequencing analysis. Subsequently, the degree of enrichment was assessed using MAPH, specific PCR, and (Sanger) sequencing. First, however, amplifiable DNA for enrichment has to be produced from the samples to be tested.

3.1 Amplifiable linkered genomic DNA

After electrophoresis on a 2% agarose gel with a 100bp ladder 500ng/µl (New England Biolabs), the sheared genomic DNA was estimated to be between 100 to 900bp in length, with the greatest concentration of DNA between 300 and 400bp.

3.1.1 PCR Efficiency

The genomic DNA, once it was end repaired and ligated, was tested in order for the optimal PCR conditions to be identified and used in future experiments. The linkered genomic DNA samples were amplified for 25 cycles and it was shown that a good yield of DNA was produced as illustrated in Figure 15.

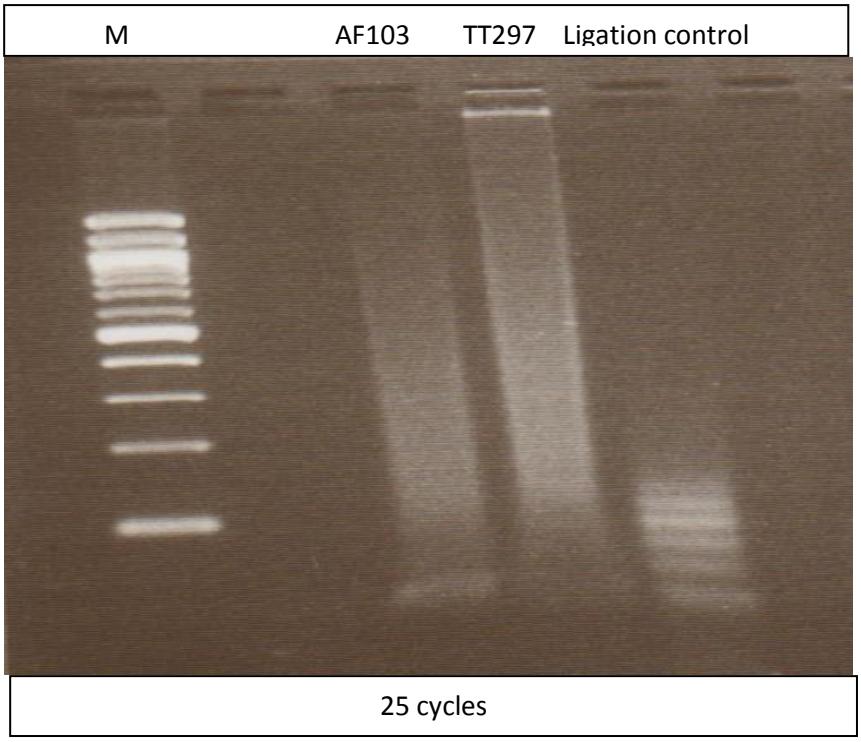


Figure 15. Genomic DNA was amplified for 25 cycles generating significant amount of DNA. The first lane consists of a 100bp DNA ladder.

The samples which had previously amplified well at 25 cycles were used at 25 cycles again but different PCR reagents were

tested as shown in Figure 16. The samples were amplified with the use of AB 10 x PCR buffer IV (ABx gene), 10x buffer (500mM Tris -HCl pH 8.8, 120mM ammonium sulphate, 50mM $MgCl_2$, 74mM 2-mercaptoethanol, 11mM dATP, 11mM dCTP, 11mM dGTP and 11mM dTTP and 1.25mg/ml BSA), 10 x buffer with 10% of HiDi formamide and 10x buffer with 50% of glycerol. Gel electrophoresis showed that the PCRs with the greatest efficiency were with the HiDi formamide and with the BSA input.

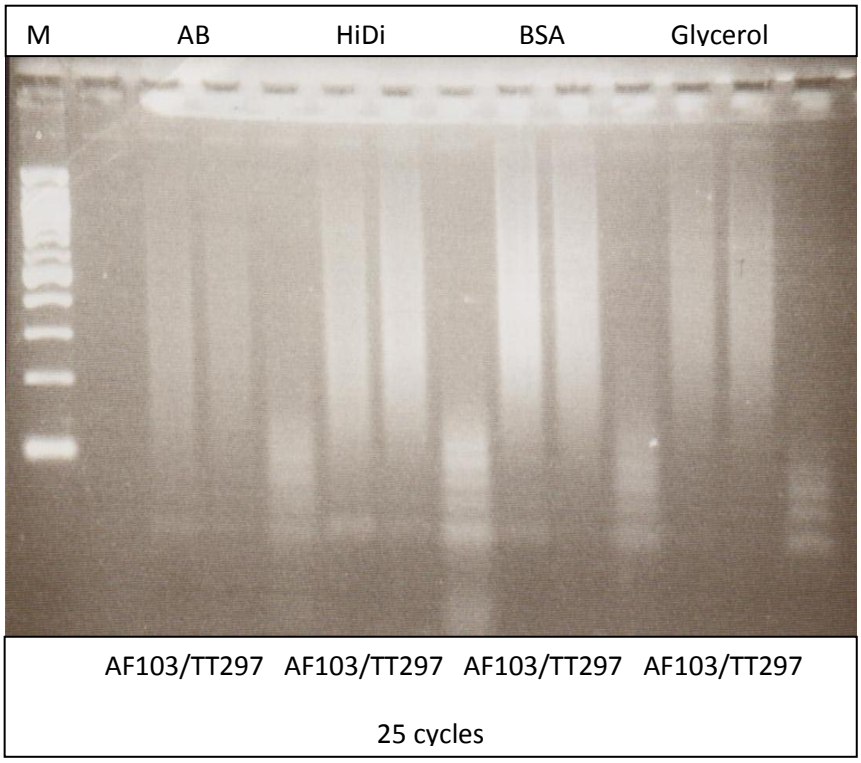


Figure 16. AF103 and TT297 genomic DNA samples were amplified for 25 cycles with the use of different PCR reagents. The lane next to the PCR reagents includes the control of the ligation mix.

The same experiment was repeated but instead of 1 minute at 95 °C at the denaturation step of PCR, 30 seconds were used (Figure 17). With the comparison of the two PCR conditions, it was identified that 30 seconds at 95°C instead of 1 minute increased the yield of product and therefore in future experiments, 10x PCR buffer was used in a three- step PCR where the denaturation step at 95°C was for 30 seconds.

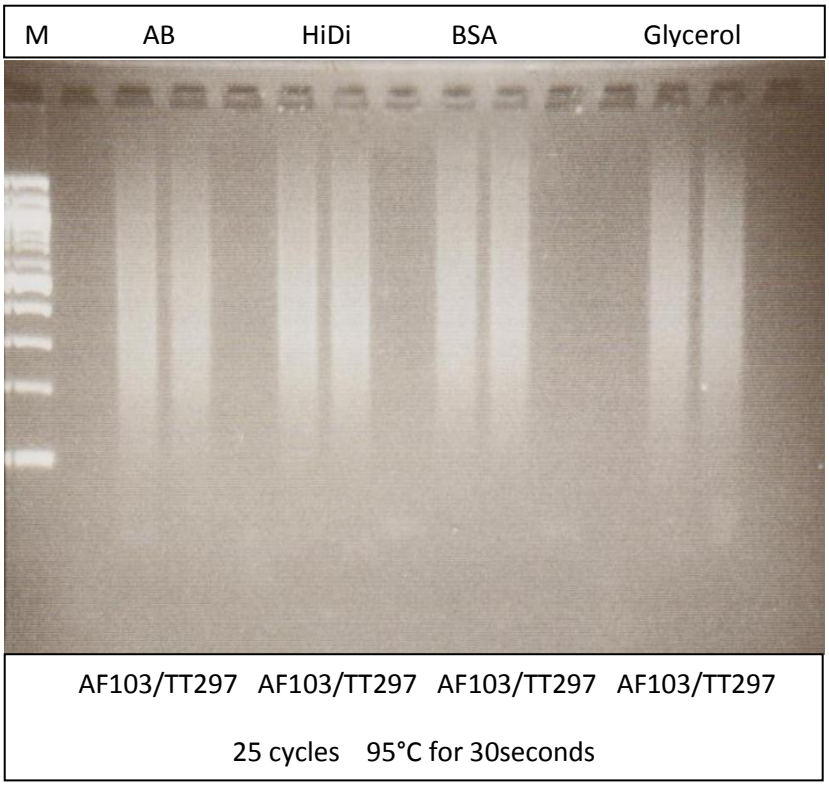


Figure 17. The genomic DNA samples were amplified with the use of different PCR reagents for 25 cycles. The PCR reaction conditions were altered as 30 seconds were used at 95°C instead of 1 minute. The first lane consists of a 100bp DNA ladder.

The samples were also tested using the previous PCR conditions for different numbers of PCR cycles as seen in Figure 18. It was found that 20 cycles were sufficient for the generation of the maximum amount of DNA.

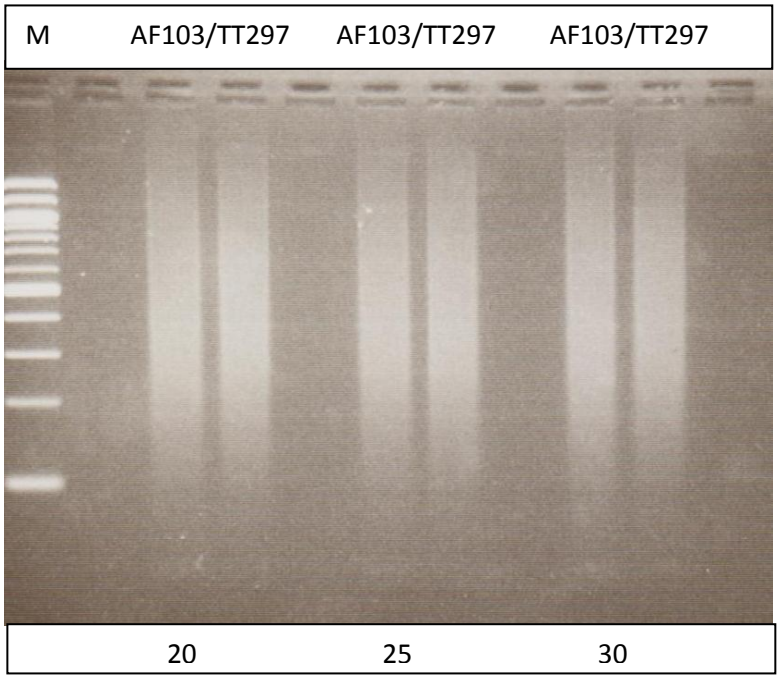


Figure 18. Series of amplifications of genomic DNA samples for 20, 25 and 30 cycles with the use of BSA and 30 seconds at 95°C for PCR conditions. The first lane consists of a 100bp DNA ladder.

3.1.2 MAPH with linkered (unenriched) genomic DNA

In the first experiment, 9 filters were prepared, 2 with TT297 linkered genomic DNA, 2 with AF103 linkered genomic DNA, 4 filters with approximately 1µg of untreated genomic DNA (TT297, AF103, AF105, TT298) and 1 blank filter. The data from the ABI electrophoresis were normalized and the standard deviations of control and of linkered genomic DNA filters were

compared. The standard deviation of the normalized readings of the peak areas from ABI 3100 of the untreated genomic DNA samples was estimated at 0.105 and of linkered genomic DNA at 0.322. The data from linkered genomic DNA varied more than the control genomic DNA, but the extra variation was relatively small. In the second experiment, filters for TT297 and with AF103 linkered genomic DNA were prepared. Filters with TT297 and AF103 linkered genomic DNA contained 1μg of DNA amplified for 10 and 20 PCR cycles. Control filters with untreated genomic DNA were also prepared exactly as in the first experiment. The standard deviation of filters with linkered genomic amplified at 10 cycles was estimated at 1.154, from the 20 cycles filters at 0.238 and from the control genomic DNA at 0.096. This shows that 20 PCR cycles was the appropriate number of cycles for the equal amplification of all the different size fragments and therefore better for hybridization and MAPH experiments. In the third MAPH experiment, filters with 1μg TT297 linkered genomic DNA and with AF103 linkered genomic DNA were prepared with DNA amplified for 25 PCR cycles. In addition to that, filters with untreated genomic DNA were prepared exactly as in the previous experiments. The standard

deviation of the linkered genomic DNA filters was 0.466 and of the control genomic DNA was 0.098. There was variation among the probes, which were not equally represented.

As shown in Figure 19, despite the fact that in 20 PCR cycles linkered genomic DNA has the lowest standard deviation, smaller probes were not represented as well as the rest of the probe set. Even in larger probes the peak height which indicates DNA intensity varies among the probes.

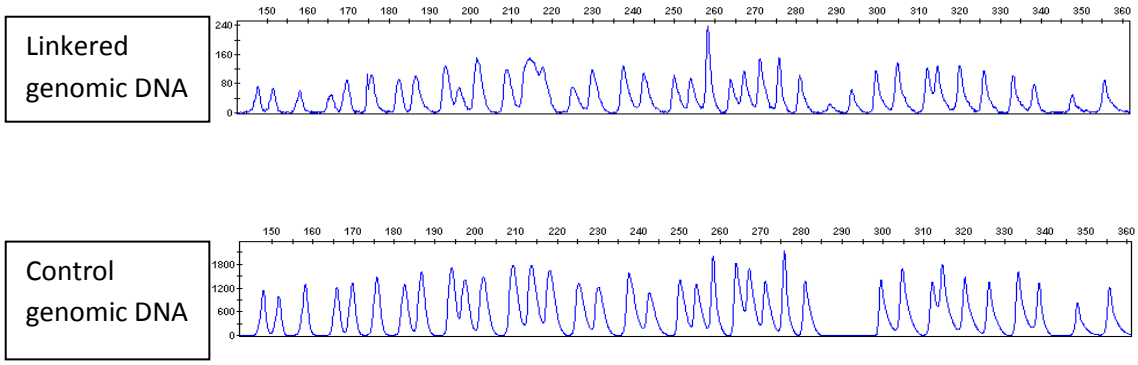


Figure 19. The comparison of the ABI readings of MAPH with linkered genomic DNA amplified for 20 PCR cycles, with untreated TT298 genomic DNA which was used as control sample. The DNA length in bp is illustrated by x axis and the peak height (intensity of DNA) by the y axis.

3.2 Enrichment

3.2.1 First Round of Enrichment

Filter based hybridization (Figure 20) was used as a capture technique in order to collect the sequences of interest from the total genomic DNA for further analysis and evaluation.

In total 11 filters were prepared, 3 for MSH2, 3 for MLH1, 3 for S11/S13 probes and 2 blank filters. Approximately 1μg of target DNA was denatured and loaded on each filter. The filters were hybridized with the denatured linkered genomic DNA, washed and denatured followed by amplification and MAPH as shown in Figures 1 and 2. The purpose of this experiment was the level of enrichment to be evaluated and the percentage of the successfully on target sequences to be estimated. Therefore, different probes were tested for their representation and hybridization efficiency.

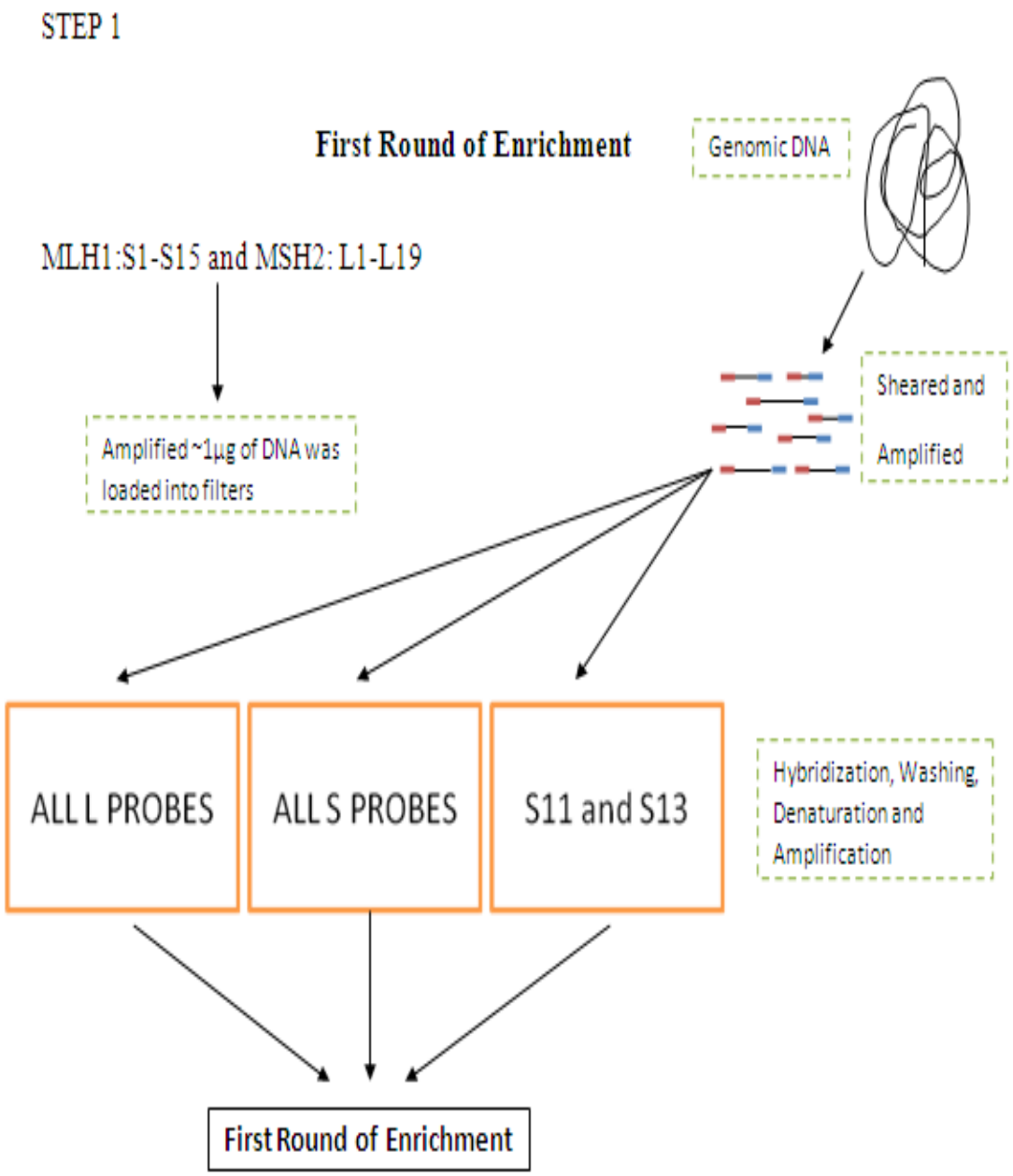


Figure 20. First round of enrichment experiment, MSH2 and MLH1 probes were loaded onto filters and were hybridized with sonicated and linkered genomic DNA. The selected DNA after 5 minutes at 95°C was released into the buffer solution and amplified with AH1F and AH2F primers.

3.2.2 MAPH

The filter based hybridization technique was used for the enrichment of the probes of interest, and MAPH for the evaluation of the enrichment. In previous studies, MAPH has been used as tool for the identification of potential deletions and duplications, but in this project the same method was used as evaluation tool. As can be observed in Figure 21, the same principle was used as described in Figure 7 (Materials and Methods chapter) but instead of untreated genomic DNA, 1μg of enriched genomic DNA from first round of enrichment was loaded onto the filters. MAPH was also repeated by loading onto the filters 100ng and 10ng of enriched genomic DNA. The ABI electrophoresis results and the level of enrichment are illustrated in Table 8. The average readings from 1μg of untreated genomic DNA were used as the control value. The level of enrichment was estimated by the comparison of the average peak height of the captured probes with the average peak height of the control genomic DNA samples. The level of enrichment which is illustrated in the sixth column shows that the use of a smaller amount of captured DNA on the filters for

MAPH enhances full hybridization and subsequently the observed level of enrichment. Also, as can be seen in Figure 22, where an evaluation of the enrichment of MLH1 and MSH2 probes is illustrated, the enrichment was successful with the sequences of interest to be much more enriched than the background DNA. However, the probes were not represented equally as smaller probes were generally underrepresented compared to the bigger probes.

STEP 2

MAPH

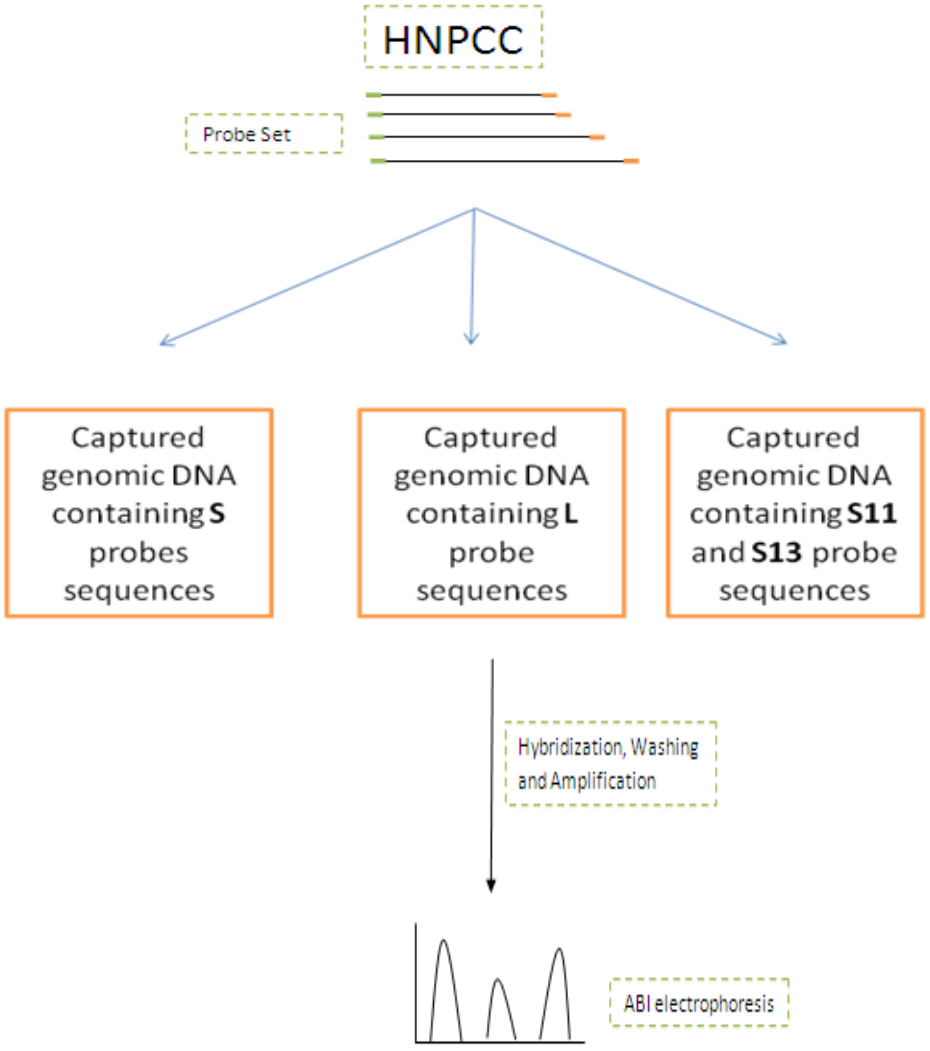


Figure 21. In MAPH, amplified captured genomic DNA from the first round of enrichment was loaded onto the filters and hybridized with the HNPCC probe set. The resulting captured DNA was amplified and analysed after electrophoresis on an ABI 3100.

Quantity	Target DNA	Average	Control Average(Untreated genomic DNA)	Average of Enriched genomic DNA/ Control Average(Untreated genomic DNA)	Enrichment
1µg	MSH2	12149.53	3516.467	3.45	3.45
1µg	MSH2	11846.1	3516.467	3.46	3.46
1µg	MLH1	8069.211	3516.467	2.29	2.29
1µg	MLH1	8291.579	3516.467	2.35	2.35
1µg	S11/S13	59711	3516.467	17	17
1µg	S11/S13	39463	3516.467	11.2	11.2
100ng	MSH2	109052	3516.467	31	31
100ng	MSH2	95927.59	3516.467	27.27	272.7
100ng	MLH1	72176.26	3516.467	20.52	205.2
100ng	MLH1	76441	3516.467	21.73	217.3
100ng	S11/S13	134993	3516.467	38.38	383.8
100ng	S11/S13	126442	3516.467	35.95	359.5
10ng	MSH2	34683.25	3516.467	9.86	986
10ng	MSH2	60901.36	3516.467	17.31	1731
10ng	MLH1	21944.74	3516.467	6.24	624
10ng	MLH1	43293.89	3516.467	12.31	1231
10ng	S11/S13	76658	3516.467	21.79	2179
10ng	S11/S13	57948	3516.467	16.47	1647

Table 8. The level of enrichment from the first round of enrichment experiments. The table illustrates a general overview of the level of enrichment at different DNA concentrations.

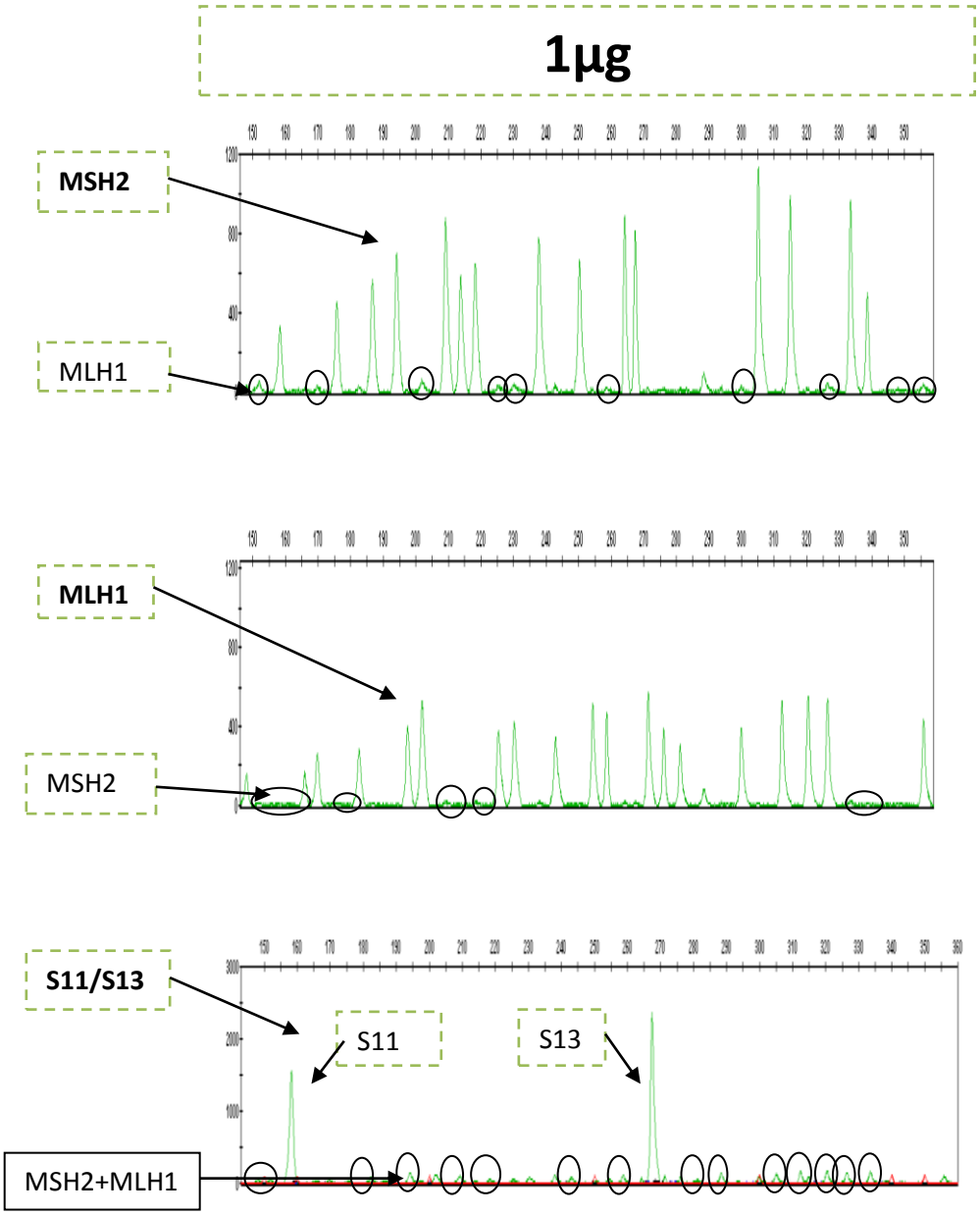


Figure 22. The ABI peaks from MHS2, MLH1 and S11/S13 probes with the use of 1µg of DNA. The x axis represents the DNA size in bp and the y axis the peak height (DNA intensity). It can be clearly seen that there is a difference between the two pictures where different length probes were captured. Therefore at the first two pictures, only MSH2 and MLH1 probes were captured respectively. In the third the

capture of the two probes S11 and S13 can be seen with relatively minor background. The circled peaks on the first picture illustrate the presence of traces of MLH1 probes as background DNA, in the second picture traces of MSH2 probes as background DNA, and in the third picture traces of both MSH2 and MLH1 probes.

3.2.3 Second Round of Enrichment

In this step as shown in Figure 23, captured genomic DNA from the first round of enrichment was used for further enrichment by hybridization, exactly as in the first round of enrichments with MSH2 and MLH1 probes. The filters were washed and denatured followed by MAPH and ABI electrophoresis as shown in Figures 20 and 21.Also, series of dilutions were prepared in order for 1ng, 100pg, and 10pg of second round enriched genomic DNA to be loaded onto filters for MAPH. In the amplified second round of enrichment DNA it was estimated that in every 3μl there are approximately 200ng of total DNA. Therefore 3μl of DNA were diluted into 600μl of deionised water in order for every 3μl to contain 1ng of DNA. Further dilutions were made to result in lower inputs, and in every dilution 1μl of NaOH was added in order for the captured DNA to be

denatured. MAPH results from the second round of enrichment experiment are shown in Figure 24. MSH2 and MLH1 probes can be easily distinguished, although the smaller probes have been found to not hybridize as efficiently as the larger probes, especially the probes at 152bp and 158bp for MSH2; the difference in hybridization efficiency can be seen at 100ng and 10ng of input where larger probes maintain the good signal in contrast with the smaller size probes which almost disappeared. Also, even though the input of DNA was decreased 1000 times and 100,000 times the average signal of the enriched probes remained higher than with 1µg unenriched genomic DNA, which not only indicates the successful enrichment but also the fact that by decreasing the input DNA, the completeness of hybridization is increased. The level of enrichment was estimated and is illustrated in Table 9 where, as in Table 8, the average intensity from untreated genomic DNA was used as a control value. The level of enrichment is illustrated in the sixth column as in first round of enrichment by reducing the amount of DNA input in MAPH enrichment is increased. An enrichment of 7,040,000 in 10pg of S11/S13 input was achieved.

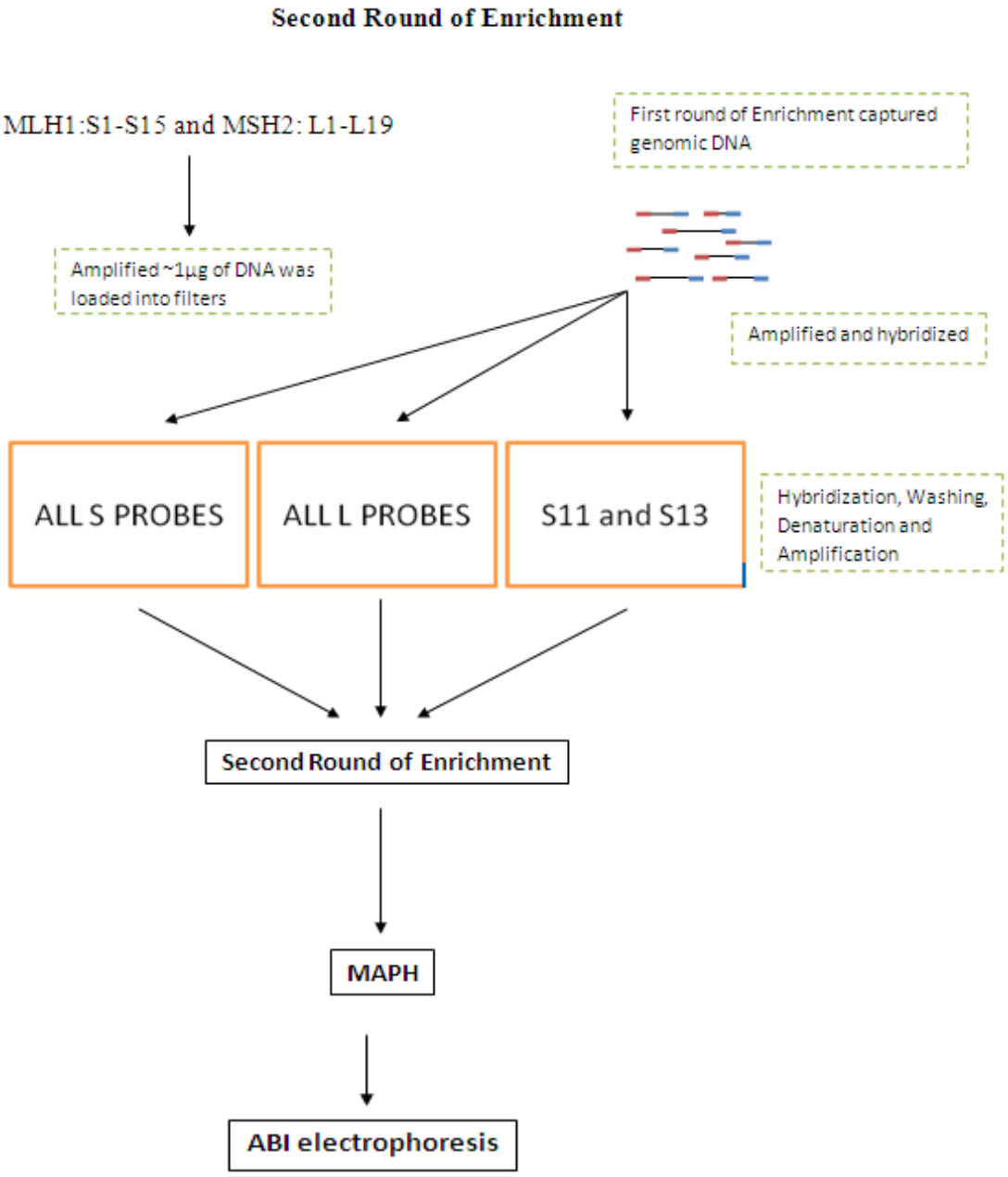


Figure 23. In the second round of enrichment MSH2 and MLH1 probes were amplified and loaded onto filters and hybridized with amplified genomic DNA captured in the first round of enrichment. The captured DNA was amplified and used for MAPH analysis followed by ABI electrophoresis.

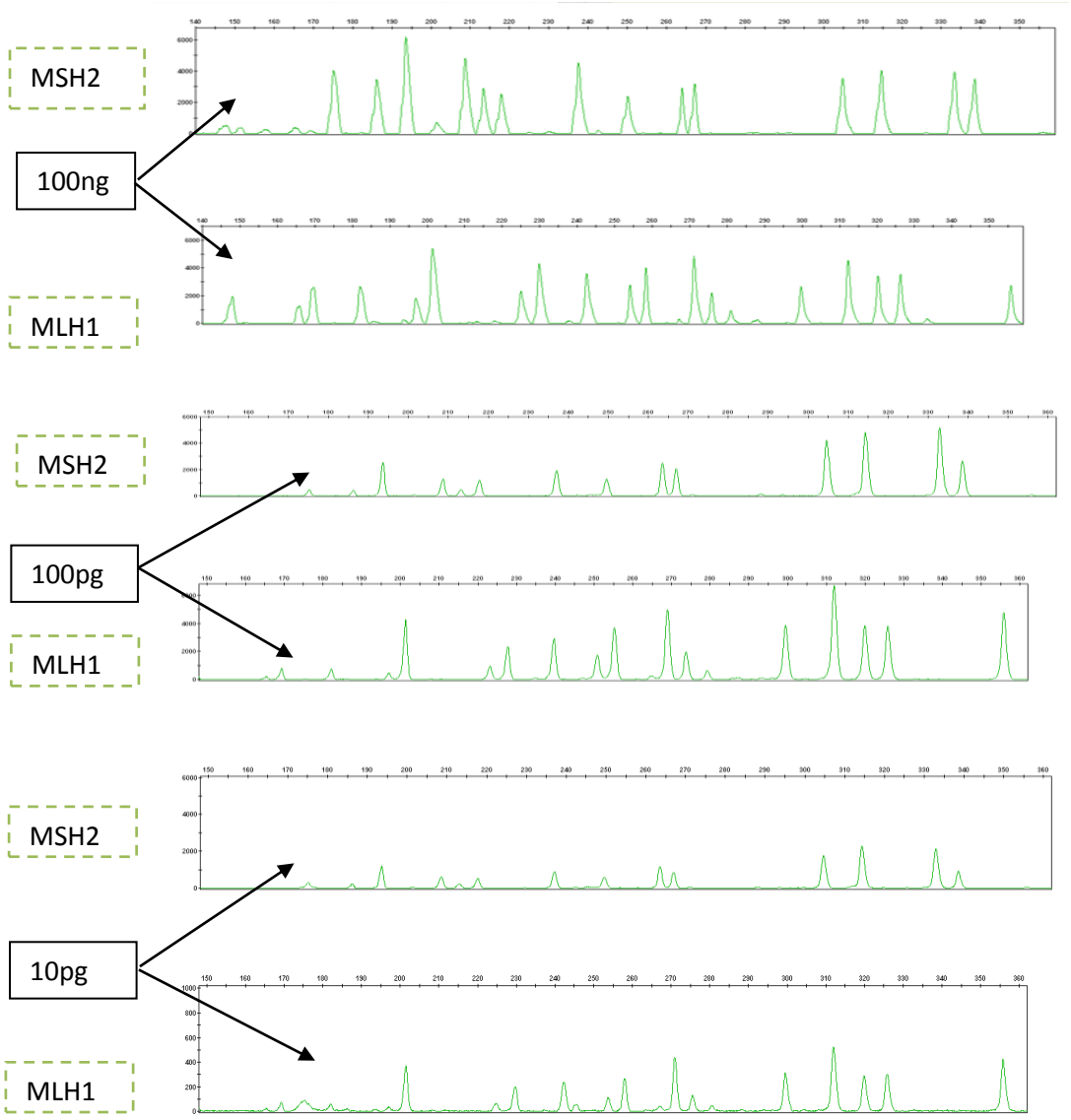


Figure 24. A comparison of the captured probes at various dilutions. At lower DNA concentrations it can be observed that smaller size probes have very weak signal and also in overall the signal is getting weaker as the input DNA concentration is reduced.

Quantity	Target DNA	Average	Control Average(Untreated genomic DNA)	Average of Enriched genomic DNA/ Control Average(Untreated genomic DNA)	Enrichment
SECOND	**ROUND**	**OF**	**ENRICHMENT**		
1ng	MSH2	72774.07	3516.467	20.75	20750
1ng	MSH2	77572.43	3516.467	22.05	22050
1ng	MLH1	42119.53	3516.467	11.97	11970
1ng	MLH1	47349.47	3516.467	13.46	13460
1ng	S11/S13	132848	3516.467	37.77	37770
1ng	S11/S13	197129	3516.467	56.05	56050
100pg	MSH2	36721.71	3516.467	10.44	104400
100pg	MSH2	34204.5	3516.467	9.72	97200
100pg	MLH1	43320.61	3516.467	12.31	123100
100pg	S11/S13	150182	3516.467	42.70	427000
100pg	S11/S13	124481	3516.467	35.39	353900
10pg	MSH2	15170.43	3516.467	4.31	431000
10pg	MSH2	4000.643	3516.467	1.137	113700
10pg	MLH1	3987.351	3516.467	1.13	113000
10pg	MLH1	3813.563	3516.467	1.08	108000
10pg	S11/S13	84756	3516.467	24.10	2410000
10pg	S11/S13	24799	3516.467	7.05	705000
10pg	S11/S13	108892	3516.467	30.9	3090000
10pg	S11/S13	247792	3516.467	70.4	7040000

Table 9. The level of enrichment in second round of enrichment. The level of enrichment is increased as the concentration is reduced with maximum enrichment achieved of 7,040,000. The average fragment size was about 300bp which is equal to 1/10,000,000 of the genome. Therefore 7,040,000 fold enrichment should equate to approximately 70% of the total DNA captured.

3.3 Single- Locus PCR of exonic regions of MLH1 and MSH2 genes

3.3.1 Single Probe Amplification Assay

In order for the level of enrichment to be further quantified, primers were designed (See Table 7 from Materials and Methods) for specific exonic regions, and the captured genomic DNA was amplified in various PCR cycles. In every amplification, unenriched linkered genomic DNA was also used, for the efficiency of the PCR to be estimated as the genomic DNA input was known (20 ng/μl). Also, captured genomic DNA from first and second round of enrichment was amplified for the same number of cycles in order the level of enrichment to be compared. All the DNA samples were amplified at 20, 23, 25, 27, 30, 32, 35 and 40 PCR cycles. With the use of a 100bp ladder (New England Biolabs), the PCR product was quantified. It was estimated for unenriched linkered genomic DNA that about 100ng of PCR product were generated after 32 PCR cycles. The average size of the sheared DNA fragments was approximately 300bp; therefore each fragment is equal to 1/10,000,000 of the length of the genomic DNA. Therefore any

300bp target in 20ng of DNA is represented at a total of 2fg and in order for 100ng of that 300bp fragment to be generated; the product was increased by 50,000,000 times. It was calculated that in every cycle the PCR product is increased by 1.74 times as 1.74^{32} are approximately equal to $5x10^{7}$ times.

3.3.2 S13 primers

For the first round of enrichment DNA, it was found that 32 PCR cycles were required for the generation of about 15ng of product DNA. Assuming that in every PCR cycle the product was increased by 1.74 times, it was estimated that in 1µl of first round of enrichment DNA S13 exon was presented at approximately 0.3fg. On the other hand, second round of enrichment DNA required only 20 cycles for the generation of approximately 10ng of DNA. Therefore, it was calculated as can be seen in Table 10 that in 1µl of second round of enrichment captured DNA, the S13 exon was represented at approximately 0.155pg.

3.3.3 S12 primers

For the S12 primer pair it was estimated that in every PCR cycle the DNA input is increased by 1.6 times. Therefore it was found that 35 cycles were required in order 15ng of DNA to be generated for first round of enrichment DNA. It was calculated that in 1μl of first round of enrichment DNA, 1fg of exon S12 is present. On the other hand, as can be observed in Figure 25, for second round of enrichment DNA, only 23 PCR cycles were required for the generation of 10ng of DNA. Therefore as shown in Table 10 in 1μl of second round of enrichment DNA, S12 exon at approximately 0.2pg was present.

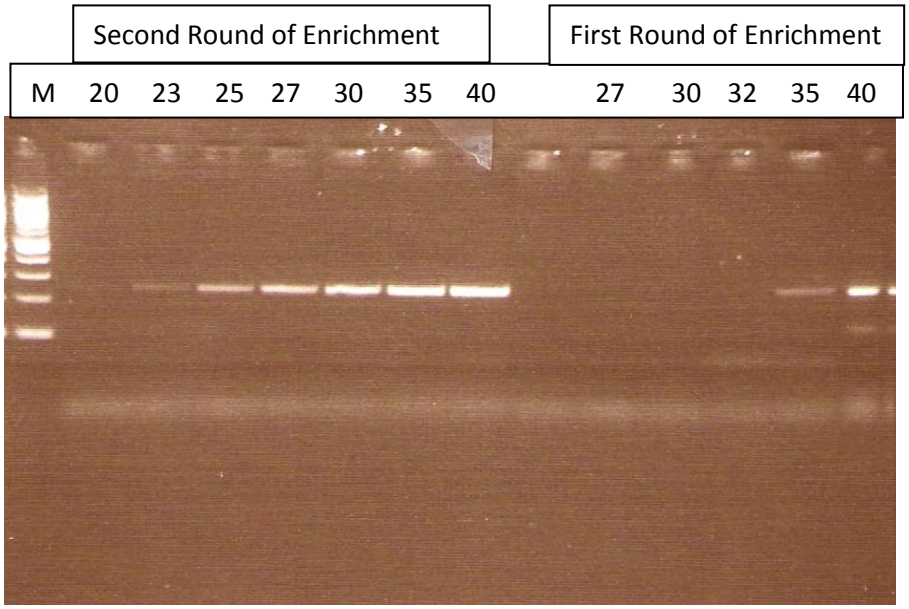


Figure 25. The PCR cycle test with the use of second round and first round of enrichment genomic DNA for amplification with S12 primers. The quantity of DNA was estimated with comparison with the 100bp marker (first lane).

3.3.4 L6 and L16 primers

The same experiments were repeated with the use of L6 and L16 primers and it was found for L6 that for the generation of 10ng of DNA 30 and 25 cycles were required for first round and second round of enrichment respectively this time using DNA enriched by hybridization to MLH1 DNA. Therefore as shown in Table 10 in 1μl of DNA input there are 1.4fg of L6 exon in the first round of enrichment, and 0.246pg for the second round of

enrichment. With L16 primers was estimated that in 1µl of DNA input there were 11fg of L16 exon for first round of enrichment DNA, and 0.25pg for second round of enrichment DNA.

3.3.5 Between S12 and S13

Primers were also designed for a region between S12 and S13 in order for any overlapping sequences captured to be identified. It was found that in 1µl of second round of enrichment DNA, approximately 2.15fg were of fragments from that region as shown in Table 10.

MSH2 probes	**AH1F/AH2F Primers**
First Round of Enrichment	5fg/μl
Second Round of Enrichment	3.2pg/μl
MSH2 probes	**S13 primers**
First Round of Enrichment	0.3fg/μl
Second Round of Enrichment	0.155pg/μl
MSH2 probes	**S12 primers**
First Round of Enrichment	1fg/μl
Second Round of Enrichment	0.2pg/μl
MSH2 probes	**S12/S13 primers**
First Round of Enrichment	-
Second Round of Enrichment	2.15fg/μl
MLH1 probes	**L6 primers**
First Round of Enrichment	1.4fg/μl
Second Round of Enrichment	0.246pg/μl
MLH1 probes	**L16 primers**
First Round of Enrichment	30fg/μl
Second Round of Enrichment	0.25pg/μl

Table 10. A summary with the results from the single probe amplification assays, together with (top lines) estimates of the total amount of DNA present from AH1/AH2 PCR. The enriched DNA was amplified and quantified in order for their representation to be evaluated. It can be seen that the different exons were represented in the captured DNA in similar amounts. The data represent the DNA quantity of the chosen enriched fragments previously described.

3.4 Nanodrop Spectrophotometry

Second round of enrichment genomic DNA for MSH2 probes which was previously amplified with AH1F/AH2F primers was quantified by gel electrophoresis. A known quantity of herring sperm DNA was used in order for the smears to be compared and the quantity of genomic DNA to be estimated. However, genomic DNA was in a form of smear and subsequently it was not possible to be measured with accuracy. Therefore Nanodrop spectrophotometry was used so a more accurate measurement of the total amount of amplified genomic DNA could be achieved. Second round of enrichment genomic DNA for MSH2 probes was amplified with AH1F/AH2F primers and then purified in order primers and nucleotides to be removed. From the purified sample, 1μl was used for Nanodrop spectrophotometry analysis (See Figure 26)

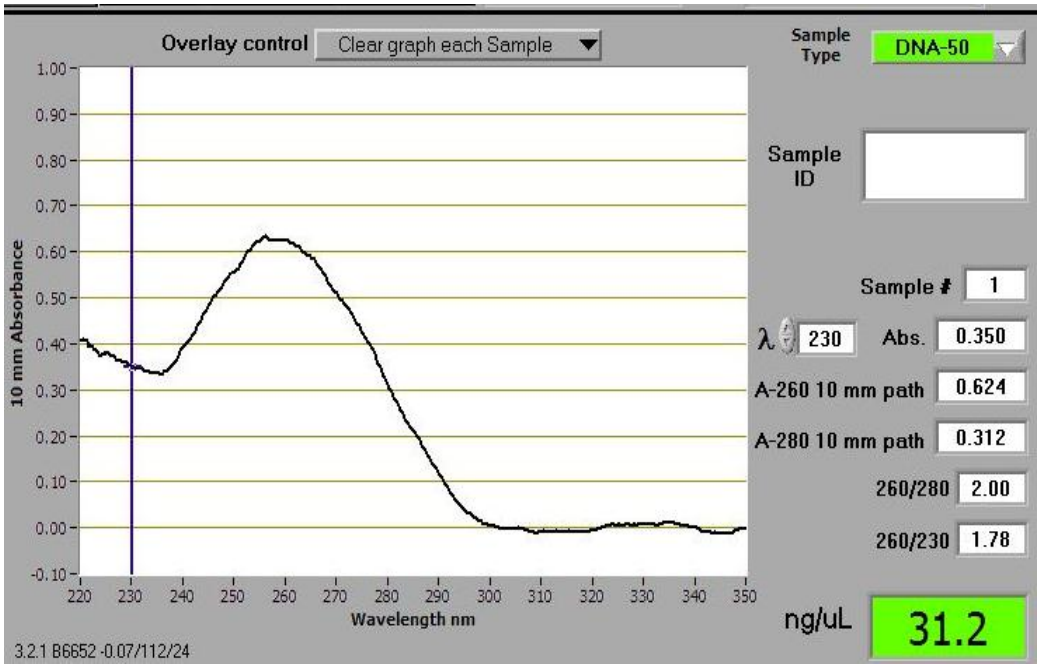


Figure 26. It was measured that in 1µl there are 31.2ng of captured DNA. The 260/280 ratio indicate the purity of the sample tested, and ratios from 1.8 to 2 are in generally considered as pure.

In total 4 DNA samples were amplified, purified and mixed together. Therefore in a mix of 30µl there are 935ng of DNA, which is approximately 234ng for each DNA sample. For the generation of 234ng of DNA 20 cycles were required from a 1µl input. It was estimated that in every PCR cycle the product was increased by 1.75 times and therefore the genomic DNA captured from the second round of enrichment in 1µl was calculated to be 3.2pg. Genomic DNA of known size was used in order for the value of 1.75 to be estimated; therefore the

samples of the same PCR experiment are expected to have approximately the same rate as the same reagents were used. Generating 234ng from 3.2pg input requires amplification of approximately 73,000 times, and 1.75^{20} is about 73,000.

3.5 Level of enrichment

The average quantity of S12 and S13 probes in 1µl of second round of enrichment DNA for MSH2 was estimated as 0.1775pg (S12 0.2pg, S13 0.155pg). There were 16 probes in total for MSH2, therefore assuming equal representation, the total amount of these probes in second round of enrichment genomic DNA for MSH2 was estimated as (16x 0.1775= 2.84pg). The total amount of MSH2 second round of enrichment DNA per 1µl was estimated to be 3.2pg, both by Nanodrop spectrophotometry and by AH1/AH2 PCR and therefore the successful on-target sequences in this enriched genomic DNA are approximately 89% of the total captured

sequences. If all the probes were enriched equally, each probe should be about 5% of the total on target 89%.

3.6 Sequencing

Second round of enrichment genomic DNA for MSH2 probes, was cloned into pGEM-T and 191 clones were sequenced for a more direct and accurate measurement of the level of enrichment. From the 191 clones, 163 of them were successfully sequenced, and 145 of them (89%) were on target sequences. In Table 11 and Figure 28 is illustrated that not all the probes were equally represented. The smaller size probes were underrepresented, and some probes (such as S8 and S16) were absent; on the other hand the larger probes were better enriched with S3 and S4 representing 22% and 18% respectively of all sequences. As can be seen in Figure 27 where there is an account of the captured probes and the region of their coverage, the fragments are overlapping with each other covering the total sequence. Also, S3 probe was investigated for the presence of any possible SNPs. It was

found that the presence of mismatches to the reference sequence was due to Taq polymerase low fidelity because mismatches were different in overlapping sequences, with no repetition. From the non-target sequences there were 5 from chromosome 19 at the NFIC gene, 5 unidentified non-human sequences, 1 from chromosomes 1, 7, 4 and 11 at ARHGEF12 gene, 3 sequences between S1 and S2 probe and 1 between S15 and S16 probe. Also the sequences captured were estimated to have an average size of 156bp, of which it was measured that 132bp were of the sequences of the target. Larger probes as mentioned above were better enriched with many sequences captured; however, each captured sequence was covering on average only 40% of the sequence of interest. On the other hand, for smaller probes fewer fragments were captured but because of their smaller size and the high specificity of the hybridization, the captured sequences were covering (on average) 80% for S10 and approximately 70% for S6 of the sequence of interest. Also, as can be seen in Figure 29, GC content was not the most important factor that affects the success of the enrichment.

Probe name	Probe size	Number of sequences captured	Percentage of Enrichment
S1	338	4	2.8%
S3	333	32	22%
S14	315	26	18%
S12	306	19	13%
S13	267	8	5.5%
S7	263	12	8.3%
S10	249	8	5.5%
S15	237	14	9.7%
S4	218	6	4.1%
S2	214	4	2.8%
S5	210	3	2%
S6	195	7	4.8%
S9	187	1	0.7%
S8	176	0	0%
S11	158	1	0.7%
S16	152	0	0%

Table 11. The list of the probes captured with their size in base pairs and the number of sequences captured for each probe

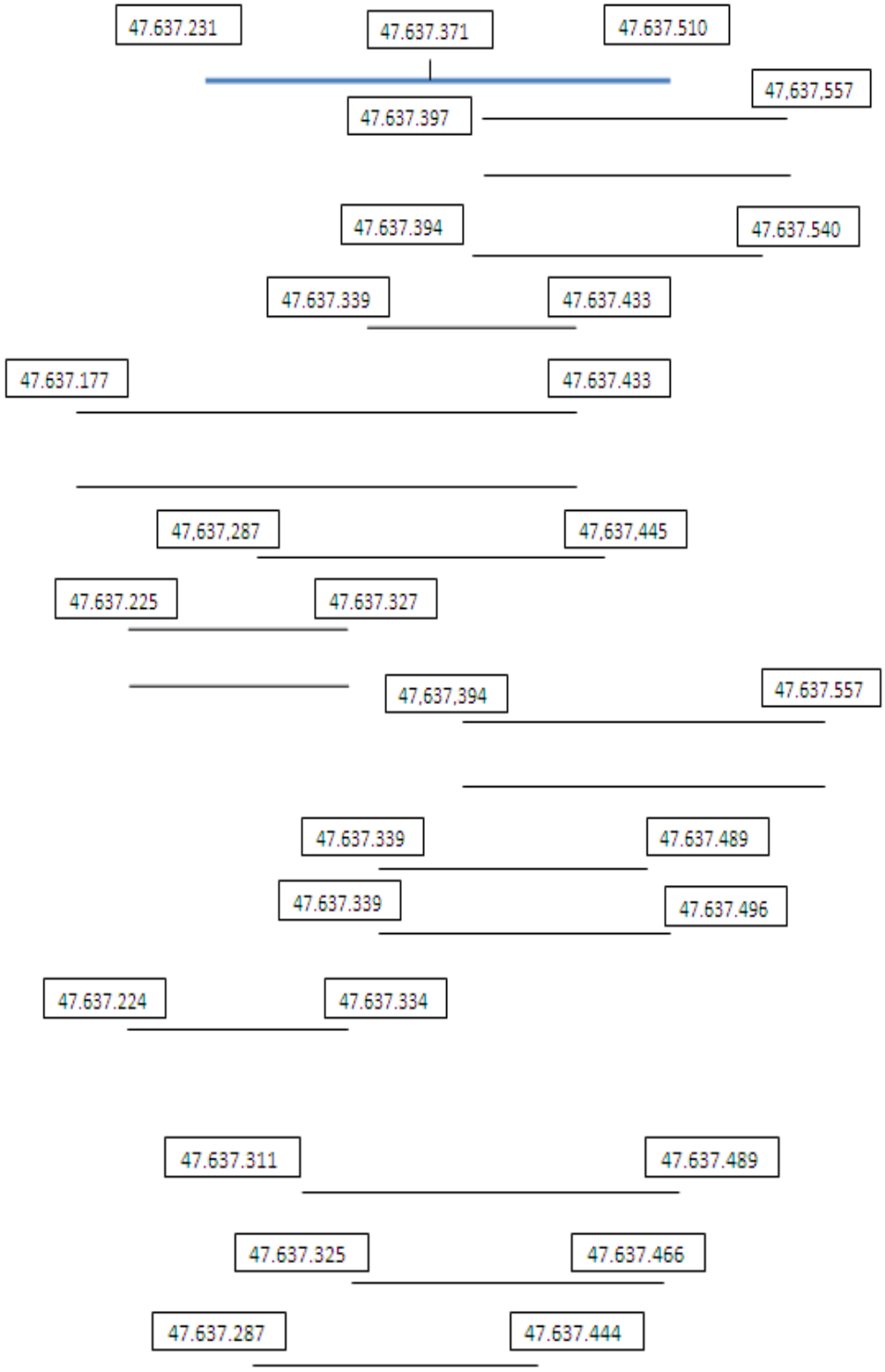


Figure.27 The captured fragments and the region of coverage for S3 probe (279bp). The top blue line indicates the S3 probe and the

black lines the captured fragments. The values on the lines indicate the position in base pairs of the fragments within chromosome 2.

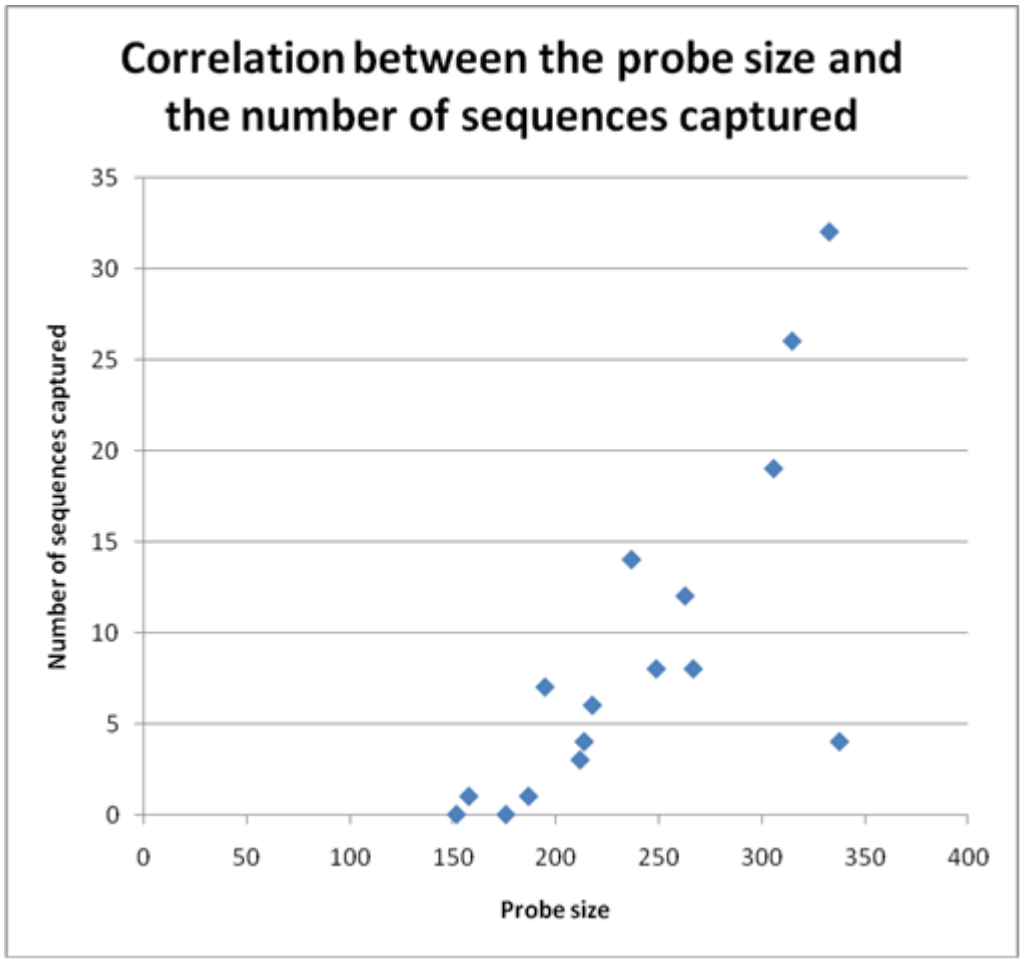


Figure 28 The correlation between the probe size and the number of sequences captured (P = 0.000914). The number of sequences captured is increased as the probe size that was used for selection is larger. However, the only exception is S1 (338bases) probe which was not enriched as well in correlation to its size.

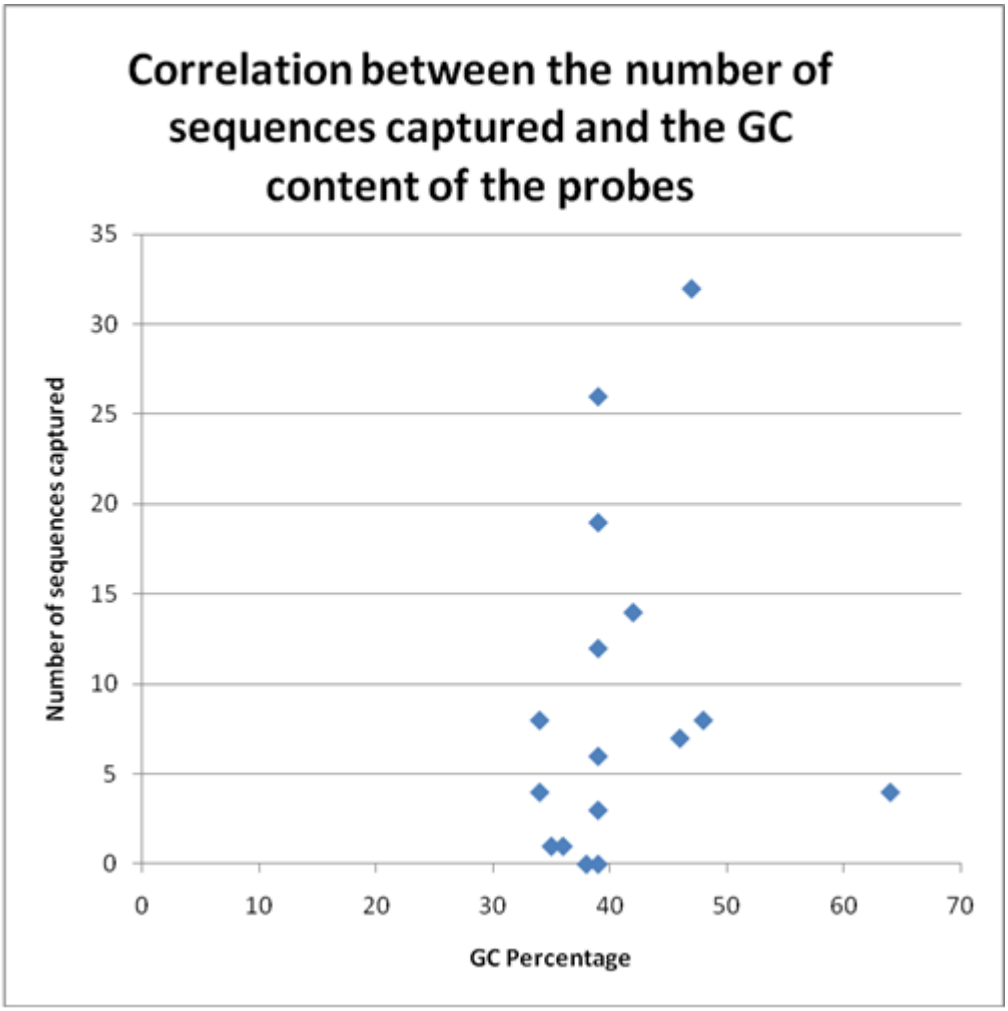


Figure 29. The correlation between the number of sequences captured and the GC content of the probes. The number of sequences captured has no correlation with GC content (P = 0.6151).

Chapter 4

Discussion

Copy number variations (CNVs) have been demonstrated that might be responsible for a wide range of genetic diseases (Shaikh *et al.* 2009). The identification of potential copy number variations in genes was a challenge in previous years. So far there is an extended record of CNVs proved to have role in genetic disease susceptibility. The challenge in this research was to establish a solid system for capturing regions in order to be sequenced and analysed for sequence variants as well as the existence of possible CNVs. This could become a future clinical diagnostic tool as CNVs can indicate the genetic disease in cases where symptoms are not clear.

In this project filter-based hybridization was used to enrich the sequences of interest (Herman *et al*. 2009) and MAPH for the evaluation of the enriched sequences. In previous studies MAPH was used for the capture of the probes of interest and investigation of potential deletions or duplications in genomic DNA (Armour *et al*. 2000). The aim of the project was for a

number of subgenomic targets to be efficiently enriched and evaluated for their level of enrichment in order that the enriched DNA could be used efficiently for next-generation sequencing analysis. The results confirmed that filter-based hybridization with two rounds of enrichment can be a powerful tool for enrichment with high specificity and high level. So far in previous research, the representation achieved by filter- based hybridization of a single round of enrichment and was from 60% to approximately 70% (Herman *et al*. 2009). On the other hand, in this project the representation which was achieved with the two round of enrichment for the capture of exons of MSH2 gene was approximately 90%. The level of enrichment after the use of only a single round of enrichment was estimated to be 600 times less than the enrichment after the second round. Therefore by comparison with the data from the MAPH analysis it was estimated that the enrichment after a single round was going to be approximately 1000 times, which indicates that about 0.15% of the total sequences, would contain sequences of interest. This is further evidence of how the second round of enrichment increased the specificity and the level of enrichment.

4.1 Generation of Amplifiable linkered genomic DNA

The genomic DNA with which the target sequences of interest were hybridized was sonicated. Therefore it was randomly sheared generating fragments from 100 to 900bp. Sonication offered the advantage of random shear at random positions, and not a method such as restriction enzymes which cut at defined positions generating fragments of known size. Also, the random size of the fragments offers the advantage in target sequences to have equal possibilities to be captured as many fragments with different size will contain the sequences of interest. In addition to that, because of the variety in the size of the captured genomic DNA many fragments will contain overlapping sequences with each other, therefore after sequencing the overlapping sequences were an additional tool for the accurate recording of the sequence of interest and overcoming the limitation of the presence of mutations induced during PCR.

4.2 Amplification of Genomic DNA

Series of PCR tests followed by MAPH were made in order for the best conditions for the amplification of genomic DNA to be identified. The aim was to find a PCR cycle number where all loci amplified almost equally, but the PCR cycle number at the same time was relatively small in order for mismatches which are incorporated during PCR to be reduced. The appropriate PCR cycle number was 20 where all the fragments seemed to where amplified similarly with a relatively low standard deviation.

4.3 MAPH

With the MAPH method and through ABI electrophoresis the representation of the captured probes was evaluated and the level of enrichment was estimated. However, for both first round and second round of enrichment captured genomic DNA, a series of dilutions were prepared and less DNA was loaded onto MAPH filters. The reduction of captured genomic DNA had

as a result the decrease in the number of sequences on the filters however the small amount of target DNA usually does not deplete the MAPH probes and therefore there was less competition among the background DNA and MAPH probes for the hybridization with the 2ng of probe set DNA. In MAPH after the reduction of amount of captured DNA on filters, because of the more successful hybridization, there was relatively stronger signal, which can be seen in Tables 8 and 9 with 5 times stronger signal than the filters with 1µg of captured genomic DNA shown in Figures 22 and 24.

4.4 ABI analysis

Captured probes from MAPH were analysed by ABI electrophoresis for the evaluation of their representation. The untreated genomic DNA samples which were used as controls were well captured and the probes equally represented which indicated that MAPH worked efficiently. However the probes from the first round and second round of enrichment were not equally represented, as can be seen in Figure 22. Larger

probes were better enriched than smaller probes, with the smallest having the weakest signal. This presumably happened because genomic DNA was fragmented randomly therefore fragments hybridized in random positions with the smaller probes and because of the small size of the probes of interest only few of the random DNA fragments were properly hybridized and not washed off after the washing step. The sequencing results support that theory, because on average each captured fragment was estimated to cover at least 60% of the sequence of interest. However, individual cases were found where the captured fragment was covering 12 bases from the sequence of interest.

4.5 Single Probe Amplification Assay

Region-specific primers were designed and series of PCR cycle tests were made in order the level of enrichment to be quantified independently of MAPH. Initially, first and second round of enrichment genomic DNA for MSH2 was amplified with the AH1F/AH2F primers in order that the total captured

sequences could be estimated. The quantification and comparison of the captured DNA between first and second round of enrichment revealed the efficiency and the importance of the use of second round enrichment hybridization. First-round enriched genomic DNA was estimated to be 5fg/µl on the other hand second round enriched DNA was found to be 3.2pg/µl, which is approximately 640 times higher concentration. The same experimental DNA was followed by using the region-specific primers. As can be seen in Table 10, there is a significant difference between the quantity of first and second round of enrichment with the concentration of specific target in second round of enriched DNA to be approximately 280 times higher than the first round.

Second round of enrichment genomic DNA was enriched with MSH2 filters, and therefore 3.2pg/µl was the quantity of captured MSH2 probes. The average of S13 and S12 probes was estimated as 0.1775pg/µl; and therefore the total amount of the 16 MSH2 probes is equal with the 89% of the total amount of the enriched genomic DNA.

4.6 Cloning and Sequencing

The previous percentage was estimated considering that all the probes were enriched equally. Therefore a cloning and sequencing experiment was undertaken for a more accurate measurement of the level of enrichment to be made, and to investigate whether all the probes were enriched equally or not. Genomic DNA from second round of enrichment with MSH2 filters was ligated into pGEM-T vector and transformed into TOP 10 cells. Once recombinant colonies were grown and selected, they were amplified with the use of PGB and 1277 primers. As it can be seen in Figure 18 (Material and Methods chapter), primer 1277 is 161 bases before the inserted sequence and PGB after the inserted sequence. These primers were selected in order to ensure that full insert sequences, including the very start of each insert, were obtained.

The results from sequencing indicated that approximately 89% of the total sequences corresponded to the MSH2 sequences of interest. As can be seen in Table 11, the probes were not enriched equally, but larger size probes were better enriched than smaller probes. This result is supported also with the ABI

peaks from MAPH analysis in Figure 22 and 24, where smaller probes were underrepresented. The fragment sizes which were captured varied from 120 to 180 bp therefore more fragments were enriched for larger probes but each fragment was covering approximately 47% (S3) of the total sequence. On the other hand fewer fragments were enriched for smaller probes but each fragment was covering approximately 70% (S6) of the total sequence. Subsequently despite the fact that not all probes were equally enriched, the approximately similar size of the captured fragments resulted in covering the total sequence of interest even for the smaller probes where less fragments were enriched. The drawback of low enrichment of small size probes can be surpassed by designing targets of a certain size preferably over 200bp in regions where there are not any dispersed repeats, which will contain the sequences that are needed to be analysed. In addition to that, it was found that the mismatches which were present were due to low fidelity of Taq polymerase and not of the presence of SNPs, therefore the use of a high fidelity polymerase and fewer PCR cycles could surpass this limitation.

4.7 Conclusion

In this project 35 exons from MSH2 and MLH1 genes were successfully enriched with approximately 90% representation for MSH2 probes equivalent to an overall enrichment of approximately 600,000 times. The method was quick with relatively low cost and proved to be a powerful tool for enriching subgenomic targets. The high specificity of the technique with the enriched fragments to cover successfully the sequences of interest suggests that it could be used as a future diagnostic tool where specific regions are required to be analysed by next generation sequencing. MLH1 probes in the MAPH and single locus PCR experiments appeared to be enriched similarly to MSH2 probes, and therefore in future experiments both MSH21 and MLH1 can be sequenced in order for an accurate comparison to be made. Also, in future research a higher number of samples can be sequenced so statistical error can be reduced, and estimation from more samples can be made. In addition to that, in this project 35 exons in total were enriched from MLH1 and MSH2 genes, and in further research a bigger fraction of the genome could be tried to be enriched by

enrichment of higher number of target sequences in a single experiment.

References

Akrami S. M. *et al*. (2005) Screening *for exonic copy number mutations at MSH2 and MLH1 by MAPH*, Familial Cancer, vol.**4** pp. 145-149

Ansorge W.J. (2009) *Next-generation DNA sequencing techniques*, New Biotechnology, vol.**25** pp.195-203

Armour J. A. L. *et al*. (2000), *Measurement of locus copy number by hybridization with amplifiable probes*, Nucleic Acids Research vol.**28** pp.605-609

Ashley E. A. *et al*. (2010) *Clinical assessment incorporating a personal genome*, Lancet, vol.**375**, pp.1525-1535

Bentley D. R. *et al*. (2008) *Accurate whole human genome sequencing using reversible terminator chemistry*, Nature, vol. **6** pp. 53-59

Biesecker L. G. (2010) *Exome sequencing makes medical genomics a reality*, Nature Genetics, vol. **42** pp. 13-14

Bunyan D.J. *et al*. (2004) *Dosage analysis of cancer predisposition genes by multiplex ligation- dependent probe amplification*, British Journal of Cancer, vol.**91**, pp.1155-1159

Cassili F. *et al*. (2002) *Rapid detection of novel BRCA1 rearrangements in high- risk breast- ovarian cancer families using multiplex PCR of short fluorescent fragments*, Human Mutation, vol.**20**, pp.218-226

Charbonnier F. *et al*. (2000) *Detection of exon deletion and duplication of the mismatch repair genes in hereditary nonpolyposis colorectal cancer families using multiplex polymerase chain reaction of short fluorescent fragments*, Cancer Research, vol.**60** pp.2760-2763

Chibon F. *et al*. (2008) *Contribution of PTEN large rearrangements in Cowden disease: a MAPH screening approach*. Journal of Medical Genetics, vol.**45**, pp.657- 665

Choi M. *et al*. (2009) *Genetic diagnosis by whole exome capture and massively parallel DNA sequencing*, Proceeding of the National Academy of Life Sciences, vol.**106**, pp.19096-19101

De Lellis L. *et al*. (2007) *Analysis of extended genomic rearrangements in oncological research*, Annals of Oncology, vol.**18** pp.vi173- vi178

Den Dunnen J. T. and White S. J. (2006) *MLPA and MAPH: Sensitive detection of deletions and duplications*, Current Protocols in Human Genetics, vol.**7** pp.7.14.1- 7.14.20

Erlich H. A. (1989) *Polymerase chain reaction,* Journal of Clinical Immunology, vol.**9** pp.437-447

Feuk L. *et al*. (2006) *Structural variation in the human genome*. Nature Reviews/Genetics, vol.**7** pp.85-97

Freeman J. L. *et al*. (2006) *Copy number variation: New insights in genome diversity,* Genome research, vol.**16** pp. 949-961

Ginsburg G. S. and McCarthy J.J. (2001) *Personalized medicine: revolutionizing drug discovery and patient care,* Trends in Biotechnology, vol.**19** pp. 491-496

Gupta P. K. (2008) *Single- molecule DNA sequencing technologies for future genomics research*, Trends in Biotechnology, vol.**26**, pp.602-611

Guttmacher A. E. (2010) *Personalized genomic information: preparing for the future of genetic medicine*, Nature Reviews/Genetics, vol.**11** pp.161-165

Herman D.S. *et al*. (2008) *Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection*, Nature Methods, vol.**6** pp.507-510

Hedges D.J. *et al*. (2009) *Exome sequencing of a multigenerational human pedigree, PLoS ONE, vol.4 pp.e8232*

Hollox E. J. *et al*. (2002) *DNA copy number analysis by MAPH: molecular diagnostic application*, Expert Review of Molecular Diagnostics, vol.**4** pp. 89-97

Hodges E. *et al.* (2007) *Genome- wide in situ exon capture for selective resequencing* , Nature Genetics, vol.**39**, pp.1522-1527

Johan T. *et al.* (2006) *MLPA and MAPH: Sensitive detection of deletions and duplications*, Current protocols in human genetics, Chapter **7**pp.7.14.1-7.14.20

Kato M. *et al.* (2009) *Population- genetic nature of copy number variations in the human genome*, Human Molecular Genetics , vol.**19**, pp.761-773

Kohlmann W, Gruber SB (Updated November 29, 2006). *Hereditary Non-Polyposis Colon Cancer*. In: GeneReviews at GeneTests: Medical Genetics Information Resource (database online). Copyright, University of Washington, Seattle. 1993-2009. Available at http://www.genetests.org. Accessed February 19, 2009.

Kousoulidou L. *et al.* (2008) *Array- MAPH: a methodology fro the detection of locus copy-number changes in complex genomes*, Nature Protocols, vol.**3** pp.849-865

Ley T. J. *et al.* (2008) *DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome*, Nature, vol.**456**, pp.66-72

Mardis E.R. (2008) *The impact of next- generation sequencing technology on genetics,* Trends in Genetics, vol.**24** pp.133-141

Mardis E. R. (2008) *Next- generation DNA sequencing methods*, Annual Review of Genomic and Human Genetics, vol.**9** pp.387-402

McCarroll S. A. (2008) *Integrated detection and population-genetic analysis of SNPs and copy number variation*, Nature Genetics, vol.**40**, pp.1166-1174

McGuire A. L*. et al.* (2008*) Research ethics and the challenge of whole-genome sequencing*, Nature Reviews Genetics, vol. **9**, pp.152-156

Ng S. B. *et al.* (2010) *Exome sequencing identifies the cause of a mendelian disorder*, Nature Genetics vol.**42** pp.30-36

Ng S. B. *et al.* (2009) *Targeted capture and massively parallel sequencing of 12 human exomes*, Nature, vol.**461** pp.272-278

Nyren, P. and Lundin, A. (1985) *Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis*. Analytical Biochemistry, vol.**151** pp.504-509

Patsalis P. C. *et al.* (2005) *MAPH: from gels to microarrays*, European journal of medical genetics, vol.**48** pp.241-249

Pleasance E.D. (2010) *A comprehensive catalogue of somatic mutations from a human cancer genome*, Nature, vol. **463**, pp.191-196

Primose S. B. and Twyman R.M. (2003) *Principles of Genome Analysis and Genomics*, 3rd edition, Blackwell edition

Ransohoff D. F. and Khoury M.J. (2009) *Personal genomics: information can be harmful*, European Journal of Clinical Investigation, vol.**40** pp.64-68

Redon R. *et al.* (2006)*, Global variation in copy number in the human genome*, Nature vol.**444**, pp. 444-454

Peter Russels (2009). *iGENETICS*. San Francisco: Pearson Education. 848.

Sanger F. *et al.* (1977) *DNA sequencing with chain-terminating inhibitors*, Proceedings of the national academy of sciences of the United States of America, vol.**74**, pp.5463-5467

Sellner L. N. and Taylor G. R. (2004) *MLPA and MAPH: New techniques for detection of gene deletions*, Human Mutation, vol.**23** pp. 413-419

Shaikh T. H. *et al.* (2009) *High- resolution mapping and analysis of copy number variations in the human genome: A data resourse for clinical and research applications*, Genome Research, vol.**19**, pp.1682-1690

Shendure J.and Ji H. (2008) *Next- generation DNA sequencing,* Nature Biotechnology, vol. **26** pp. 1135-1145

Snyder M. *et al.* (2010) *Personal genome sequencing: Current approaches and challenges,* Genes and development, vol.**24** pp. 423-431

Stevens V. L. *et al.* (2008) *Nicotinic Receptor Gene Variants Influence Susceptibility to Heavy Smoking, Cancer Epidemiology, Biomarkers and Prevention*, vol.**17** pp.3517-3525

"The Johns Hopkins Guide for Patients and Families: Hereditary Nonpolyposis Colorectal Cancer." The Johns Hopkins University (1995).

Tyson J. *et al.* (2009) *Quadruplex MAPH: Improvement of throughput in high-resolution copy number screening*, BioMed Central Genomics, vol.**10** pp. 1-9

Van Eijk R. *et al*. (2010) *MLPAinter for MLPA interpretation: an integrated approach for the analysis, visualisation and data*

management of Multiplex Ligation-dependent Probe Amplification, BioMed Central Bioinformatics, vol.**11** pp.1471-2105

Via M. *et al.* (2010) *The 1000 Genomes Project: new opportunities for research and social challenges,* Genome Medicine, vol.**2** pp.1-3

Vissers L. E. L. M. *et al*. (2003) *Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities*, American Journal of Human Genetics, vol.**73** pp.1261-1270

White S. *et al.* (2002*), comprehensive detection of genomic duplication and deletion in the DMD gene, by use of multiplex amplifiable probe hybridization*, American Journal of Human Genetics, vol.**71** pp. 365-374

XueMei W. and HuaSheng X. (2009) *Progress in the detection of human genome structural variations, Science in China Series C: Life Sciences*, vol.**52**, pp.560-567

Zhou X. G*. et al.* (2009) *The next- generation sequencing technology: A technology review and future perspective*, Science China Life Sciences, vol.**53**, pp.44-57