

OPTIMIZATION OF SAMPLE PREPARATION FOR NEXT- GENERATION SEQUENCING WITH ILLUMINA GENOME ANALYZER II

M. Järvinen, July 2010

Maija Järvinen

Master's thesis

University of Helsinki

HEBIOT

Biotechnology

July 2010

Tiedekunta/Osasto — Fakultet/Sektion — Faculty Bio- ja ympäristötieteiden laitos		Laitos — Institution — Department HEBIOT	
Tekijä — Författare — Author Maija Järvinen			
Työn nimi — Arbetets titel — Title Näytteiden esikäsittelyn optimointi uuden polven sekvensointiin Illuminan Genome Analyzer II:lla			
Oppiaine — Läroämne — Subject Bioteknologia			
Työn laji — Arbetets art — Level Pro Gradu		Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Tiivistelmä — Referat — Abstract <p>Viimeisen vuosikymmenen aikana sekvensointisovellukset ovat kehittyneet tehokkaammiksi ja vastaamaan paremmin alati kasvavaa kysyntää. Tämä tutkielma keskittyy DNA näytteiden esikäsittelyn optimointiin Illuminan Genome analyzer II – sekvensointiin. Näytteiden esikäsittelyvaiheista optimoitiin fragmentaatio, PCR-reaktion jälkeinen puhdistus sekä kvantitointi. Ihmis- ja bakteeri-DNA:ta fragmentoitiin käyttämällä kohdistettua ääniaaltoa. Fragmentointi testattiin eri aikapisteiden ja näytemäärien suhteen. Puhdistusmenetelmistä verrattiin kahta eri pylväspuhdistusmenetelmää, geelipylvästä resiinillä ja ilman sekä magneettihelmiin perustuvaa puhdistusta. Kvantitatiivista PCR:ää ja geelielektroforeesia sirulla verrattiin DNA-määrän mittaamiseen.</p> <p>Puhdistusmenetelmistä magneettihelmpuhdistus toimi tehokkaimmin ja on parhaiten muokattavissa. Fragmentointi optimoitiin isommille fragmenteille ja se on joustavammin muokattavissa. Kvantitointimenetelmistä kvantitatiivinen PCR korreloi parhaiten syntyneiden klustereiden kanssa. Tämän tutkielman tuloksena sekvensointiajot tuottavat enemmän dataa edullisemmin. Lisäksi laadunvarmistuspisteet helpottavat vianmäärittystä. Uudet sekvensointilaitteet ja –sovellukset tulevat vaatimaan optimointia myös jatkossa.</p>			
Avainsanat — Nyckelord — Key words Uuden polven sekvensointi, näytteiden esikäsittely			
Säilytyspaikka — Fövaringsställe — Where deposited			
Muita tietoja — Övriga uppgifter — Further information Ohjaaja ja vastuuproessori: Professori: Professori Tapio Palva; Ohjaaja: Janna Saarela, LT; Pekka Ellonen, Ins.			

Tiedekunta/Osasto — Fakultet/Sektion — Faculty Faculty of Biological and Environmental Sciences		Laitos — Institution — Department HEBIOT	
Tekijä — Författare — Author Maija Järvinen			
Työn nimi — Arbetets titel — Title Optimization of sample preparation for next-generation sequencing with Illumina Genome Analyzer II			
Oppiaine — Läroämne — Subject Biotechnology			
Työn laji — Arbetets art — Level Master's Thesis		Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Tiivistelmä — Referat — Abstract <p>The growing interest for sequencing with higher throughput in the last decade has led to the development of new sequencing applications. This thesis concentrates on optimizing DNA library preparation for Illumina Genome Analyzer II sequencer. The library preparation steps that were optimized include fragmentation, PCR purification and quantification. DNA fragmentation was performed with focused sonication in different concentrations and durations. Two column based PCR purification method, gel matrix method and magnetic bead based method were compared. Quantitative PCR and gel electrophoresis in a chip were compared for DNA quantification.</p> <p>The magnetic bead purification was found to be the most efficient and flexible purification method. The fragmentation protocol was changed to produce longer fragments to be compatible with longer sequencing reads. Quantitative PCR correlates better with the cluster number and should thus be considered to be the default quantification method for sequencing. As a result of this study more data have been acquired from sequencing with lower costs and troubleshooting has become easier as qualification steps have been added to the protocol. New sequencing instruments and applications will create a demand for further optimizations in future.</p>			
Avainsanat — Nyckelord — Key words Next-generation sequencing, sample prep, sample preparation, library preparation			
Säilytyspaikka — Fövaringsställe — Where deposited			
Muita tietoja — Öriga uppgifter — Further information Supervisor and responsible professor: Professor: Professor Tapio Palva Supervisors: Janna Saarela, PhD, MD; Pekka Ellonen, BSc, Biotechnology			

Table of Contents

Table of Contents	i
Abbreviations	iii
Introduction	1
Sequencing with Genome Analyzer II	2
Sample preparation	2
Fragmentation.....	3
End repair	3
A-tailing	4
Ligation	4
Purification.....	4
Silica- membrane purification	5
Gel filtration with resin	5
Paramagnetic bead purification	5
Sequence capture	6
Array capture	6
In-solution capture	7
DNA quantification.....	8
Spectrophotometric method.....	8
Microchannel based electrophoresis.....	9
Real-time PCR.....	10
Cluster amplification.....	11
Sequencing	12
Data analysis	13
Aims of the study	14
Materials and methods	15
Fragmentation.....	15
The evaluation of the current protocols using Covaris fragmentation (Series A) ...	16
The effect of the fragmentation length to product size (Series B)	16
The effect of the concentration in fragmentation to product size (Series C)	16

Purification	16
The ability of purification methods to remove small fragments (Series D)	17
The efficiency of purification methods to remove primer dimers (Series E)	18
The ability of purification methods to remove excess amount of primer dimers from sequencing library (Series F)	18
Comparison of the ability of purification methods to enrich optimal fragment sizes (Series G)	19
Amplification.....	19
Quantification	20
Results	22
Fragmentation.....	22
The evaluation of the current protocols using Covaris fragmentation (Series A) ...	22
The effect of the fragmentation length to product size (Series B)	23
The effect of the concentration in fragmentation to product size (Series C)	24
Purification	26
The ability of purification methods to remove small fragments (Series D)	26
The efficiency of purification methods to remove primer dimers (Series E)	29
The ability of purification methods to remove excess amount of primer dimers from sequencing library (Series F)	31
Comparison of the ability of purification methods to enrich optimal fragment sizes (Series G)	34
Quantification	36
Amplification.....	37
The impact of the optimization to sequencing yields	39
Discussion	41
Acknowledgements	46
References	48
APPENDIX 1: Primer and adapter sequences	54
APPENDIX 2: Protocols	54
APPENDIX 3: Covaris Protocols	54
APPENDIX4: Thermal amplification protocol.....	55

Abbreviations

CCD	Electron multiplying charge-coupled device
PCR	Polymerase chain reaction
dAMP	Deoxyadenosine monophosphate
dATP	Deoxyadenosine triphosphate
dTMP	Deoxythymidine monophosphate
dNTP	Deoxyribonucleotide triphosphate
SPRI	Solid Phase Reversible Immobilization
UV	Ultraviolet
RTA	Real time analysis
SNP	Single nucleotide polymorphism
CV	Coefficient of variation
DTR	Dye terminator removal
BWA	Burrows-Wheeler Aligner
PF	Purity filtered
<i>E. Coli</i>	<i>Escherichia coli</i>

Introduction

The Sanger method (Sanger, Nicklen & Coulson 1977) has been the golden standard for DNA sequencing for the last few decades. It has played a major role in understanding genomic sequence in human (Lander et al 2001, Venter et al 2001) and other species such as *Escherichia coli* (Blattner et al. 1997) and hence has substantially contributed to the studies of medicine and biotechnology. The growing interest for sequencing with higher throughput and lower costs in the last decade has led to development of new sequencing applications. The currently leading platforms are Genome Analyzer (Illumina Inc., San Diego, CA, USA), Solid (Applied Biosystems, Foster City, CA, USA) and 454 (454 Life Sciences, a Roche company, Branford, CT, USA). Each platform is based on specific technological innovations (Margulies et al. 2005, Bentley et al. 2008, Fedurco et al. 2006, Shendure et al. 2005). On the other hand all of them rely on high quality nucleic acid libraries.

The main weakness of current library preparation protocols is that they have not been optimized for the variation between samples. The amplification, for example, is often performed blindfolded without a prior knowledge of the sample concentration. This creates fluctuation between samples and complicates possible troubleshooting. The presence of primers dimers hinders the accuracy of quantification and should thus be avoided. The purification methods used in current protocols are not efficient enough for this and other methods should be looked into. The quantification for sequencing should be repeatable and precise in order to maximize the data yield. At the moment there are no guidelines for quantification method from the manufacturers thus creating the need for optimization. (Appendix 2)

This Master's thesis concentrates on optimizing DNA library preparation for Illumina Genome Analyzer II sequencer with the amount of sample sufficient for sequence capture. However, the sample library steps described here are partially adaptable also for 454 and Solid.

Sequencing with Genome Analyzer II

The genome Analyzer II workflow consists of sample preparation, cluster amplification, sequencing and data analysis.

Sample preparation

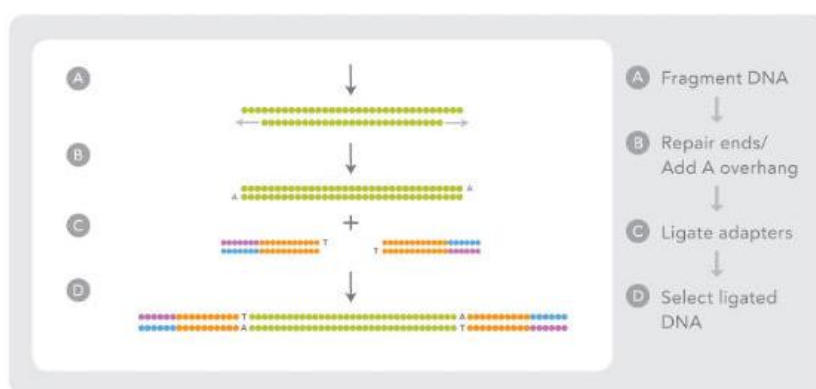


Figure 1 Sample preparation workflow. In sample preparation DNA is fragmented, end repaired, A-tailed, ligated and amplified. (Adapted from “Go where biology takes you” brochure, <http://www.illumina.com>, Publication Number 770-2009-034)

Genome Analyzer II can utilize nucleic acids from various resources if it is provided in a compatible format (de Magalhães, Finch & Janssens 2010, Schmidt et al. 2009, Linnarsson 2010). The library consists of short fragments of double-stranded DNA. These fragments should be approximately 150 to 500 base pairs long to have uniform performance in cluster amplification. The fragments have adapters on both ends of the fragment complementary to both adapters on flow cell and sequencing primers (Linnarsson 2010, Pettersson, Lundeberg & Ahmadian 2009, Mardis 2008).

The primary step in standard library preparation protocol is fragmentation in which the genomic DNA is sheared to short fragments of double-stranded DNA. The fragments that have single-stranded overhangs at this point are then end repaired to generate blunt-ended fragments. The addition of deoxyadenosine monophosphate (dAMP) to the 3'-end

of the strands enhances the efficiency of the correct ligation (Linnarsson 2010). The polymerase chain reaction (PCR) amplification is also needed in order to enrich the fragments that have been successfully prepared and thus have the adapters in both ends and are capable of amplifying and attaching to a flow cell. (Figure 1)

Fragmentation

The fragmentation can be performed either by mechanical or sonar force or by enzymatic reaction. The official Illumina sample preparation protocol uses pressurized air based method called nebulization to fragment the DNA. Other means for fragmentation are focused sonication (Covaris Inc., Woburn, MA, USA) and enzymatic fragmentation with Fragmentase (New England Biolabs, Ipswich, MA, USA). In enzymatic fragmentation random nicks are created by nuclease and the nicks are recognized by endonuclease which cleaves the other strand (New England Biolabs 2010). Also DNaseI in the presence of Mn^{2+} fragments the DNA (Linnarsson 2010).

Sample yields from focused sonication can be more than four-fold compared to nebulization since the fragment size distribution is narrower. The enzymatic methods have wider size distribution than sonication but have better yields than nebulization (Linnarsson 2010).

End repair

After the fragmentation the fragments have 3' and 5' overhangs which must be cut off in order to have blunt-ended fragments which can be further prepared for adapter ligation. T4 DNA polymerase catalyzes the synthesis of DNA in the 5' -> 3' direction thus filling the 5' overhangs and cleaves the 3' overhangs with its 3' -> 5' exonuclease activity (Orkin 1990).

A-tailing

The aim of incorporating deoxyadenosine monophosphate (dAMP) to the 3' end of the blunt-ended DNA fragment is to prevent the formation of concatemers in ligation and to enhance the ligation efficiency. The adapters have complementary deoxythymidine monophosphate (dTMP) in the 3' end of the ligating end of the adapter which enables successful ligation of the adapter. The deoxyadenosine triphosphates (dATP) are incorporated by Klenow fragment which is an N-terminal truncation of DNA Polymerase I which retains polymerase activity, but has lost the 5' → 3' exonuclease activity (Clark, Joyce & Beardsley 1987).

Ligation

The ligation process incorporates adapters to both ends of the blunt-ended DNA fragments. Successful ligation should result in different adapters in different ends of the fragments. A dTMP on the 3' end of the ligating end of the adapter enables the ligation to the DNA fragment with 3' end dAMP overhang. The DNA strands of the adapters are complementary only on the ligating end of the adapter resulting in Y-shape. This significantly reduces the formation of fragments with the same adapter in both ends of the fragments (Linnarsson 2010). The phosphorothioate modifications in the ends of the adapters protect the adapters from endonucleases. The ligation is performed by T4 DNA ligase. T4 DNA ligase is an enzyme purified from *E. coli* C600 pC1857 pPLc28 lig8. It is capable of catalyzing the phosphodiester bond formation of juxtaposed 5' phosphate and 3' hydroxyl termini of DNA and RNA. It can incorporate both blunt-ended and cohesive-ended termini. (Engler, Richardson 1982)

Purification

After each sample preparation step the library must be purified from salts, enzymes, surplus of adapters and other reagents. There are several approaches for purification from

which silica-membrane-based method, gel-filtration with resin and paramagnetic bead purification are covered in this thesis.

Silica-membrane purification

DNA in a high-salt buffer binds to silica-gel-membrane. The additional substances are removed, and the DNA eluted in a low-salt buffer. This method is susceptible to pH changes. (Hamaguchi, Geiduschek 1962, Vogelstein, Gillespie 1979)

Gel filtration with resin

Gel filtration purification utilizes different method compared to the other purification approaches. As in other methods DNA is bound to beads or membranes, in gel filtration additional substances in the solution are separated from DNA during gel filtration. Primers, short single stranded DNA strands, salts, buffers, deoxyribonucleotide triphosphates (dNTP) and other small molecules are eluted in larger volume than large molecules such as DNA. (EdgeBio 2010)

Paramagnetic bead purification

Solid Phase Reversible Immobilization (SPRI) beads have polystyrene core coated with a layer of iron (Figure 2). The iron grants the beads their paramagnetic properties. The surface of the bead is a polymer layer containing carboxyl functional groups. The SPRI beads bind the negatively charged nucleic acid with the carboxyl groups leaving other substances in the solution. The beads carrying the nucleic acid can be separated from the solution by external magnetic field. This enables efficient washing and elution of the DNA or RNA in proper elution buffer. (Beckman Coulter Genomics 2010, Deangelis, Wang & Hawkins 1995)

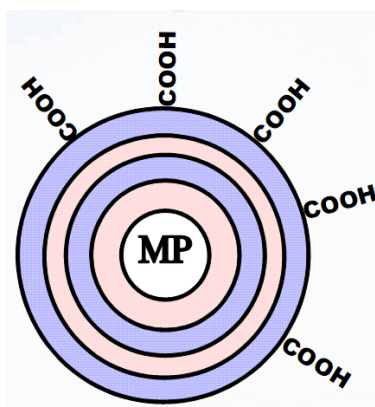


Figure 2 Agencourt AMPure XP SPRI bead. White illustrates the polystyrene core, pink the magnetite and purple the polymer layer and the outermost layer contains carboxyl groups. (Adapted from Agencourt AMPure SPRI bead brochure, <http://www.agencourt.com>)

Sequence capture

The next-generation sequencing platforms such as HiSeq 2000 (Illumina) are aiming at the whole genome sequencing in a single run of the instrument. However, the Genome Analyzer II does not yet have such capacity. This compels to the selection of regions of interest and their enrichment in order to keep the sequencing costs and resources in a feasible level. At present selected regions of the genome can be captured using microarray or a solution phase method. (Mamanova et al. 2010b) Agilent and Roche Nimblegen also provide capture probes for the whole human exome.

Array capture

In an array capture application the target specific probes are immobilized on a microarray surface. The sample library is hybridized with these probes, nonspecific DNA is washed away and the captured DNA is eluted. At present, array capacities varies from 244000 probes for smaller targets to arrays of 2,1 million probes for exomes. (Mamanova et al. 2010b, Agilent 2010, Roche NimbleGen 2010)

In-solution capture

In an in-solution capture the target regions are captured with biotin attached probes. Of the two present in-solution capture manufacturers Agilent uses 150-mer RNA probes and Nimblegen 60-90-mer DNA probes (Mamanova et al. 2010b, Gnirke et al. 2009). After the hybridization of the probes to the target DNA the captured DNA is separated from the non-captured DNA by binding the biotin of the probes to streptavidin coated paramagnetic beads. Using external magnetic field the beads and hence the targeted regions can be recovered. (Figure 3) (Mamanova et al. 2010b, Invitrogen 2010, Gnirke et al. 2009)

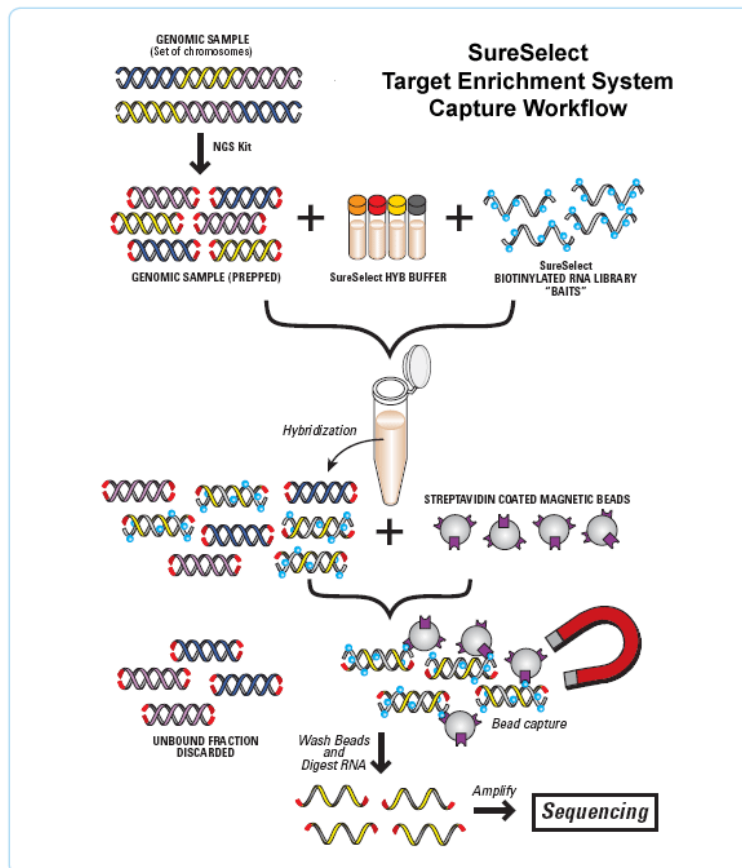


Figure 3 Agilent SureSelect Target Enrichment workflow illustrating sequence capture in solution. Sample library is hybridized with biotinylated probes and then captured with Streptavidin coated magnetic beads and magnet. (Adapted from Agilent SureSelect Target Enrichment protocol, www.chem.agilent.com, Publication Number G3360-90010).

DNA quantification

Accurate DNA quantification is crucial in sequencing with Illumina Genome Analyzer II. If concentration is too high clusters will overlap and the signals from these different clusters cannot be distinguished hence resulting in reduced data acquisition. The full capacity is not used if the concentration is too low, since the clusters will not be as dense as possible (Linnarsson 2010). The raw cluster number, given by the primary data analysis of a sequencing experiment, indicates the total cluster number. The purity filtered (PF) cluster number is an indicator of the number of clusters that are not overlapping (Illumina 2010).

Quantification of DNA is also crucial before sequence capture experiment. A target area might not be fully covered if sufficient amount of DNA is not available for capture probes. This would result in too few template molecules for amplification after capture and creates a risk of producing clonal molecules and a biased library. These clonal molecules reduce the data yield of the sequencing run since clonal molecules from different clusters are discarded in the data analysis.

Spectrophotometric method

Spectrophotometer measures the absorption of light at different wavelengths. DNA and RNA absorbs ultraviolet (UV) light with an absorption peak at 260 nanometers of wavelength as proteins have absorption peak at 280 nanometers of wavelength and phenol and polysaccharides at 230 nanometers of wavelength. Thus the ratios 260/280 and 260/230 can be used as an indicator of the purity of the sample. However, the 260/280 ratio has been noted to be insufficient as an indicator of sample purity since even significant protein amounts reduces the ratio only slightly (Linnarsson 2010, Orkin 1990, Gallagher SR, FAU - Desjardins & Desjardins PR 2008).

Spectrophotometer is a rapid method for DNA quantification and NanoDrop instrument from Thermo Scientific also reduces the volume needed for the analysis to 1 to 1,5 microlitres compared to traditional spectrophotometers with analysis volume of 1,5 to 4,5 millilitres in a cuvette. With NanoDrop the sample is pipetted directly to the pedestal hence eliminating the need for cuvettes. (NanoDrop 2010)

The additional primers, adapters and free nucleotides cannot be distinguished from the sample library and hence they increase the total concentration. Further since spectrophotometrically all nucleic acids are measured also the fragments of the sample library that are not able to amplify in flow cell, i.e. lack the adapters from either one or both ends, are measured. (Linnarsson 2010)

Microchannel based electrophoresis

Microchannel based electrophoresis platforms (Lab-on-a-chip) such as the 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA), are based on fluorescent labeling of the sample, electrophoretical separation of the DNA in microfluidic channels (Figure 4) and sample analysis and quantification based on internal standards. This method reduces the sample volume needed for the analysis and is less time consuming than conventional electrophoresis. The advantage of Bioanalyzer compared to other DNA quantification methods is the fragment size information which is important in next-generation sequencing. Also primers and adapters can be distinguished by their fragment size from the sample library while in the other methods these are included in the sample concentration. (Kuschel, Buhlmann & Preckel 2005) Bioanalyzer is still unable to distinguish the fragments of a sample library that can be sequenced from those that cannot (Linnarsson 2010). However, this is not a major issue as the library is amplified and the fragments able to be sequenced are enriched from those that are not.

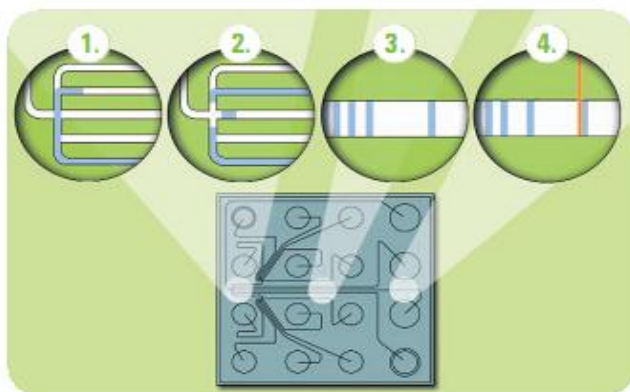


Figure 4 Illustration of microchannels on Agilent Bioanalyzer chip. 1. The sample moves through the microchannels from the sample well. 2. The sample is injected into the separation channel. 3. Sample components are electrophoretically separated. 4. Components are detected by their fluorescence and translated into gel-like images (bands) and electropherograms (peaks). (Adapted from Agilent Bioanalyzer brochure, www.chem.agilent.com, Publication Number 5989-7725EN).

Real-time PCR

In real-time qPCR the sample library is amplified with appropriate primers, in this case the primers complementary to ends of the library generated in the PCR amplification step. The amplification takes place in the presence of SYBR Green I fluorescent dye. SYBR Green binds specifically to double-stranded DNA and hence enables the quantification of DNA after each cycle of PCR by electron multiplying charge-coupled device (CCD). (Finnzymes 2010) Absolute DNA concentration can be achieved by comparing the samples with unknown concentration to a dilution series of a sample with a known concentration. The dilution series of a known sample will give a standard curve where initial sample concentration is plotted against cycle threshold. Plotting the cycle thresholds of samples being studied to the standard curve will give the initial concentration. (Finnzymes 2010, Sellars et al. 2007)

The benefit of an absolute quantification with real-time PCR compared to the other quantification methods is that it simulates the bridge-amplification in sequencing since the primers in real-time PCR are the same as the ones attached to flow cell surface in which the sample library is bound. Hence only the molecules in the library that can be

sequenced are counted. (Meyer et al. 2008) However, this includes the conjoined primers and hence these primer dimers are indistinguishable from the sample library.

Cluster amplification

Sequencing takes place in a microarray surface called flow cell (Fuller et al. 2009). Flow cells have 8 individual lanes which enables running 8 physically separated DNA libraries (Mardis 2008). The sample library is hybridized to oligonucleotide adapters on the flow cell surface and clonally amplified before sequencing.

Each DNA fragment that has the ability to be cluster amplified has adapters complementary to those in the flow cell surface in both ends of the fragment (Linnarsson 2010, Pettersson, Lundeberg & Ahmadian 2009, Ansorge 2009). Adapters within flow cell are covalently bound to the flow cell surface (Mardis 2008). DNA molecules of a sample library are hybridized to the probes on the flow cell from 3' end of the fragment and the 5' end of the fragment is attached to an adapter complementary to this end of the fragment creating a bridge-like structure. Fragments are being amplified with a method called bridge amplification to generate clusters containing identical copies of the original molecule (Pettersson, Lundeberg & Ahmadian 2009, Ansorge 2009). This step is needed in order to detect the signal with CCD camera (Fuller et al. 2009). One flow cell contains hundreds of millions of these clusters in random positions, each cluster containing approximately 1000 copies of one clonal DNA fragment (Pettersson, Lundeberg & Ahmadian 2009). After cluster generation the flow cell contains heterogeneous population of clusters as each cluster contains clonal copies of one fragment (Illumina 2010). (Figure 5)

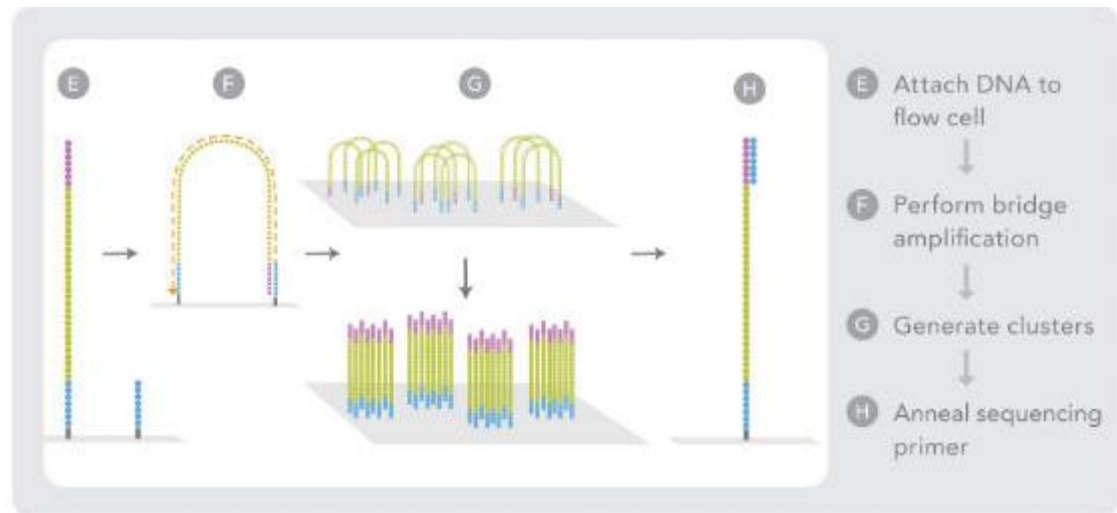


Figure 5 Cluster amplification workflow. DNA is hybridized to the oligonucleotides on the surface of the flow cell. Single molecules are then amplified to generate clusters. Each cluster consists of identical copies of the original template. (Adapted from “Go where biology takes you” brochure, <http://www.illumina.com>, Publication Number 770-2009-034)

Sequencing

Illumina Genome Analyzer II sequencing is based on sequencing-by-synthesis method in which fluorescent labelled, reversible terminated nucleotides are incorporated in a cyclic manner (Ansorge 2009).

After the cluster amplification the fragments are linearized and sequencing primers are added (Illumina 2010). Sequencing is performed in cyclic manner. Each cycle consist of addition of polymerase and fluorescently labelled nucleotides with chemically inactivated 3'-OH (Mardis 2008), incorporation of the reversibly terminated nucleotide, washing of excess reagents, imaging of the flow cell to determine the incorporated nucleotide and finally removing fluorescent dye and blocking group from the 3'-end of the base (Mardis 2008, Ansorge 2009). These cycles are repeated up to 100 times. (Figure 6) Library molecules can also be sequenced from the other end of the molecule which results in 2 times more data. The paired reads facilitate the alignment of the short reads to genome and enable more specific mapping of fragments to the reference. These paired-end reads also enable detection of structural rearrangements (Tucker, Marra &

Friedman 2009). Over 20 gigabases of sequence can be generated with Genome Analyzer II in a paired-end sequencing run with 100 cycles (Illumina 2010).

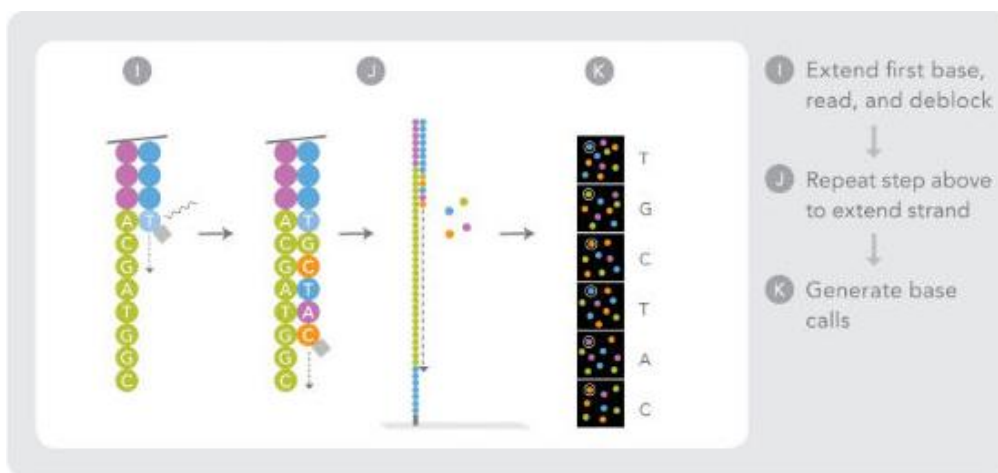


Figure 6 Sequencing-by-synthesis workflow. After the addition of reversible terminated base the flow cell surface is washed and imaged. The fluorescent dye and blocking agent is then removed and next base is added. These cycles can be repeated up to 100 times. (Adapted from “Go where biology takes you” brochure, <http://www.illumina.com>, Publication Number 770-2009-034)

Data analysis

The raw data from a sequencing run is images illustrating clusters after each cycle in each tile of the lane. Four images are taken per cycle for each tile, one for each nucleotide. The images are converted to sequence using Sequencing Control Software real time analysis (RTA) (Illumina). RTA defines bases on the basis of cluster intensities and noise estimates in a process called base calling. Up to 10 gigabases of sequence can be produced on one flow cell with a paired-end sequencing run with 50 basepairs read length (Illumina). The produced sequences can then be aligned to a reference sequence using CASAVA GERALD module (Illumina 2010). Alignment and further analyses can also be done with third party software, such as Burrows-Wheeler Aligner (BWA) (Li, Durbin 2009).


Aims of the study

The protocols provided by instrument manufacturers and application providers are not fully optimized and protocols from different manufacturers have different approaches for sample preparation. This leads to the need for optimized and standardized protocol. An important aspect for a sample preparation protocol is sufficient quality control steps which are needed in troubleshooting and also when deciding whether to continue with a sample or not.

The aim of this Master's thesis was to find optimal sample preparation protocol which would lead to more uniform and higher data acquisition from sequencing with lower costs (Table 1). Methods optimized were DNA quantification and purification for sample amounts sufficient for sequence capture. The purification optimization concentrates on finding a method which is capable of purifying enzymes and other reagents that might interfere the following sample preparation steps but also able to purify primer dimers after PCR and preferably is size selective to fragment sizes from 150 to 500 base pairs. The fragmentation parameters for sonication were also tested.

The aim was not to do basic research but rather to streamline production scale laboratory process.

Table 1 Aims of this Master's thesis.

Step	Current protocol	Aim
Library preparation	Illumina Sample preparation kit	Reduce costs.
Purification	QIAquick PCR purification column	1) Purify enzymes and other reagents to prevent interference in following steps 2) Purify primer dimers 3) Size selective for fragments from 200 to 500 base pairs
Amplification	Unknown amount of sample to amplification. Fixed number of amplification cycles. Inconsistent behaviour between samples. Fragments concatenating, artefact formation.	Uniform performance between samples. Lower number of cycles to prevent artefacts.
Quantification	Bioanalyzer	Produce more uniform data yield in sequencing by more accurate quantification
 <p>More uniform library preparation. DNA libraries with higher quality. Higher data yield. Cost reduction.</p>		

Materials and methods

Several types of sample material, including human and bacterial DNA, were used for constructing paired-end sample libraries as a part of ongoing projects. All libraries were constructed using NEBNext DNA sample preparation kits (New England Biolabs). The oligonucleotides in adapters and PCR primers were custom made by Sigma-Aldrich (St. Louis, MO, USA) according to sequence information provided by Illumina (Appendix 1). The tests conducted are listed in Table 2.

Table 2 List of tests conducted with the series number.

	Aim	Series
Fragmentation	Evaluate current fragmentation protocols	A
	Test the effect of the fragmentation length to product size	B
	Test the effect of concentration to product size	C
Purification	Test the ability of purification methods to remove small fragments	D
	Evaluate the efficiency of purification methods to remove primer dimers	E
	Test the ability of purification methods to remove small fragments from sequencing library	F
	Compare the ability of current purification method and beads to enrich optimal fragment sizes	G
Quantification	Compare qPCR and Bioanalyzer for library quantification	
Amplification	Test how amplification cycles effects the artefact production and find optimal cycle number	

Fragmentation

Bacterial and human genomic DNA from ongoing projects was fragmented with Covaris S2 –instrument with Snap-Cap microTube with AFA fiber and pre-split silicone septa (Covaris Inc). All samples were diluted in distilled water (Sigma Aldrich). Each test was conducted with only one fragmentation method per sample. Concentrations were measured with NanoDrop ND-1000 (Thermo Scientific, Wilmington, DE, USA). After the fragmentation the samples were purified with QIAquick PCR purification column (Qiagen, Venlo, The Netherlands) and one microlitre of each sample mixed with 4 microlitres of FlashGel loading dye was run in a 2,2% FlashGel (Lonza, Basel, Switzerland) according to the protocol (Appendix 2). FlashGel DNA Quant Ladder 100-1500 bp was used to determine the fragment sizes in concentration and time test and for Covaris protocols the FlashGel DNA Marker 100-4000 bp was used.

The evaluation of the current protocols using Covaris fragmentation (Series A)

The Covaris protocols for target base pair peaks of 200, 300, 500 and 700 were tested with a human genomic DNA sample with the sample volume of 120 microlitres and concentration of 25 nanograms per microlitre and 3000 nanograms of DNA in total.

The effect of the fragmentation length to product size (Series B)

In time change test (Series B) 3000 nanograms of human and bacterial DNA was fragmented according to Covaris protocol (Appendix 2 and 3) for target peak of 200 base pairs which is the current protocol in SureSelect Target Enrichment System (Appendix 2). Fragmentation times of 90, 180 and 270 seconds were used. Sample volume was 100 microlitres.

The effect of the concentration in fragmentation to product size (Series C)

In dose dependence test bacterial DNA was sufficient for sample amounts of 1500 and 3000 nanograms and human DNA for 1500, 3000 and 4500 nanograms. Sample volume was 120 microlitres due to a Covaris protocol change during these tests. The DNA was fragmented according to Covaris protocol for target peak of 200 base pairs with the fragmentation time of 180 seconds.

Purification

PCR purification was tested with two different human DNA libraries with high concentration of primer dimers and a low molecular weight DNA ladder. Each test was conducted with one purification per method. Series D, E and F were run with Bioanalyzer High Sensitivity kit after the purifications and Series G with DNA 1000 kit. The results were exported to Microsoft Office Excel (Microsoft, Redmond, WA, USA)

and cut-off size of 200 base pairs was defined in order to sort out primer dimers from sample library.

The ability of purification methods to remove small fragments (Series D)

Low Molecular Weight DNA Ladder (New England Biolabs) was diluted to 500 picograms per microlitre. This sample was purified with QIAquick PCR purification column (Qiagen), Performa DTR (dye terminator removal) gel filtration cartridge with and without SOPE resin (EdgeBio, Gaithersburg, MD, USA), NucleoSpin Extract (Macherey-Nagel, Düren, Germany) and Agencourt AMPure XP SPRI beads (Beckman Coulter Genomics, Brea, CA, USA). The control for this purification test was the DNA ladder before purification (Figure 7).

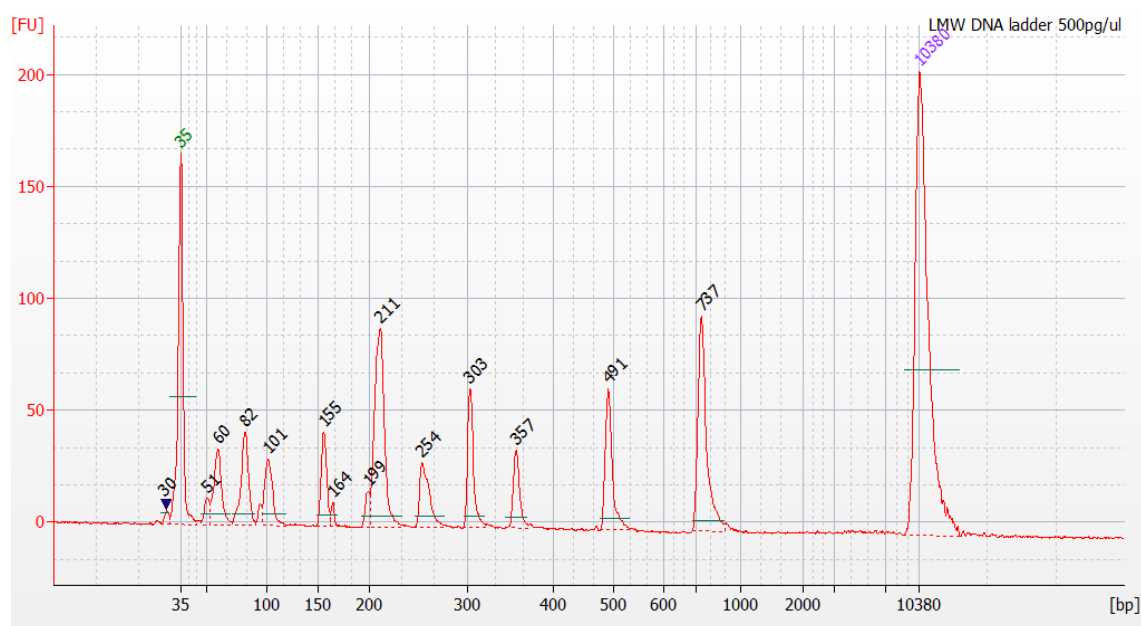


Figure 7 Bioanalyzer electropherogram illustrating DNA ladder before purifications.

The efficiency of purification methods to remove primer dimers (Series E)

One of the DNA libraries was purpose-made for testing the purification. This library was constructed from human DNA according to SureSelect Target Enrichment protocol from Agilent with the exception of using 100 picomoles of primers instead of the 25 picomoles used in the SureSelect protocol in order to create excess primer dimers. Six parallel PCR reactions with ten amplification cycles were performed to obtain sufficient amount of sample for five different purifications. After the PCR the six reactions were pooled and then evenly divided to five aliquots, one for each purification method tested. The purification methods tested with this sample were QIAquick PCR purification column, Performa DTR Gel Filtration Cartridges with and without SOPE Resin (EdgeBio), NucleoSpin Extract II columns (Macherey-Nagel) and Agencourt AMPure XP SPRI beads (Beckman Coulter Genomics). The purifications were performed according to manufacturer's protocols (Appendix 2). The control for this purification test was the QIAquick purification as it is the standard purification method in the Illumina, Agilent and Nimblegen sample preparation protocols. The sample preparation steps were done with NEBNext DNA Sample Prep Master Mix Set 1 (New England Biolabs).

The ability of purification methods to remove excess amount of primer dimers from sequencing library (Series F)

The libraries were constructed with NEBNext DNA Sample Prep Master Mix Set 1 and the capture was performed according to SureSelect Target Enrichment protocol. The libraries from this pool were purified after 10 cycles of PCR according to SureSelect protocol with QIAquick PCR purification column. 24 microlitres of each of these libraries were pooled and diluted in molecular grade water to obtain adequate amount and volume of sample for three different purification methods. The primer dimers are shown in Figure 8. The purification methods used with this sample were NucleoSpin Extract II, Performa DTR Gel Filtration cartridge with SOPE resin and Agencourt AMPure XP SPRI beads (Beckman Coulter Genomics). The samples were purified according to the protocols (Appendix 2). The control for this test was the pool of Qiagen PCR purification column purified samples before other purifications. The QIAquick

purified sample and the NucleoSpin Extract II –purified sample were eluted in 50 microlitres of 10 nanomolar Tris-HCl (pH 8) as the Agencourt AMPure XP SPRI bead – purified sample was eluted in 100 microlitres of molecular biology grade water.

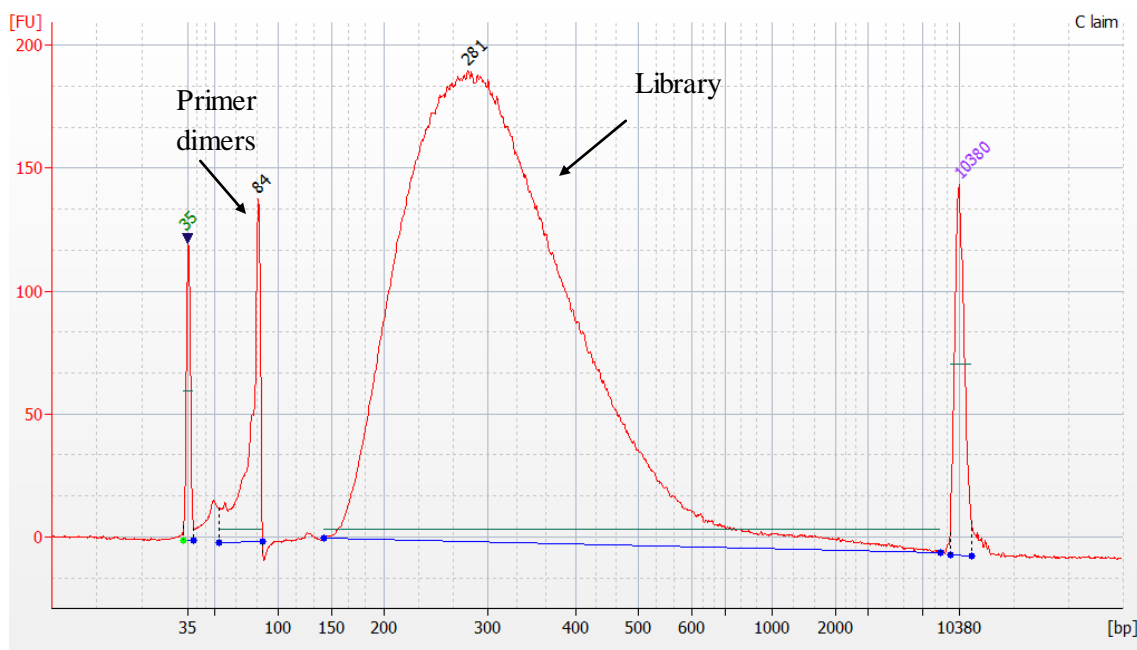


Figure 8 Bioanalyzer Electropherogram illustrating high concentration of primer dimers in a sample library before purification.

Comparison of the ability of purification methods to enrich optimal fragment sizes (Series G)

The Series A sample with target peak of 200 was purified with QIAquick PCR purification column after end repair with NEBNext DNA Sample Prep Master Mix Set 1. The sample was then purified with Agencourt AMPure XP SPRI beads following with A-tailing and ligation. After each step sample was purified with SPRI beads and after the purifications the sample was run with Agilent Bioanalyzer DNA 1000 kit. The SPRI bead purified sample after end repair was compared to the end repaired sample that was purified with QIAquick. Also the size distributions after each SPRI bead purification were compared.

Amplification

The Series G sample with target peak of 200 base pairs after ligation was amplified according to thermal amplification protocol shown in Appendix 4. The samples were purified according to QIAquick PCR purification protocol and ran in Agilent Bioanalyzer DNA 1000 kit.

Quantification

Libraries from ongoing projects were used in quantification tests. Samples were quantified with qPCR and Bioanalyzer High Sensitivity kit. Libraries were prepared according to SureSelect Human All Exon protocol (Agilent) and SeqCap EZ exome protocol (Roche Nimblegen, Madison, WI, USA) (Appendix 2). Standard curve was a 2-fold dilution series of a genomic human paired-end library prepared according to Agilent SureSelect protocol. Both the samples and the standard curve were prepared with NEBNext DNA Sample Prep Master Mix Set 1.

qPCR reactions were run with 10 picomoles of primers complementary to the PCR tails in the ends of the fragments (Appendix 1) using DyNAmo HS SYBR Green (Finnzymes, Espoo, Finland). Amplification was run according to Finnzymes protocol (Appendix 2) with LightCycler 480 Real-Time PCR System (Roche, Penzberg, Germany). Both the standard and the libraries were diluted to 10 nanomoles per litre concentration. Concentrations were then verified with Bioanalyzer High Sensitivity kit. A dilution series of 100, 50, 25, 12,5, 6,25, 3,12 and 1,56 picomoles per litre was made from the standard sample. 500-fold dilution was made from the 10 nanomoles per litre concentration sample libraries. One microlitre of each standard curve and samples were used in each 20 microlitre reaction in triplicates. The data was analyzed with LightCycler480 software (Roche) and exported to Excel.

Samples diluted to approximately 10 nanomoles per litre according to Bioanalyzer DNA 1000 kit were also quantified with Bioanalyzer High Sensitivity kit according to protocol.

The qPCR concentration was compared to the concentration obtained from Bioanalyzer. The Bioanalyzer results were used for determining the amount of library required for sequencing. The amount of library used for sequencing according to qPCR was also retraced and both methods were compared to the number of clusters in sequencing.

Results

Fragmentation

The evaluation of the current protocols using Covaris fragmentation (Series A)

The purpose of the Covaris protocol test was to evaluate the fragment size distributions produced with the protocols provided by the manufacturer. This will further help to optimize the fragmentation for longer sequencing reads. The main difference in the protocols was that as the target base pair increases the duty cycle, intensity and fragmentation time decreases (Appendix 3).

Samples fragmented according to Covaris protocol (Appendix 2) produced fragments that corresponds the target peaks given in the Covaris protocol (Table 3). Protocol for target peak of 200 base pairs however produced fragments with an average more c lose to 250 base pairs than 200 as is shown in Figure 9.

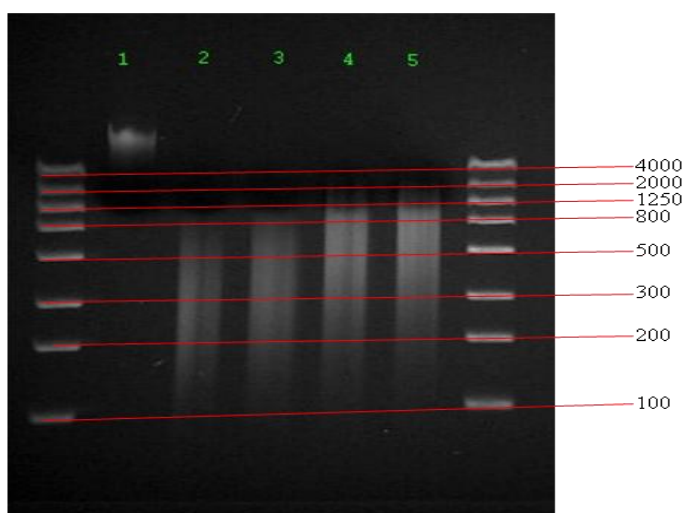


Figure 9 FlashGel after fragmentation with Covaris protocols. In lane 1 is human DNA before fragmentation. In lanes 2 to 5 are fragmented human DNA samples. In lane 2 sample fragmented with Covaris protocol with target peak 200 base pairs, lane 3 target peak 300, lane 4 target peak 500 and lane 5 target peak 700.

Table 3 Fragment sizes after Covaris protocol fragmentations. All of the protocols produced average fragment sizes equivalent to protocol.

Sample	Target Peak	Minimum fragment size	Mazimum fragment size	Average fragment size	Equivalent to average base pairs given in protocol
Human	200 bp	150	600	250	Yes
	300 bp	175	900	300	Yes
	500 bp	190	1200	500	Yes
	700 bp	190	1200	700	Yes

The effect of the fragmentation length to product size (Series B)

Some DNA samples fragmented differently from most of the samples. Fragmentations with different time lengths were tested in order to be able to adjust the fragmentation for these samples.

Human and bacterial samples have similar results. After 90 seconds treatment the fragment sizes vary from less than 100 base pairs to approximately 1500 base pairs. The majority of the fragments were in the size range from 300 to 600 base pairs. The default fragmentation time of 180 seconds and 270 seconds produces similar fragments with the mean size of 150 to 400 base pairs (Table 4). The 270 seconds produces only slightly smaller fragments than 180 seconds. (Figure 10)

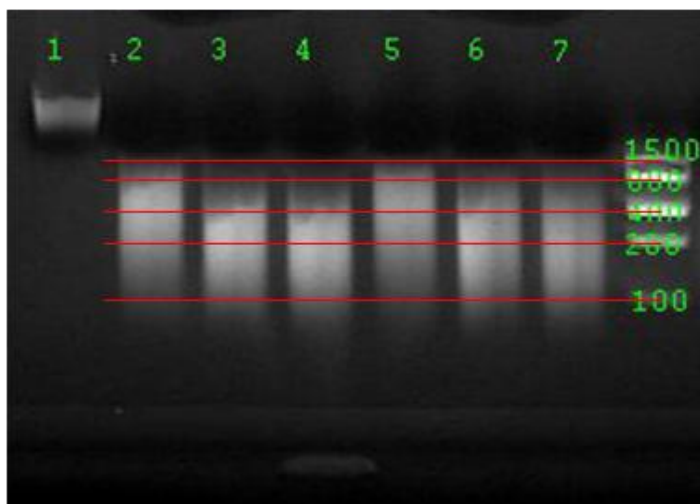


Figure 10 FlashGel after time test in fragmentation. Lane 1 is human DNA before fragmentation. Lanes 2, 3 and 4 are human DNA after 90 (2), 180 (3) and 270 (4) seconds of fragmentation. Lanes 5, 6 and 7 are bacterial DNA after after 90 (5), 180 (6) and 270 (7) seconds of fragmentation.

Table 4 Fragment sizes after time change test in fragmentation. After 90 seconds fragmentation the fragments are slightly above optimal as other fragmentation times produce optimal fragments.

Sample	Fragmentation time	Minimum fragment size	Mazimum fragment size	Average fragment size	Acceptable range
Human	90 s	<100	1500	500	Above
	180 s	<100	800	300	Yes
	270 s	<100	800	200	Yes
Bacterial	90 s	<100	1500	500	Above
	180 s	<100	800	300	Yes
	270 s	<100	800	200	Yes

The effect of the concentration in fragmentation to product size (Series C)

The purpose of the dose dependence test was to test the effect of 2-fold concentration changes to the fragmentation since the quantification methods show fluctuation between measurements and hence the concentrations between samples vary. This test will help to determine the accepted concentration range for fragmentation.

The concentration test revealed that increasing the concentration resulted in smaller fragment sizes. The fragment sizes vary from less than 100 base pairs to 800 base pairs in each concentration as is shown in Figure 11 and Table 5. The mean fragment sizes in the 1500 nanogram samples were from 200 to 400 base pairs, in the 3000 nanogram samples from 150 to 300 base pairs and in 4500 nanogram sample from 100 to 200 base pairs. Human and bacterial samples have similar fragment size distributions and this suggests that the concentration of the starting material, rather than the size of the genome affect the fragmentation efficiency.

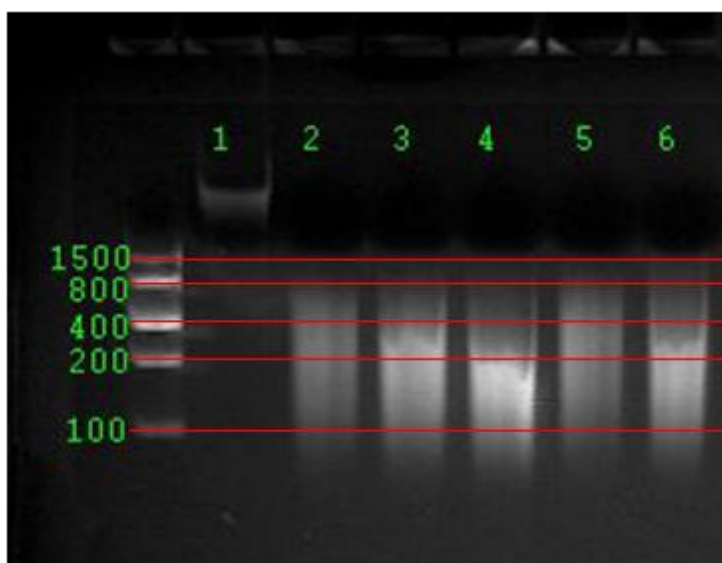


Figure 11 FlashGel after DNA concentration test in fragmentation. In lane 1 is human DNA before fragmentation, lanes 2, 3 and 4 are human DNA with 1500 (2), 3000 (3) and 4500 (4) nanograms of starting material. Lanes 5 and 6 are bacterial DNA with 1500 (5) and 3000 (6) nanograms of starting material.

Table 5 Fragment sizes after concentration change test in fragmentation. All concentrations produce fragment sizes optimal for sequencing

Sample	Amount of DNA	Minimum fragment size	Mazimum fragment size	Average fragment size	Acceptable range
Human	1500 ng	<100	800	300	Yes
	3000 ng	<100	700	175	Yes
	4500 ng	<100	600	150	Yes
Bacterial	1500 ng	<100	800	300	Yes
	3000 ng	<100	700	175	Yes

Purification

The purpose of the purification tests was to evaluate the different purification methods. The evaluation criteria were the capability to purify enzymes and other reagents, the ability to remove primer dimers and to have size selectivity from 200 base pairs to 500 base pairs. The removal of the primer dimers enables more precise quantification. As the read lengths grow, it is important to exclude fragment sizes less than the read length as the current analysis algorithms discards the reads that have been read through and also as part of the paired-end sequencing capacity is lost if the paired-end reads overlap. Current protocols either use size selection from gel to prevent this or ignore the issue. As the size selection improves the sequencing yields and the gel selection is laborious the ability of purification methods to discard fragment sizes under 200 base pairs was also evaluated in order to improve the library preparation.

The ability of purification methods to remove small fragments (Series D)

To evaluate the capability of different purification methods to remove fragment sizes less than 200 base pairs a DNA size marker was purified with NucleoSpin Extract II, Performa DTR with and without resin, AMPure XP beads and QIAquick column.

As is shown in Figure 13, Performa DTR with resin and Agencourt AMPure XP SPRI beads have the highest proportional yield of fragment sizes larger than 200 base pairs as Performa DTR without resin purifies more efficiently the larger fragments. QIAquick purifies only slightly more the smaller fragments than the larger ones. The purification efficiencies are also illustrated in Bioanalyzer electropherograms in Figure 14.

NucleoSpin Extract II column gave highest DNA yield after purification with 86 per cent yield (Table 6). Also Performa DTR without resin (83%), AMPure XP beads (81%) and QIAquick (82%) gave good yields. The yield with Performa DTR with resin was less

than half of the starting material as is shown in Figure 14. Non-purified DNA marker was used as a control.

Table 6 Table representing the data from DNA marker purification.

	Control		Performa DTR		Performa DTR with Resin		QIAquick		NucleoSpin Extract II		AMPure XP SPRI beads	
	Smaller than 200 bp	Larger than 200 bp	Smaller than 200 bp	Larger than 200 bp	Smaller than 200 bp	Larger than 200 bp	Smaller than 200 bp	Larger than 200 bp				
	154,66	204,75	136,49	162,2	48,86	108,47	113,34	182,25	119,77	189,29	101,21	189,57
Yield pg/ul	359,41		298,69		157,33		295,59		309,06		290,78	
Yield from control	100 %		83 %		44 %		82 %		86 %		81 %	
Per cent from control	100 %	100 %	88 %	79 %	32 %	53 %	73 %	89 %	77 %	92 %	65 %	93 %
Per cent of larger than 200 bp fragments	57,0 %		54,3 %		68,9 %		61,7 %		61,2 %		65,2 %	

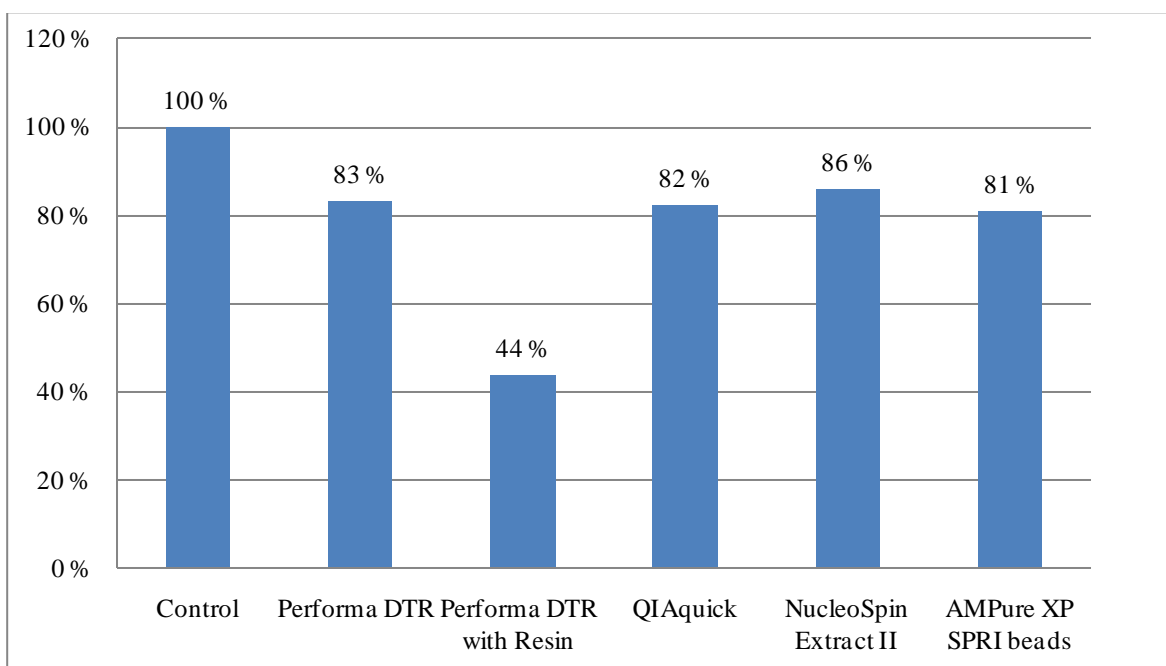


Figure 12 DNA yield after purification in DNA marker test. Yields with all purification methods are good except with the Performa DTR with resin.

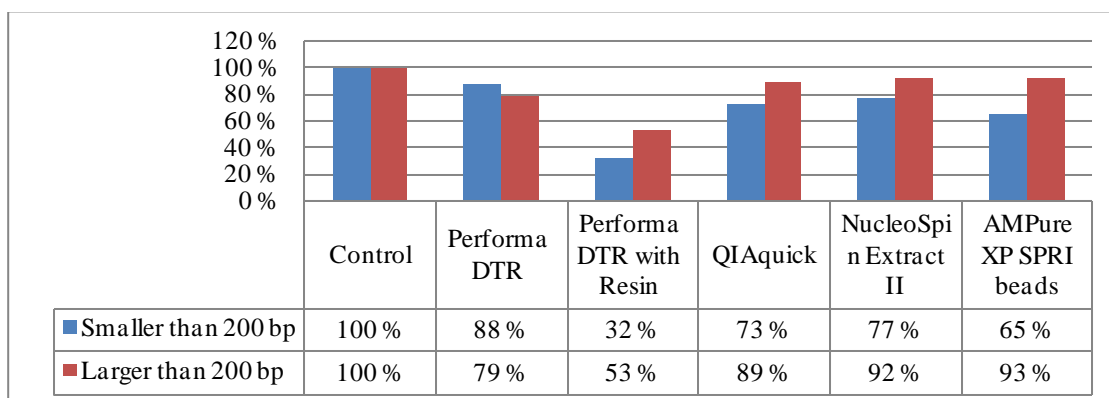


Figure 13 The ratio of smaller and larger than 200 base pair fragments after purification in DNA marker test. Performa DTR without resin purifies larger fragments more efficiently than smaller. Other purification methods purify smaller fragment more efficiently than larger fragments. From these AMPure XP SPRI beads and Performa DTR with resin most efficiently.

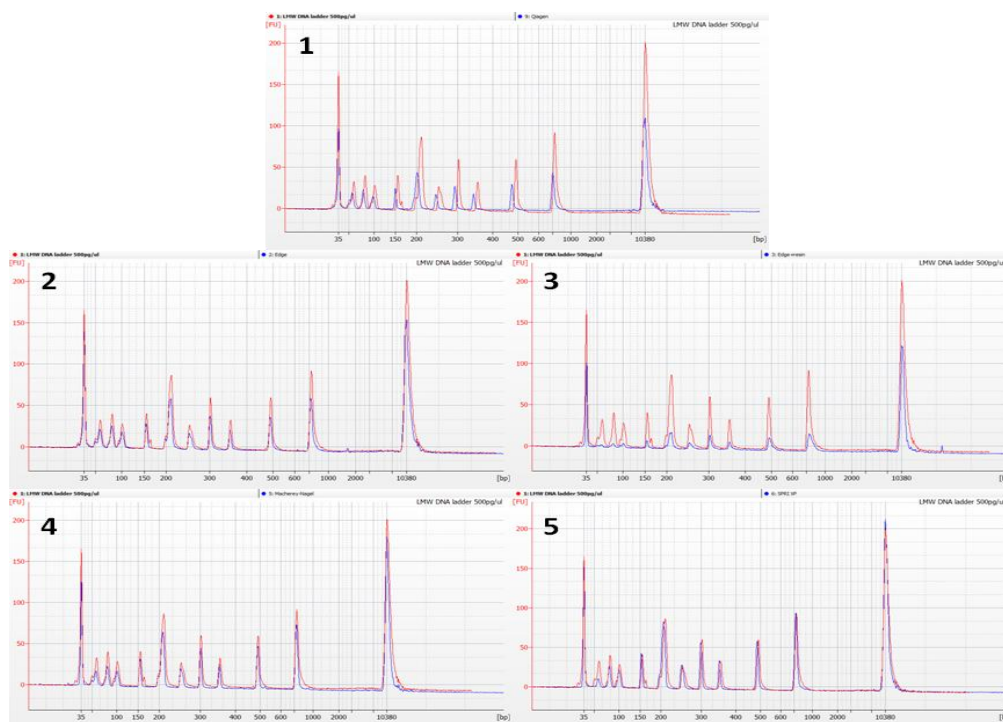


Figure 14 Bioanalyzer electropherogram illustrating the difference between DNA ladder before purification (red) and after purification with the purification method (blue). Uppermost electropherogram (1) illustrates the QIAquick column purification, 2 purification with Performa DTR, 3 Performa DTR with resin, 4 Nucleospin Extract II and 5 Agencourt AMPure XP SPRI beads. Performa DTR with resin removes the small fragments most efficiently.

The efficiency of purification methods to remove primer dimers (Series E)

The purpose-made library was a sample library amplified with the excess of primers in order to obtain a library with primer dimers. The purpose of purifying this library was to evaluate the efficiency of different purification methods to remove primer dimers.

The QIAquick column purification results were used as a control in the purpose-made library purification test since the QIAquick is the standard purification method and non-purified sample would not have been good control since the reagents might have had impact in the Bioanalyzer run.

All the purification methods except Performa DTR without resin purify the smaller fragments more efficiently than QIAquick. From these NucleoSpin Extract has the best overall yield (Table 7). Performa DTR without resin gave the highest yield (Figure 15) but most of the yield was fragments smaller than 200 base pairs (Figure 16) thus meaning that Performa DTR is purifying relatively more of the sample library than the primer dimers.

The Bioanalyzer electropherograms in Figure 17 illustrates the purification efficiencies in which the NucleoSpin Extract II purifies the primer dimers most efficiently.

Table 7 Table representing the data from purpose-made library purification.

	Control		Performa DTR		Performa DTR with resin		NucleoSpin Extract II		Agencourt AMPure XP SPRI beads	
	Smaller than 200 bp	Larger than 200 bp	Smaller than 200 bp	Larger than 200 bp	Smaller than 200 bp	Larger than 200 bp	Smaller than 200 bp	Larger than 200 bp	Smaller than 200 bp	Larger than 200 bp
	4752,7	11286,77	16089,53	6466,43	1851,47	7906,45	2298,52	11894,46	1585,46	11655,59
Yield pg/µl	16039,47		22555,96		9757,92		14192,98		13241,05	
Yield from control	100 %		141 %		61 %		88 %		83 %	
Per cent from control	100 %	100 %	339 %	57 %	39 %	70 %	48 %	105 %	33 %	103 %
Per cent of larger than 200 bp fragments	70 %		29 %		81 %		84 %		88 %	

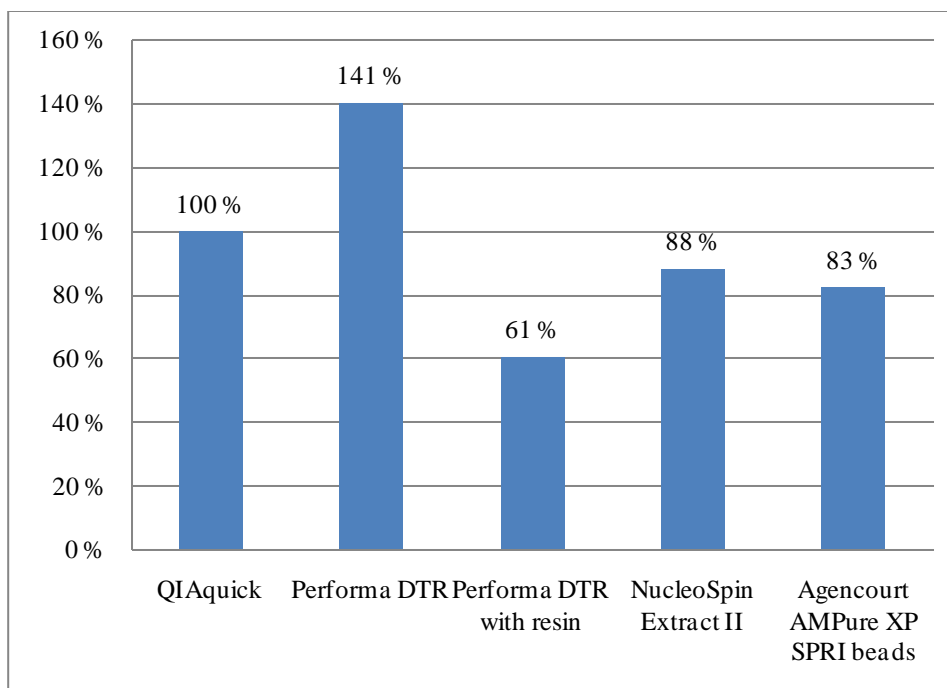


Figure 15 DNA yield after purification in purpose-made sample library test. Performa DTR without resin has the highest yield and Performa DTR with resin the lowest. NucleoSpin Extract II and AMPure XP SPRI beads both have good yield.

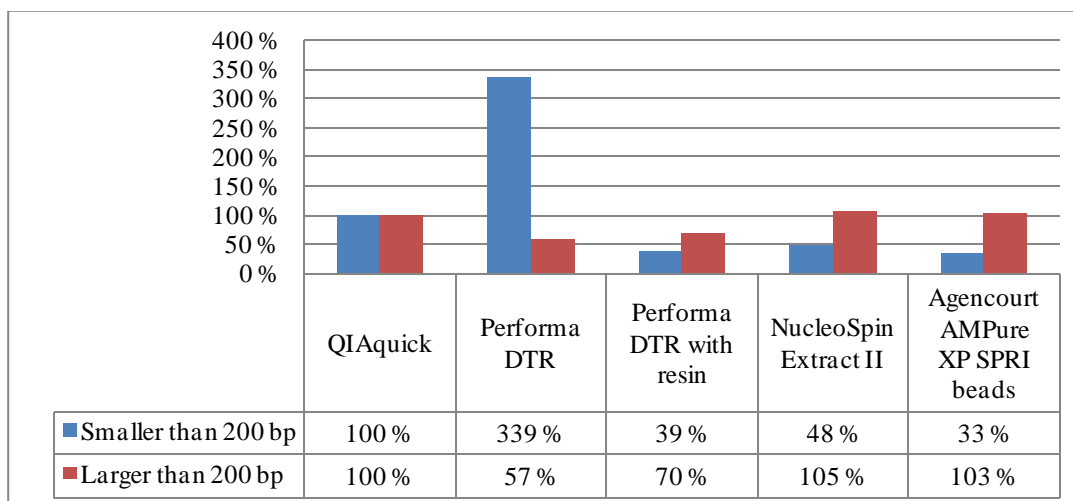


Figure 16 The ratio of smaller and larger than 200 bp fragments after purification in purpose-made library test. Performa DTR without resin purifies larger fragments more efficiently than smaller. Other purification methods purify smaller fragment more efficiently than larger fragments. From these AMPure XP SPRI beads most efficiently.

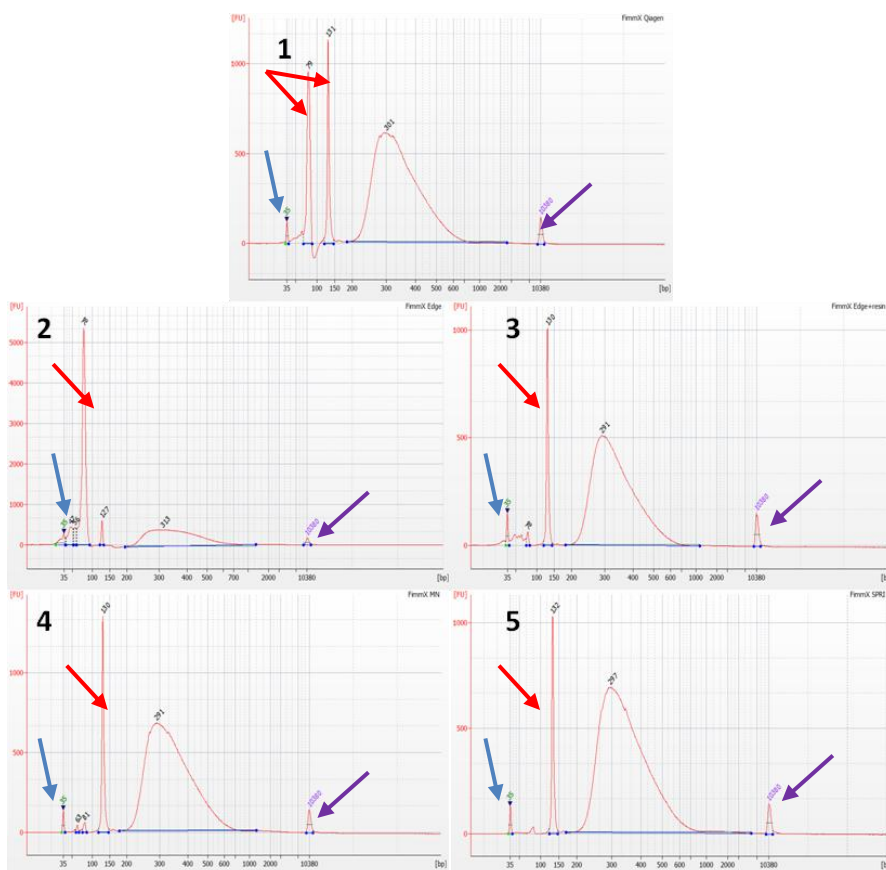


Figure 17 Bioanalyzer electropherograms illustrating the purpose-made sample after the purifications with Qiaquick (1), Performa DTR (2), Performa DTR with resin (3), NucleoSpin Extract II (4) and Agencourt AMPure XP SPRI beads (5). **Red arrow** indicates the primer dimers. **Blue arrow** indicates the lower marker with size of 35 base pairs and **violet arrow** points the upper marker with the size of 10380 base pairs.

The ability of purification methods to remove excess amount of primer dimers from sequencing library (Series F)

The Bioanalyzer electropherograms from the sample pool after purifications shows that all the methods tested with this sample removed the primer dimers (Figure 18, 20). From these the NucleoSpin Extract II columns and Agencourt AMPure XP SPRI beads have the highest yield (Figure 19, Table 8).

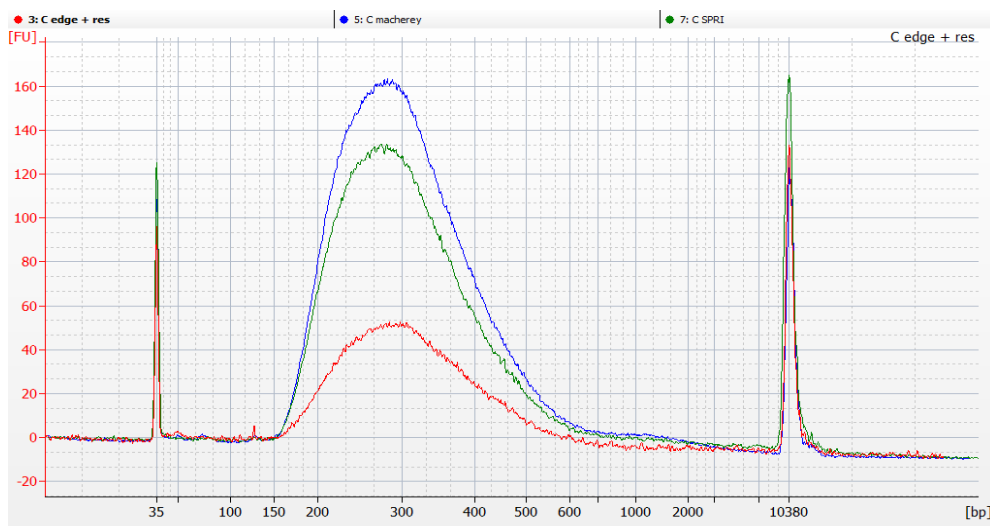


Figure 18 Overlay of Bioanalyzer electropherograms of sample pool purification test. NucleoSpin Extract II (Blue) and SPRI beads (green) gave the highest yields but Performa DTR cartridges (red) also removed the primer dimers but with lower yield.

Table 8 Table representing the data from sample library pool purification.

	Control		Performa DTR with Resin		NucleoSpin Extract II		AMPure SPRI beads	
	QIAquick		Smaller than 200 bp	Larger than 200 bp	Smaller than 200 bp	Larger than 200 bp	Smaller than 200 bp	Larger than 200 bp
	Smaller than 200 bp	Larger than 200 bp	Smaller than 200 bp	Larger than 200 bp	Smaller than 200 bp	Larger than 200 bp	Smaller than 200 bp	Larger than 200 bp
Yield	318,65	4467,88	0	1383,24	0	4416,16	0	2226,95
Yield pg/ μ l	4786,53		1383,24		4416,16		2226,95	
Yield from control	100 %		29 %		92 %		93 %	
Yield from control	100 %	100 %	0 %	31 %	0 %	99 %	0 %	100 %
Per cent of larger than 200 bp fragments	93 %		100 %		100 %		100 %	

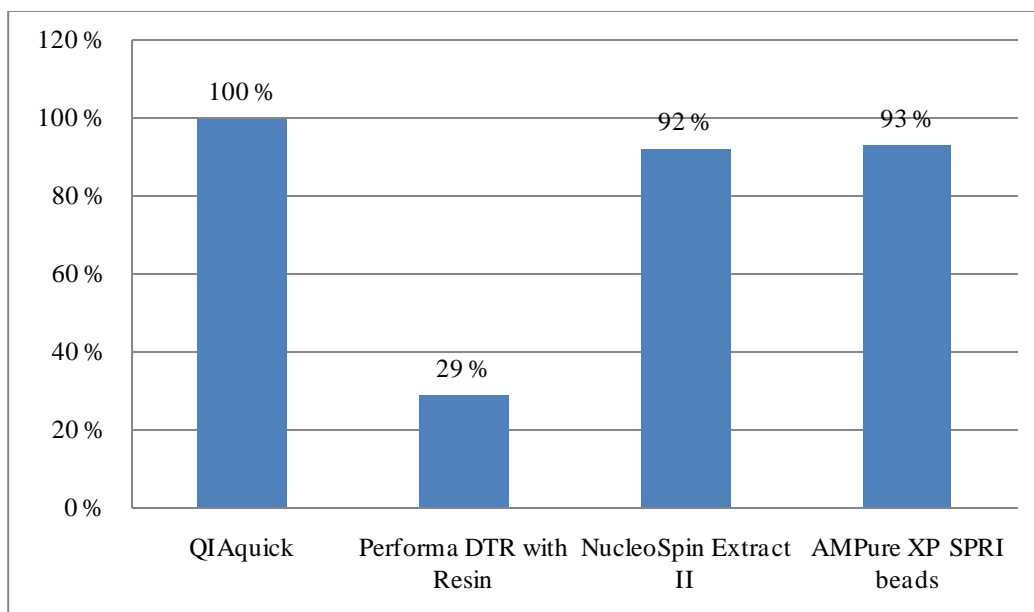


Figure 19 DNA yield after purification in project sample pool test. Performa DTR with resin gave the lowest yield as NucleoSpin Extract II and AMPure XP SPRI beads both gave high yields.

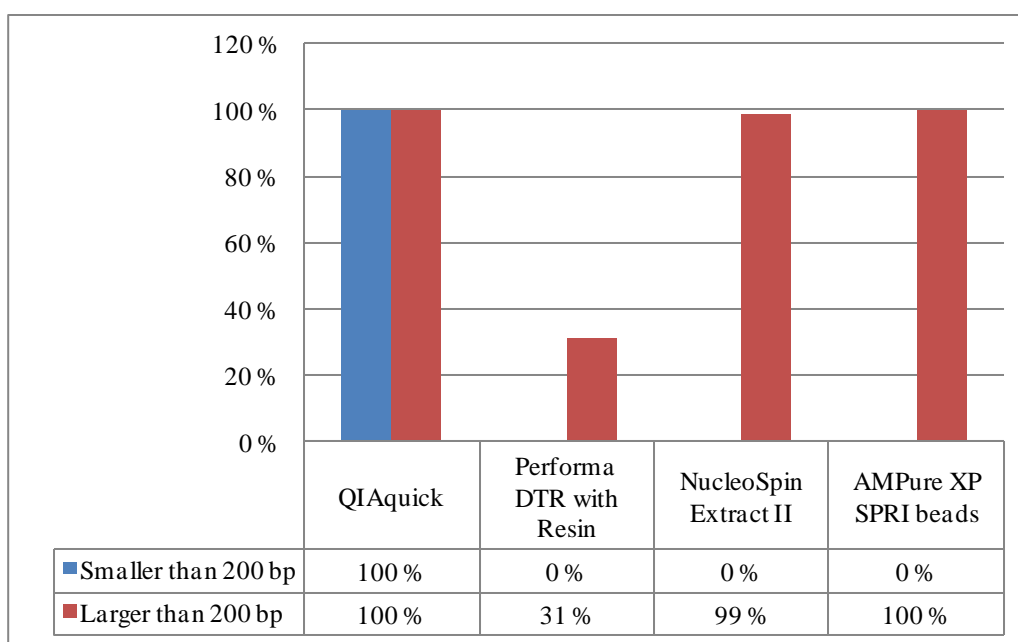


Figure 20 The ratio of smaller and larger than 200 bp fragments after purification in project sample pool test. All of the tested purification methods removed the small fragments.

Comparison of the ability of purification methods to enrich optimal fragment sizes (Series G)

The purpose of the Agencourt AMPure XP SPRI bead test was to test the size selectivity of the beads compared to the QIAquick column. The test sample was normal sequencing library. The SPRI bead size selection test reveals that SPRI beads enriches for the fragments sizes larger than 100 base pairs as QIAquick removes all fragment sizes equally (Figure 21). However, SPRI beads have lower yield in all fragment sizes.

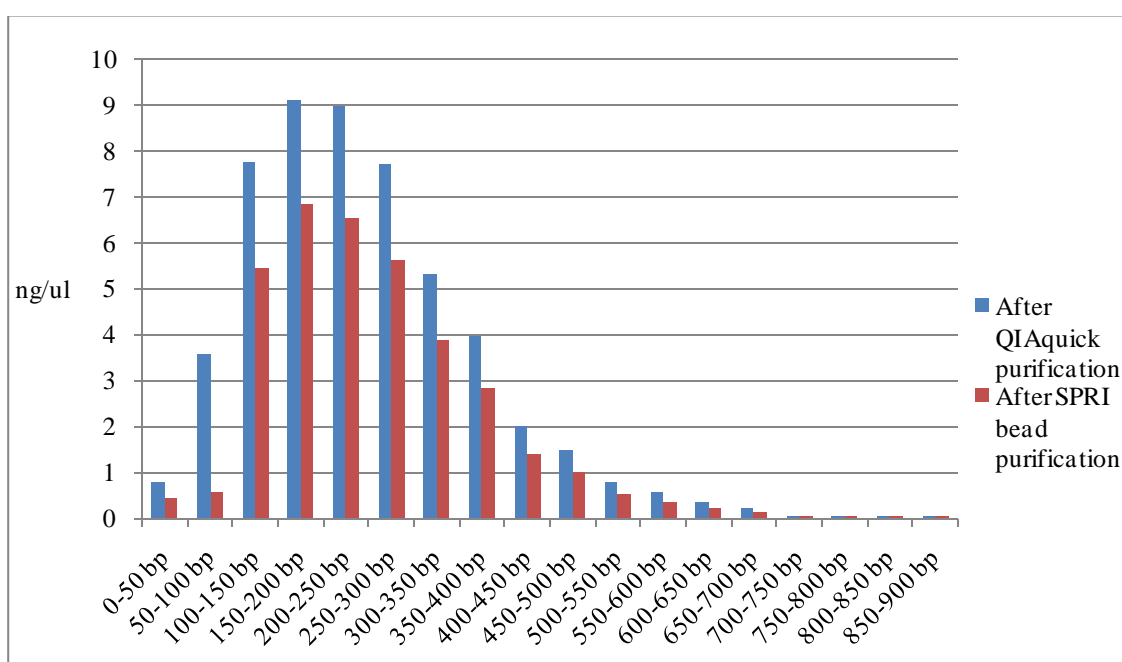


Figure 21 Covaris sample with target peak 200 after QIAquick (blue) and SPRI bead (red) purifications. SPRI beads have lower overall yield than QIAquick but SPRI beads have size selectivity to optimal fragment sizes.

The SPRI bead purification after each sample preparation step narrowed the library which can be concluded from the diminishing standard deviation after each consecutive SPRI bead purification (Table 9). This can be seen also in the Bioanalyzer overlay of each step (Figure 22).

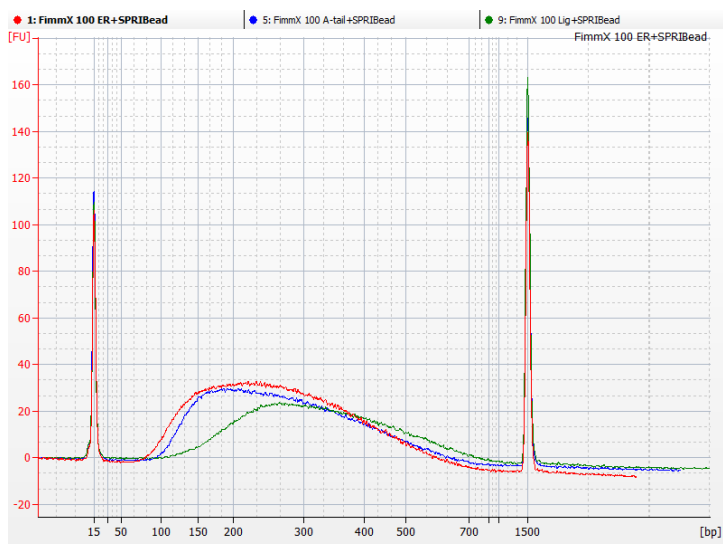


Figure 22 SPRI bead purifications after end repair (red), A-tailing (blue) and ligation (green). The size shift after ligation is due to the addition of the adapters to the ends of the fragments.

Table 9 Concentrations of different fragment sizes after each consecutive SPRI bead purification. Target fragment sizes are highlighted in green.

		Concentration		
From [bp]	To [bp]	After end repair (first SPRI bead purification)	After A-tailing (second SPRI bead purification)	After ligation (third SPRI bead purification)
50	100	2 %	0 %	0 %
100	150	15 %	13 %	3 %
150	200	19 %	21 %	11 %
200	250	18 %	19 %	18 %
250	300	16 %	16 %	20 %
300	350	11 %	11 %	15 %
350	400	8 %	8 %	12 %
400	450	4 %	4 %	7 %
450	500	3 %	3 %	5 %
500	550	1 %	2 %	3 %
550	600	1 %	1 %	3 %
600	650	1 %	1 %	2 %
650	700	0 %	0 %	1 %
700	750	0 %	0 %	0 %
750	800	0 %	0 %	0 %
800	850	0 %	0 %	0 %
850	900	0 %	0 %	0 %
900	950	0 %	0 %	0 %
950	1 000	0 %	0 %	0 %
Standard deviation		2,49	1,99	1,17

Quantification

Concentrations measured by qPCR were repeatedly lower than those measured with Bioanalyzer except for two individual samples. This is expected since qPCR counts only the molecules that have adapters ligated to both ends and are thus able to amplify in the flow cell while Bioanalyzer measures all the double stranded DNA fragments.

The average difference between qPCR and Bioanalyzer was 2,01 nanomoles per litre as the median was 2,43 and standard deviation was 2,12. These values are quite high but as the methods measures different molecules the values are tolerable. Concentrations from both methods are illustrated in Figure 23. The number of sequencing clusters better correlates with qPCR than Bioanalyzer (Figure 24) with the R^2 value of 0,628 and 0,1765 respectively.

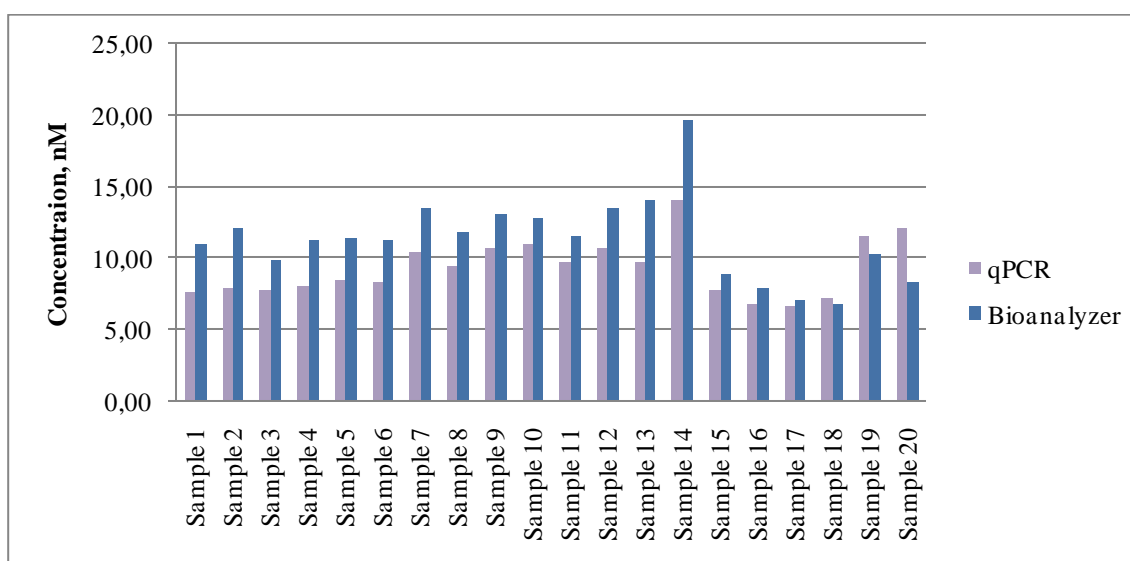


Figure 23 Library concentrations measured with qPCR (purple) and Bioanalyzer (blue). The concentrations measured with qPCR are repeatedly lower than those from Bioanalyzer.

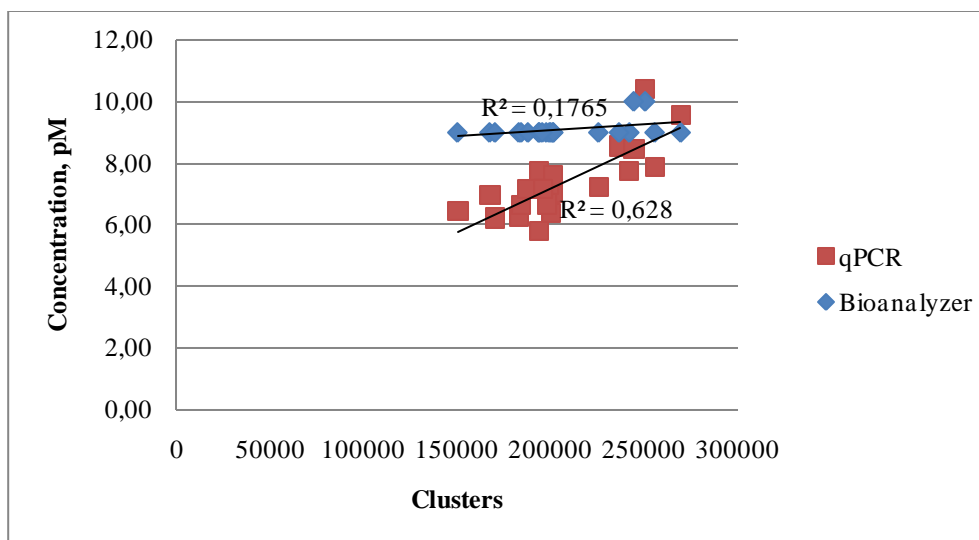


Figure 24 Clusters generated per tile according to qPCR (red) and Bioanalyzer (blue). The correlation between concentration and clusters are higher with qPCR than Bioanalyzer.

Amplification

The libraries must be amplified in order to produce sequence to the ends of the molecules which enable the attachment of the fragment to the flow cell. Amplification is also needed in order to produce sufficient amount of DNA to the capture. The amplification takes place after the ligation as the amplification primers are complementary to the adapters. Recent studies also suggest that transcriptome sequencing is possible without amplification (Mamanova et al. 2010a). This creates further interest in studying amplification free library preparation methods.

The amplification test shows that using fewer cycles, less material will be produced but also using too many cycles will produce amplification artefacts that are conjoined fragments and are shown in a Bioanalyzer electropherogram as a second peak with the fragment size double the original. The artefacts begin to form after 16 cycles of amplification and in Figure 27 this can be seen as the change in the electropherogram profile. In Figure 28 which illustrates the sample after 20 cycles of amplification the second peak is evident.

The concentration after amplification and purification is highest after ten cycles with 18,28 nanograms per microlitre as is shown in Figure 29. This is due to the reason that fragments starts to concatenate and form artefacts. This artefact formation should be avoided by using as few cycles as possible. Electropherogram of sample after 10 cycles is illustrated in Figure 26. Amplification with 6 cycles (Figure 25) has the lowest yield.

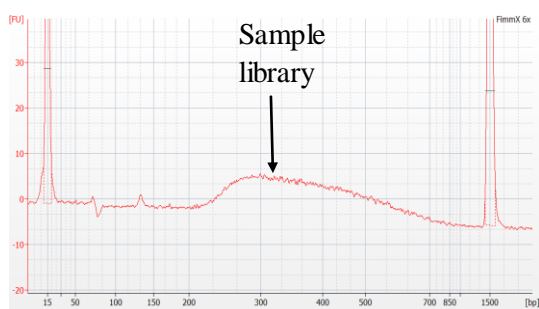


Figure 25 Bioanalyzer electropherogram from human DNA library after 6 cycles of amplification.

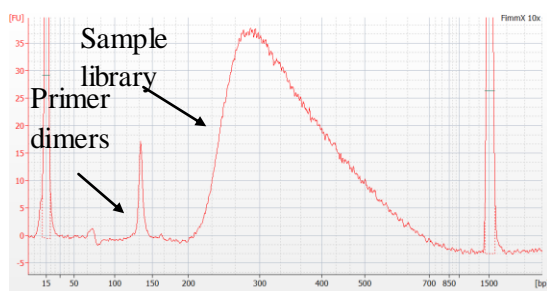


Figure 26 Bioanalyzer electropherogram from human DNA library after 10 cycles of amplification.

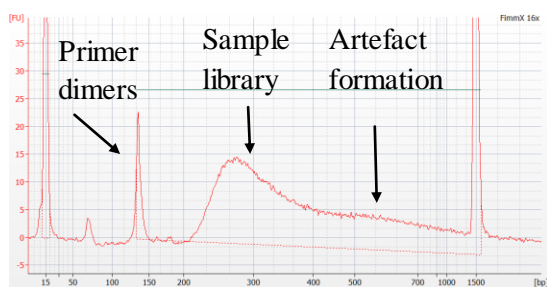


Figure 27 Bioanalyzer electropherogram from human DNA library after 16 cycles of amplification.

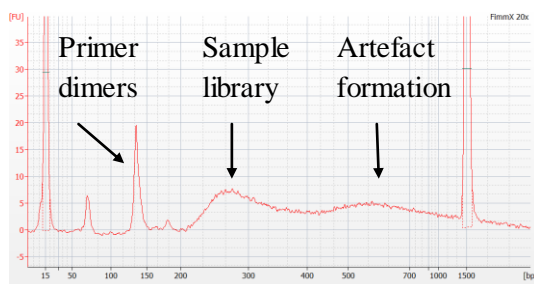


Figure 28 Bioanalyzer electropherogram from human DNA library after 20 cycles of amplification.

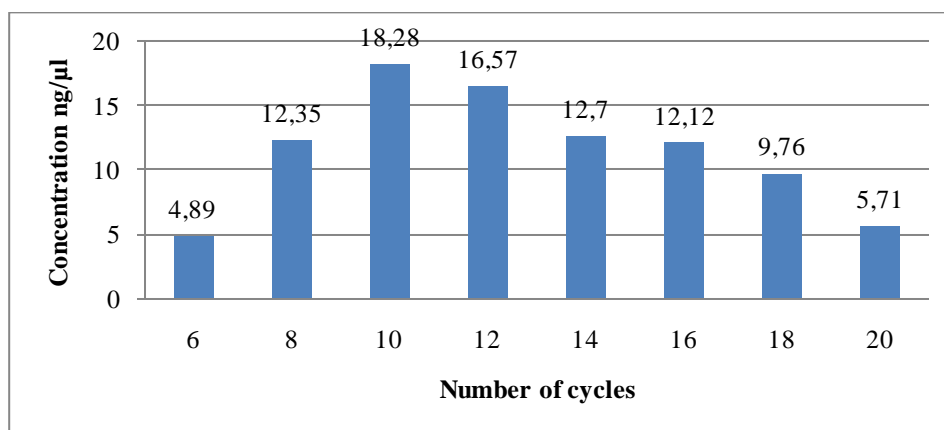


Figure 29 Concentrations of the purified samples after amplification with different cycles.

The impact of the optimization to sequencing yields

The overall library preparation costs have reduced one fifth due to the change of sample preparation kit. The purification method has been changed to Agencourt AMPure XP SPRI beads since they are easy to automate and are size selective without significant loss in sample material. The samples have more uniform performance in amplification and reduced amount of artefacts are produced. The Bioanalyzer will still be used in the sample preparation protocol for quantification and quality assessment, but the qPCR will be used for the quantification prior to sequencing. The sequencing yield has grown from the average of 550 megabases before optimization to an average of 930 megabases after optimization (Table 11). The yield has also become more uniform (Figure 30). The results and the alterations to the original protocol are represented also in Table 10.

Table 10 Alterations made to the original protocol and their affect.

Step	Original protocol	Current protocol	Improvements to original protocol
Library preparation	Illumina Sample preparation kit	NEBNext DNA Sample Preparation kit	Costs reduced to one fifth.
Purification	QIAquick PCR purification column	SPRI bead purification	1) Purify enzymes and other reagents 2) Purify primer dimers 3) Size selective for fragments from 200 to 500 base pairs 4) Automatable
Amplification	Unknown amount of sample to amplification. Fixed number of amplification cycles. Inconsistent behaviour between samples. Fragments concatenating, artefact formation.	Known concentration of DNA to amplification.	Uniform performance between samples.
		Lowered amplification cycles	Reduction of artefact production.
Quantification	Bioanalyzer	Bioanalyzer and qPCR	More uniform data yield from sequencing.

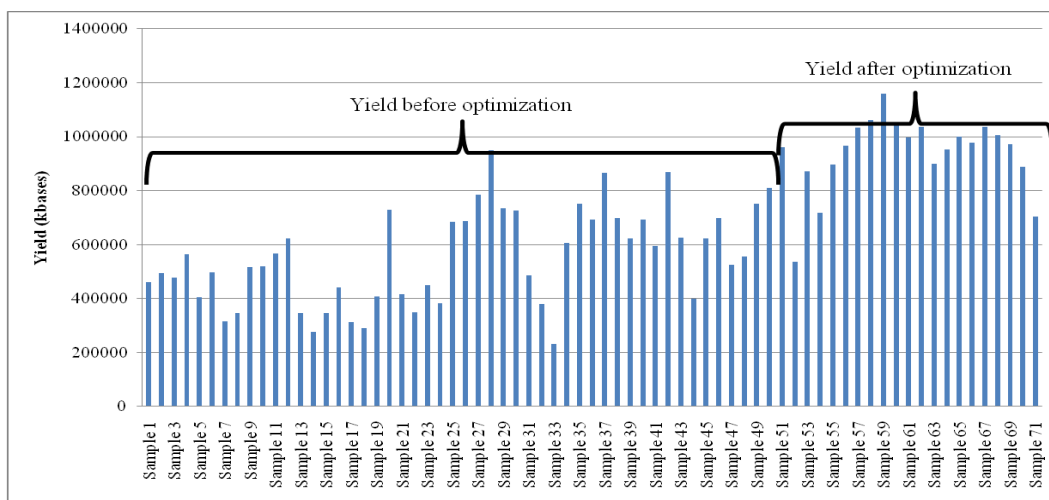


Figure 30 Yield (kbases) per lane before and after optimization. The yield is higher and more uniform after the optimization.

Table 11 Average yield (kbases) per lane before and after optimization.

	Average lane yield (kbases)	Std deviation
Before optimization	546240	173708
After optimization	933172	140286

Discussion

Since the DNA quantification methods for genomic DNA all have their defects and each method may give divergent concentration for one sample it was important to know how 2-fold changes in DNA concentration effects the fragmentation. This study suggests that within 2-fold concentration range the fragments produced do not differ dramatically and all the libraries produced contained fragment sizes suitable for sequencing.

As the fragmentation time increases the fragment sizes decrease. However, the change is not linear as halving and doubling the time does not have the same shift in the size. This information may be helpful in the future for optimizing protocols for challenging samples, that are difficult to fragment albeit having good purity and being intact before fragmentation. These challenging samples may be lost in fragmentation when the default fragmentation protocol is used, while some samples were not fragmented at all. These challenging samples are subject to future studies.

The read lengths in sequencing with Illumina Genome Analyzer II have grown from 36 base pairs to 100 base pairs. Additionally, instead of reading the sequence only from one end (single read), one can today read the fragments from both ends (paired end). This leads to the need for longer fragments in order not to sequence the insert through. This would mean gaining overlapping sequences in the middle of the sequence which would diminish the overall sequence gained. In the Covaris protocol test the fragment size distributions produced with the protocols provided by Covaris were tested. The produced fragment sizes corresponded the presumed target peaks. These protocols and the effect of the variables to the fragment size will be important information in the future as the sequencing reads lengthen. During this study the fragmentation protocol from target base pair of 200 has already grown to 300.

Another problem with small fragment sizes is the purification issue, since fragmentation methods covered in this thesis and mentioned in the introduction are also producing fragments smaller than the read length in sequencing. The main focus on purification tests in this thesis was to identify a method capable of removing primer dimers. The primer dimers interfere with the concentration measurement as they are also nucleic acid and for example qPCR is unable to distinguish primer dimers from the sample library. The primer dimers can be distinguished from the sample library with Bioanalyzer but as the primers are complementary to the adapters in the flow cell it is possible that these primers block the adapters which results to lower sequencing yield. The NucleoSpin Extract II and SPRI beads were found to most efficiently remove the primer dimers and still remain a good DNA yield. From these two the SPRI beads were selected as they are universally applicable to each step of the process, easy to automate and affordable.

In the current protocols agarose gel selection is used for selecting fragments sizes of interest. This is laborious, time consuming, results in a significant DNA loss and also introduces the possibility of contamination. Hence more efficient size selection methods were studied. Since the SPRI beads are size selective it was also tested how well the beads can produce sample library with tighter fragment distribution. The SPRI bead size selection test indicated that using SPRI beads to purify the sample after end repair, A-tailing and ligation narrowed the fragment distribution, indicating removal of both small and large fragments. This is important for the reasons described earlier and also as smaller fragments are sequenced more efficiently due to more efficient cluster amplification of smaller fragments. However, the data analysis on the necessity of also removing the larger fragments is still ongoing.

The DTR gel filtration cartridges were used in the purification tests instead of PCR purification cartridges as previous studies in the capillary sequencing group suggests that DTR cartridges with resin perform as well or even better than PCR purification cartridges.

As amplification is exponential it is important to use as few amplification cycles as possible but yet produce sufficient amount of sample for sequence capture and sequencing. In this thesis the threshold for amplification artefact production was determined. Other vital information would be the correlation between starting material, number of cycles and per cent of clonal molecules. The purpose of amplification after library preparation is to enrich perfect fragments and to gain enough sample to be able to quantify it accurately for sequencing after sequence capture and not to produce clonal molecules. These clonal molecules are discarded in the primary sequencing data analysis as they do not provide any new information. Thus they lower the overall sequencing yield and hence it is important to use as few cycles as possible, and instead, prepare parallel reactions for each sample to prevent the production of clonal molecules. This was not, however, studied in this thesis, but will be looked into in the future. Here the libraries from samples amplified with 10, 15 or 20 nanograms of DNA, using 8 to 10 amplification cycles in 3 to 5 parallel reactions produced good quality sequence with negligible amount of clonal molecules. This is visualised in a screen shot from SeqMonk software (Babraham Institute, Cambridge, UK) which illustrates reads covering exons from EDNRB gene captured with Agilent and Nimblegen exome capture kits (Figure 31). The samples were part of an ongoing project but not this thesis.

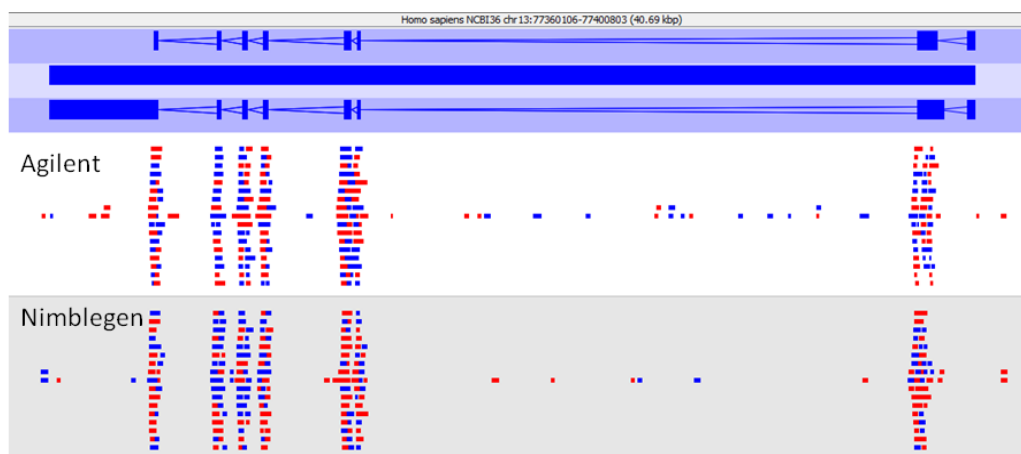


Figure 31 Reads covering the exons of EDNRB gene captured with Agilent All Exon kit and Nimblegen EZ Exome kit (Courtesy of Dr Janna Saarela and Anna-Maija Sulonen).

In order to gain maximum data yield with good repeatability it is important to accurately quantify the library. This is due to the fact that if the concentration is too high clusters will overlap and reduced amount of data is acquired. If the concentration is too low the maximum sequencing capacity will be lost.

The quantification test was only conducted with Bioanalyzer and qPCR as these were considered to produce more information on the sample. NanoDrop is unable to distinguish between substances absorbing UV light at the wavelength of 260 from DNA and hence it is not considered reliable method especially in low concentration (Gallagher SR & Desjardins PR 2008).

Comparing the two DNA quantification methods showed that the DNA concentrations measured by qPCR were consistently lower than those from Bioanalyzer. qPCR only measures fragments that have adapters in both ends and are thus able to amplify in a flow cell. Bioanalyzer High Sensitivity kit has quantitation accuracy of 20% CV and qPCR protocol includes dilutions that produce pipetting variance. This explains the variation in the concentration differences between these two methods and also explains two samples in which the concentration was higher when measured with qPCR. Thus far the amount of DNA for sequencing has been measured with Bioanalyzer but as the qPCR has a better correlation to the number of clusters it should be considered as a default method in the future. This results in even more uniform cluster numbers and reaches the highest possible cluster number. However, Bioanalyzer will still be used to determine the fragment distribution and the possible presence of primer dimers.

Real-time PCR absolute quantification of genomic DNA could be possible with universal primers but this would require a standard sample with known concentration. This concentration would require quantification with another known method e.g. with NanoDrop and thus this method would be redundant. Hoechst stains which are fluorescent dyes could be used as another method for quantification of genomic DNA. These stains bind to double stranded DNA and the fluorescent emission can be measured with standard spectrofluorometer (Invitrogen 2010).

As a result of this Master's thesis the sample preparation costs have reduced to one fifth of the original sample preparation process despite the fact that quality control steps have been added to the protocol to ensure more uniform sample libraries. The quality control steps added are Bioanalyzer DNA 1000 verification of the library after end repair, ligation and amplification to ensure the addition of the adapters and PCR tails. In the Bioanalyzer this can be seen as a size shift with the size of adapters and PCR tails. This results in lower sequencing costs with higher quality. Since these data the Genome Analyzer II has been upgraded to IIx and the data yields are even higher.

The sequencing starting from sample preparation was a slow process taking approximately three weeks from sample preparation to raw data acquisition for only a single sample. This was followed by data analysis which was the most time consuming part. The sequencing of one sequence captured sample costs almost 3000 euros. Thus the resources available for optimization, including time, funding and material, were limited. Since the aim of this study was to streamline production scale laboratory process the steps with the potential of immediate enhancement for sequencing quality were studied.

Sequencing applications and methods are an important field of study that is continuously developing and new applications are coming to market all the time. At the time of finishing this Master's thesis automated sample preparation instruments are about to come to market as well as kits that are automatable.

Sample preparation method using novel and innovative approach has also come to market. Nextera (Epicentre Biotechnologies, Madison, WI, USA) technology is based on *in vitro* transposition in which transposon complex nicks both strands of the DNA and attaches transposon sequence to the nicked end (Syed, Grunenwald & Caruccio 2009). This method is equivalent to fragmentation, end repair and ligation in the conventional sample preparation but the construction of library with adapters in the ends with Nextera is significantly faster. The low amount of starting material (50 nanograms) and quick sample preparation makes it promising method in near future. Figure 32 illustrates the Nextera workflow.

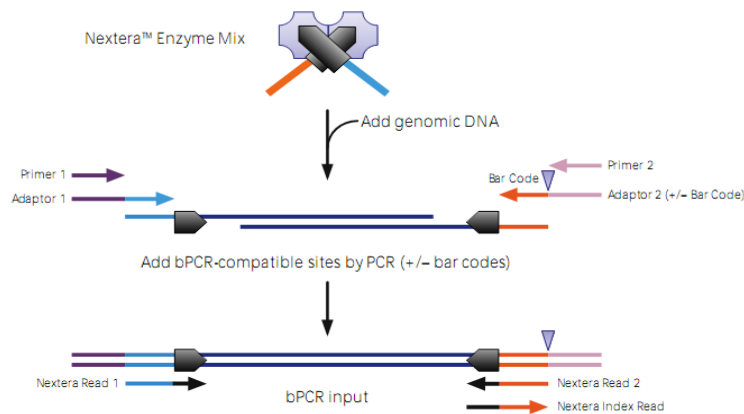


Figure 32 Nextera transposition sample preparation overview. Target DNA is fragmented and tagged with Nextera Enzyme Mix containing transposon ends appended with sequencing primer sites (blue and orange). Limited-cycle PCR with a four-primer reaction adds bridge PCR (bPCR)-compatible adaptors (purple and pink) to the core sequencing library. (Adapted from Nextera protocol, www.epibio.com)

The latest sequencing instruments at the market such as HiSeq 2000 from Illumina is capable of sequencing two whole human genomes in a single run with less than \$10,000 per sample (Illumina 2010).

The next application in sequencing will be single molecule sequencing. Several approaches have already been announced. Single molecule sequencing will reduce time expenditure and costs per nucleotide in sequencing as well as lowers the sample volumes (Treffer, Deckert 2010). It will also remove the need for amplification which introduces the risk of contamination and potentially creates biases (Fuller et al. 2009).

Even as the platforms develop the need for high quality libraries still remain in order to have high quality data.

Acknowledgements

I would like to thank my supervisors Janna Saarela and Pekka Ellonen for their irreplaceable help during this process as well as Sari Hannula, Tiina Hannunen and Suvi Kyttänen, my colleagues at FIMM Technology Centre. Arto Orpana and Petri Auvinen, thank you for providing guidance and grading my thesis. Great thanks also belong to Risto Ranta from Biotop, Isto Jänönen from Finnzymes, Christine Laine from Folkhälsan, Suvi Koskela from HUS and Inga-Lill Svedberg from University of Helsinki for providing reagents and materials.

I also want to thank my family, Paula and last but not least Antti for being there for me and pushing me forward.

References

- Ansorge, W.J. 2009, "Next-generation DNA sequencing techniques", *New Biotechnology*, vol. 25, no. 4, pp. 195-203.
- Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Cheetham, R.K., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M.J., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M.D., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Catenazzi, M.C.E., Chang, S., Cooley, R.N., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fajardo, K.V.F., Furey, W.S., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Jones, T.A.H., Kang, G., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ng, B.L., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Pike, A.C., Pinkard, D.C., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Rodriguez, A.C., Roe, P.M., Rogers, J., Bacigalupo, M.C.R., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Sohna, J.E.S., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., vandeVondele, S., Verhovskiy, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R. & Smith, A.J. 2008, "Accurate whole human genome sequencing using reversible terminator chemistry", *Nature*, vol. 456, no. 7218, pp. 53-59.
- Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., ColladoVides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. & Shao, Y. 1997, "The complete genome sequence of *Escherichia coli* K-12", *Science*, vol. 277, no. 5331, pp. 1453-&.

- Clark, J.M., Joyce, C.M. & Beardsley, G.P. 1987, "Novel blunt-end addition reactions catalyzed by DNA polymerase I of *Escherichia coli*", *Journal of Molecular Biology*, vol. 198, no. 1, pp. 123-127.
- de Magalhães, J.P., Finch, C.E. & Janssens, G. 2010, "Next-generation sequencing in aging research: Emerging applications, problems, pitfalls and possible solutions", *Ageing Research Reviews*, vol. 9, no. 3, pp. 315-323.
- Deangelis, M.M., Wang, D.G. & Hawkins, T.L. 1995, "Solid-Phase Reversible Immobilization for the Isolation of PCR Products", *Nucleic acids research*, vol. 23, no. 22, pp. 4742-4743.
- Engler, M.J. & Richardson, C.C. 1982, "I DNA Ligases" in *The Enzymes*, ed. Paul D. Boyer, Academic Press, , pp. 3-29.
- Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. 2006, "BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies", *Nucleic acids research*, vol. 34, no. 3, pp. e22.
- Fuller, C.W., Middendorf, L.R., Benner, S.A., Church, G.M., Harris, T., Huang, X., Jovanovich, S.B., Nelson, J.R., Schloss, J.A., Schwartz, D.C. & Vezenov, D.V. 2009, "The challenges of sequencing by synthesis", *Nature Biotechnology*, vol. 27, no. 11, pp. 1013-1023.
- Gallagher S,R & Desjardins P,R "Quantitation of DNA and RNA with absorption and fluorescence spectroscopy.", - *Curr Protoc Protein Sci.2008 May;Appendix 3:Appendix 4K.*, , no. 1934-3663 (Electronic).
- Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D.B., Lander, E.S. & Nusbaum, C. 2009, "Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing", *Nature Biotechnology*, vol. 27, no. 2, pp. 182-189.
- Hamaguchi, K. & Geiduschek, E.P. 1962, "Effect of Electrolytes on Stability of Deoxyribonucleate Helix", *Journal of the American Chemical Society*, vol. 84, no. 8, pp.
- Kuschel, M., Buhlmann, C. & Preckel, T. 2005, "High-throughput Protein and DNA Analysis Based on Microfluidic On-chip Electrophoresis", *Journal of the Association for Laboratory Automation*, vol. 10, no. 5, pp. 319-326.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczy, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D.,

Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissole, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H.M., Yu, J., Wang, J., Huang, G.Y., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S.Z., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H.Q., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglu, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W.H., Johnson, L.S., Jones, T.A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J.R., Slater, G., Smit, A.F.A., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J. & Int Human Genome Sequencing Conso 2001, "Initial sequencing and analysis of the human genome", *Nature*, vol. 409, no. 6822, pp. 860-921.

Li, H. & Durbin, R. 2009, "Fast and accurate short read alignment with Burrows-Wheeler transform", *Bioinformatics*, vol. 25, no. 14, pp. 1754-1760.

Linnarsson, S. 2010, "Recent advances in DNA sequencing methods – general principles of sample preparation", *Experimental cell research*, vol. 316, no. 8, pp. 1339-1343.

Mamanova, L., Andrews, R.M., James, K.D., Sheridan, E.M., Ellis, P.D., Langford, C.F., Ost, T.W.B., Collins, J.E. & Turner, D.J. 2010a, "FRT-seq: amplification-free, strand-specific transcriptome sequencing", *Nature Methods*, vol. 7, no. 2, pp. 130-U63.

- Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J. & Turner, D.J. 2010b, "Target-enrichment strategies for next-generation sequencing", *Nature Methods*, vol. 7, no. 2, pp. 111-118.
- Mardis, E.R. 2008, "The impact of next-generation sequencing technology on genetics", *Trends in Genetics*, vol. 24, no. 3, pp. 133-141.
- Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. & Rothberg, J.M. 2005, "Genome sequencing in microfabricated high-density picolitre reactors", vol. 437, no. 7057, pp. 380.
- Meyer, M., Briggs, A.W., Maricic, T., Hoeber, B., Hoeffner, B.H., Krause, J., Weihmann, A., Paeaebo, S. & Hofreiter, M. 2008, "From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing", *Nucleic acids research*, vol. 36, no. 1, pp. e5.
- Orkin, S. 1990, "Molecular-Cloning - a Laboratory Manual, 2nd Edition - Sambrook, J., Fritsch, E., Maniatis, T.", *Nature*, vol. 343, no. 6259, pp. 604-605.
- Pettersson, E., Lundeberg, J. & Ahmadian, A. 2009, "Generations of sequencing technologies", *Genomics*, vol. 93, no. 2, pp. 105-111.
- Sanger, F., Nicklen, S. & Coulson, A.R. 1977, "Dna Sequencing with Chain-Terminating Inhibitors", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 12, pp. 5463-5467.
- Schmidt, D., Wilson, M.D., Spyrou, C., Brown, G.D., Hadfield, J. & Odom, D.T. 2009, "ChIP-seq: Using high-throughput sequencing to discover protein-DNA interactions", *Methods*, vol. 48, no. 3, pp. 240-248.
- Sellars, M.J., Vuocolo, T., Leeton, L.A., Coman, G.J., Degnan, B.M. & Preston, N.P. 2007, "Real-time RT-PCR quantification of Kuruma shrimp transcripts: A comparison of relative and absolute quantification procedures", *Journal of Biotechnology*, vol. 129, no. 3, pp. 391-399.
- Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X.X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. & Church, G.M. 2005, "Accurate multiplex polony sequencing of an evolved bacterial genome", *Science*, vol. 309, no. 5741, pp. 1728-1732.
- Syed, F., Grunenwald, H. & Caruccio, N. 2009, "Optimized library preparation method for next-generation sequencing", *Nature Methods*, vol. 6, no. 10, pp. I-II.

- Treffer, R. & Deckert, V. 2010, "Recent advances in single-molecule sequencing", *Current opinion in biotechnology*, vol. 21, no. 1, pp. 4-11.
- Tucker, T., Marra, M. & Friedman, J.M. 2009, "Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine", *The American Journal of Human Genetics*, vol. 85, no. 2, pp. 142-154.
- Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.Q.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J.H., Miklos, G.L.G., Nelson, C., Broder, S., Clark, A.G., Nadeau, C., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z.M., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W.M., Gong, F.C., Gu, Z.P., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z.X., Ketchum, K.A., Lai, Z.W., Lei, Y.D., Li, Z.Y., Li, J.Y., Liang, Y., Lin, X.Y., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nuskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B.X., Sun, J.T., Wang, Z.Y., Wang, A.H., Wang, X., Wang, J., Wei, M.H., Wides, R., Xiao, C.L., Yan, C.H., Yao, A., Ye, J., Zhan, M., Zhang, W.Q., Zhang, H.Y., Zhao, Q., Zheng, L.S., Zhong, F., Zhong, W.Y., Zhu, S.P.C., Zhao, S.Y., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H.J., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferreira, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfé, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H.Y., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A.D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X.J., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M.Y., Wu, D., Wu, M.,

Xia, A., Zandieh, A. & Zhu, X.H. 2001, "The sequence of the human genome", *Science*, vol. 291, no. 5507, pp. 1304-+.

Vogelstein, B. & Gillespie, D. 1979, "Preparative and Analytical Purification of Dna from Agarose", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 76, no. 2, pp. 615-619.

Agilent, <http://www.home.agilent.com/agilent/home.jsp?cc=US&lc=eng> [2010, 3/30/2010] .

EdgeBio, <http://www.edgebio.com/index.php> [2010, 3/30/2010] .

Finnzymes, <http://www.finnzymes.com/> [2010, 4/4/2010] .

Illumina, Inc, <http://illumina.com/> [2010, 3/29/2010] .

Invitrogen, <http://www.invitrogen.com/site/us/en/home.html> [2010, 4/25/2010] .

NanoDrop, <http://www.nanodrop.com/> [2010, 4/4/2010] .

New England Biolabs, <http://www.neb.com/nebecomm/default.asp> [2010, 3/29/2010] .

Roche NimbleGen, <http://www.nimblegen.com/> [2010, 3/30/2010] .

Beckman Coulter Genomics, <http://www.beckmangenomics.com/> [2010, 3/30/2010] .

APPENDIX 1: Primer and adapter sequences

Oligonucleotide	Sequence
Adapter downstream	5'GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG3'
Adapter upstream	5'ACACTCTTTCCCTACACGACGCTCTTCCGATCT3'
PCR Primer forward	5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT3'
PCR Primer reverse	5'CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATCT3'
Sequencing Primer Read 1	5'ACACTCTTTCCCTACACGACGCTCTTCCGATCT3'
Sequencing Primer Read 2	5'CGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATCT3'
Oligonucleotide sequences © 2006-2008 Illumina, Inc. All rights reserved.	
qPCR Primer forward	5'ATACGGCGACCACCGAGAT3'
qPCR Primer reverse	5'AGCAGAAGACGGCATACGAG3'

APPENDIX 2: Protocols

Protocols	Part number	Reference
DNA Shearing with microTubes (<1.5kb fragments)	400056	http://www.covarisinc.com/
FlashGel® System	00521123-0209-01	http://www.lonza.com/
SureSelect Target Enrichment System Protocol	G3360-90010	http://chem.agilent.com/
PCR clean-up Gel Extraction User Manual NucleoSpin® Extract II		http://www.mn-net.com/
QIAquick® Spin Handbook		http://www.qiagen.com/
Performa® DTR Gel Filtration Cartridges		http://www.edgebio.com/
Agencourt® AMPure XP PCR purification	000387v001	http://beckmangenomics.com/
NimbleGen SeqCap EZ Exome Library SR User's Guide	5987407001	http://www.nimblegen.com/
SureSelect Human All Exon Kit	G3362-90001	http://chem.agilent.com/
DyNAmo™ HS SYBR® Green qPCR Kit		http://www.finnzymes.com/

APPENDIX 3: Covaris Protocols

COVARIS PROTOCOLS (Part number 400056)					Temperature	6° to 8°C (chiller set to 4°C)
Target Base Pair (Peak)	200 *	300	500	700	Power mode	Frequency Sweeping
Duty Cycle	10%	10%	5%	5%	Degassing mode	Continuous
Intensity	5	4	3	3	Volume	120 µl
Cycles per Burst	200	200	200	200	Buffer	Tris EDTA, pH 8.0 (no glycerol)
Time (seconds)	180	120	90	75	Mass (DNA)	<3µg/100 µl
* Protocol used in SureSelect Target Enrichment protocol					Water level (RUN)	S2 level 12 (E210 level 6)

APPENDIX4: Thermal amplification protocol

PCR Master Mix	1x	Final amount of substance
Phusion High-Fidelity PCR Master mix	25 μ l	
PCR Primer Forward (20 μ M)	1,2 μ l	24 picomoles
PCR Primer Reverse (20 μ M)	1,2 μ l	24 picomoles
Water	X	
DNA	15 ng	15 ng
Volume	50 μ l	

Thermal cycling		
Step	Temperature	Time
1	98 °C	2 min
2	98 °C	20 s
3	65 °C	30 s
4	72 °C	30 s
5	go to step 2 X more times	
6	72 °C	5 min
7	10 °C	forever