UPPSALA
UNIVERSITET

UPTEC X 13 011
Examensarbete 30 hp
Juni 2013

Optimisation of HaloPlex PCR technology for low input DNA resequencing

Somar Al-walai

UPPSALA UNIVERSITET

Molecular Biotechnology Programme

Uppsala University School of Engineering

UPTEC X 13 011	**Date of issue 2013-06**

Author

Somar Al-walai

Title (English)

Optimisation of HaloPlex PCR technology for low input DNA resequencing

Title (Swedish)

Abstract

A great future application for targeted resequencing is diagnostics for which the amount of isolated DNA are typically very low, such as after a biopsy. A difficulty is that cancer cells are mixed with normal cells so the frequency of mutation by sequencing is often low. Therefore, a more sensitive method is needed to avoid missing any information and to be able to distinguish sequencing errors from actual mutations. In this study, I have developed and optimised a new faster version of the HaloPlex Target enrichment technique that allows the implementation of molecular barcodes used to increase the sensitivity for rare alleles and reduce errors in sequencing data.

Keywords

PCR, next-generation sequencing, target enrichment, HaloPlex, molecular barcodes, rare allelic variants, protocol optimisation, cancer

Supervisors

Fredrik Roos
Agilent Technologies

Scientific reviewer

Lotte Moens
Uppsala University

Project name	Sponsors
Language **English**	Security **Secret until 2018-06**
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **51**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 471 4687

Optimisation of HaloPlex PCR technology for low input DNA resequencing

Somar Al-walai

Populärvetenskaplig sammanfattning

I början av 2000-talet slutfördes det humana genomprojektet HUGO som gick ut på att sekvensera människans 3 miljarder långa DNA sekvensen. Detta jätteprojekt som var ett samarbete mellan forskare från flera olika länder tog 10 år att slutföra och kostade flera miljarder dollar. Idag är sekvenseringsteknikerna betydligt snabbare och effektivare. Dessa tekniker går under samlingsnamnet nästa generationens sekvenseringstekniker. Att kartlägga ett mänskligt genom görs numera i en enda körning till ett pris om ca $5000.

HaloPlex är en produkt för provberedning från Agilent Technologies, och den uppfanns vid Uppsala universitet. I jämförelse med den klassiska PCR så är HaloPlex PCR bättre anpassat till nästa generationens sekvenseringstekniker där flera miljontals olika PCR reaktioner kan göras i ett och samma provrör. Denna metod kan användas av cancerforskare för att till exempel studera delar av arvsmassan som är kopplade till sjukdomstillstånd och för att exempelvis studera varför vissa par får sjuka barn.

Under provberedning med HaloPlex PCR så kopieras DNA molekyler till hundratusentals kopior för att möjliggöra sekvensering. Eftersom effektiviteten av DNA kopieringen kan variera för olika molekylerna kan förhållandet av variationer i genomet bli felaktigt representerade. Dessutom så kan kopieringsfel ske vilket kan leda till att en falsk variation detekteras.

För att lösa dessa problem har det utvecklats speciella sonder som letar upp och markerar DNA molekylerna med en streckkod innan kopieringen. Dessa kallas för molekylära streckkoder och med deras hjälp kan man korrigera problemen som uppstår.

I detta examensarbete har en ny snabbare version av HaloPlex utvecklats med implementeringen av molekylära streckkoder. Molekylära streckkoder ger noggrannare DNA sekvenseringen som öppnar nya möjligheter inom både cancerforskning och annan forskning på arvsmassan.

Examensarbete 30 hp
Civilingenjörsprogrammet Molekylär bioteknik
Uppsala Universitet, Juni 2013

Contents

Abbreviations

bp	base pair
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
FU	fluorescence unit
NGS	next-generation sequencing
RE	restriction enzymes
ROI	region of interest
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
U	enzyme unit
WGS	whole genome sequencing

Introduction

An international scientific project with the name Human Genome Project was initialized in 1990 with the aim of determining the 3 billion bases long human DNA sequence(HATTORI 2005). The project was a collaboration between many groups from different countries using the method of sanger sequencing. Ten years and nearly 3 billion USD later, the project was completed and the full genome was presented and made publicly available (LANDER *et al.* 2001; VENTER *et al.* 2001). Since then, the sequencing technologies have become much faster and more effective and are commonly referred to as *next-generation sequencing* technologies. Despite this rapid growth, whole genome DNA sequencing remains expensive and challenges in the analysis of raw sequencing data remains (ANSORGE 2009; GLENN 2011). Therefore, for some applications it is more useful to only sequence a targeted portion of the genome, practice commonly referred to as *targeted resequencing* (JOHANSSON *et al.* 2011). HaloPlex Target Enrichment is a product by Agilent Technologies that performs targeted enrichment and library preparation for next-generation sequencing.

During the HaloPlex Target Enrichment protocol, targeted regions, e.g. a set of cancer associated genes, are captured and amplified by PCR to increase the number of molecules for sequencing. A randomly selected subset of the amplified copies of each molecule is then sequenced (SHENDURE and LIEBERMAN AIDEN 2012). Ideally, for most applications, only one copy of each original molecule is sequenced as all amplified copies contain the same information. During the capture and amplification steps, the representation of DNA molecules of different subpopulations within the sample (e.g. the two alleles in a germline sample or mutations in a heterogeneous cancer sample) can be skewed which means that several copies of some molecules are sequenced while other molecules are not sequenced at all, leading to a potential risk of missing important sequence information about the sample. By tagging each original DNA molecule with a unique molecular barcode before PCR, all copies of this molecule will have the same barcode. The sequence reads can subsequently be tracked to their original molecules and duplicated reads can either be discarded or used to increase the confidence that the sequence read from this molecule is correct, leading to more accurate variant calls(CASBON *et al.* 2011). The introduction of molecular barcodes will also bring additional benefits to the HaloPlex technology as it provides a more accurate way of measuring efficiency during protocol development (how many unique molecules are present after PCR).

A molecular barcode consist of a library of synthetic oligonucleotides with randomly generated sequences of a certain length, in this study the length is ten bases. Since there are four kinds of nucleotides each position may obtain, the total number of unique barcode combinations will theoretically be 4^{10}= 1 048 576. (ADLER *et al.* 2008).

In the current HaloPlex protocol, implementation of molecular barcodes was not possible due to technical limitations. The project for my thesis was to modify the current protocol to make the implementation of molecular barcodes possible for the HaloPlex Target Enrichment technology.

Background

Next generation sequencing

DNA sequencing technologies have evolved much since the first sequencing technique (Sanger sequencing) was presented in 1975 (SANGER *et al.* 1977). *Next-generation sequencing* (NGS) technologies are a group of sequencing technologies performing massively parallel sequencing of genomic targets resulting in high throughput. With the introduction of next-generation sequencing technologies the cost of sequencing has been dramatically reduced. Sanger sequencing has been improved over the years and automation of the technique has been done. NGS in comparison with sanger sequencing provides higher throughput for a significantly lower cost with an output of 600 Gbp/run (ANSORGE 2009; GLENN 2011; MAINLAND *et al.* 2013). NGS platforms use the concept of cyclic-array sequencing which can be summarized as iterative cycles of enzymatic manipulation and collection of data based on imaging or chemical measurement. The sequencing procedure for NGS platforms are performed in parallel with millions of DNA fragment immobilized to a surface (SHENDURE and JI 2008). The first NGS platform named 454 GS 20 was launched in 2005 by 454 Life science and was based on massively parallel pyrosequencing. Pyrosequencing is a sequencing-by-synthesis technique where the sequencing can be performed by measuring the emitted light during the incorporation of a new nucleotide (MARGULIES *et al.* 2005). Today, the two major instrument providers are Illumina and Life Technologies. Both companies market booth high-output large machines (Illumina's HiSeq 2500 and Life Technologies' SOLiD 5500) and smaller lower output benchtop machines (Illumina's MiSeq and Life Technologies' Ion Torrent Proton). Illumina sequencing platforms use a sequence-by-synthesis technology with reversible nucleotide terminators coupled with fluorescent molecules that are released and imaged upon incorporation of a nucleotide (ANSORGE 2009) while the sequencing technology used by Ion Torrent platforms are based on flowing the reaction chamber with one nucleotide type at a time followed by detection of the hydrogen ion that is released if the nucleotide is incorporated. Both NGS platforms have much in common but also differ much in terms of run time, sequenced read length, cost per sequenced nucleotide, output, accuracy and more. Therefore, depending on the application, a different platform can be the optimal choice (ANSORGE 2009; GLENN 2011; SHENDURE 2011; SHENDURE and JI 2008; VASTA *et al.* 2009).

Targeted enrichment

For studies involving many samples or applications that require really deep sequencing to find rare variants, whole genome sequencing (WGS) can become too expensive and laborious for most labs. In that case targeted resequencing can be a better alternative especially when the application allows for limiting the study to specific regions of a genome based on prior knowledge (MAMANOVA *et al.* 2010). By using targeted resequencing instead of WGS less sequencing capacity is needed per sample making it possible to multiplex more samples or to achieve higher coverage at a lower cost (LI *et al.* 2012).

To allow targeted resequencing several technologies have been developed to capture and isolate specific regions of the genome and prepare them for sequencing. The first and most famous target enrichment method is the polymerase chain reaction, PCR, where amplification of a target region guided by a specific primer pair. But as one PCR reaction is required for each region, and the number of regions per study can be several hundred, the throughput of PCR does not match the NGS platforms. Multiplexing the number of PCR primers

per reaction can be one solution but is associated with cross reactivity leading to formation of primer-dimer and non-specific amplification (HOLLELEY and GEERTS 2009; MEUZELAAR *et al.* 2007).

For the evaluation of the performance for different target enrichment techniques different parameters are used for the comparison. The fraction of the ROI (region of interest) that has been sequenced is referred to the term coverage and is often presented in percent. Specificity is a term referring to the percentage of the sequenced reads that can be correctly aligned to the targeted regions which is a measurement of how accurate the enrichment method is. Sequencing irrelevant fragments takes unnecessary sequencing capacity that can be used elsewhere, for example to achieve higher sequence depth. Sequencing depth is how many times a certain region has been sequenced (e.g. how many reads have aligned to this region). Uniformity is a term referring to the variation of the sequence depth between the different targeted regions. Perfect uniformity is archived when all bases are covered with the same sequence depth. Reproducibility is a measurement of the target enrichments robustness, how sensitive the result is to changes in conditions. Other important parameters to take account for are cost, ease of use and amount of input DNA needed (ALBERT *et al.* 2011; HODGES *et al.* 2007; JOHANSSON *et al.* 2011).

HaloPlex target enrichment

The HaloPlex target enrichment technology provided by Agilent Technologies uses specially designed probes for the capture of ROIs. The first step (figure 1) of the HaloPlex target enrichment is digestion of the DNA sample using restrictions enzymes (RE). The RE digestion is done in 8 reactions using two enzyme pairs for each reaction (table 1). Since the reference genome is known one can predict where in the genome the RE will cleave and *in silico* design of probes can be done for the capture of specific fragments originating from the target region.

Each probe library consists of probes between a few hundreds to a few millions. These different probes are designed to target different short DNA fragment. In the second step (figure 1) a probe library containing a biotin group is added and hybridized to the target fragments. The HaloPlex probes are single stranded and biotinylated at their 5' end. For different probe libraries, different set of probes targeting specific ROIs are used. The probe arms are complementary to the targeted fragments and the centre portion of the probes is complementary to the two index primer vectors. The two index primer vectors contain universal primer motifs, sequence adapter and index barcode sequence. The sample barcode are located on vector 2 and for the new HaloPlex 2.0 Alpha target enrichment protocol developed in this study, molecular barcode sequence are located on vector 1. The index barcode is an eight nucleotide long sequence tagging fragments from each sample making it possible to sequence up to 96 samples in one sequencing run. Molecular barcodes consist of a library of randomly generated nucleotide sequences of a certain length. Incorporation of the two index primer vectors is done during the hybridization step. DNA fragments with the ligated vectors are separated from the probes by NaOH eluation and the released DNA fragment are PCR amplified making the DNA library ready for sequencing. To ensure specific capture, only correctly hybridized and ligated fragments are amplified. By using general primers designed against the index primer vectors, amplification of millions of different fragments can be done in parallel. The sequencing adapters used are specific for the used sequencing platform and since the ligation of the index primer vectors is made, fragments are ready for sequencing on the chosen platform.

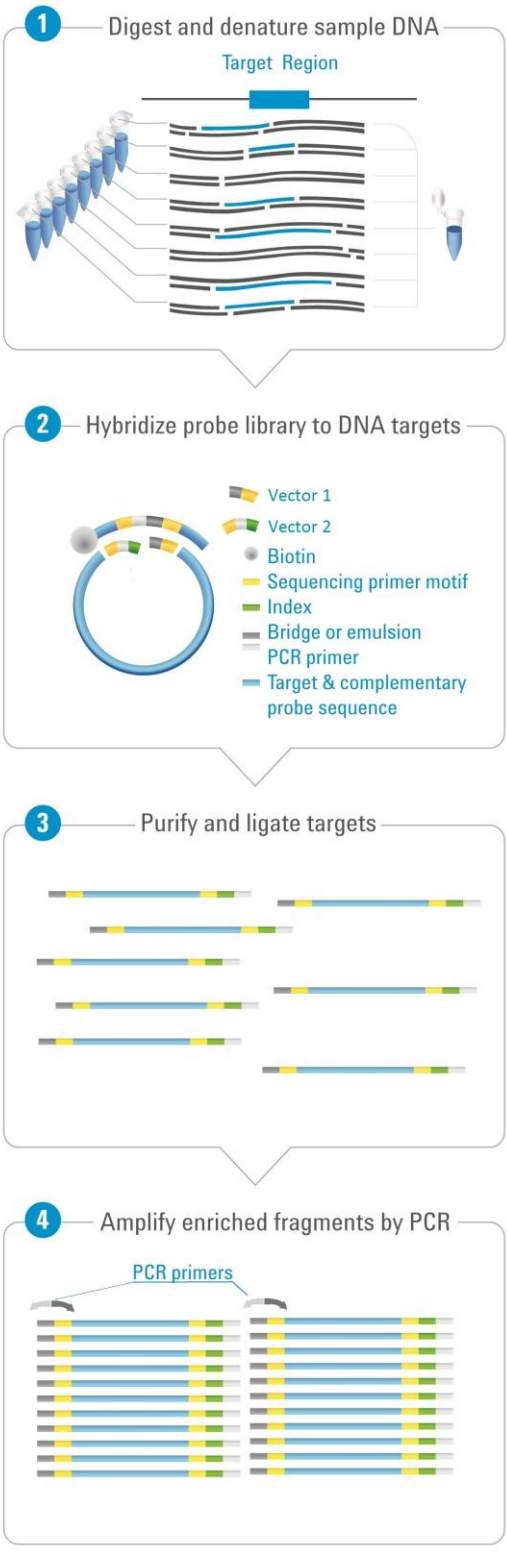

Figure 1. The steps for the current HaloPlex Target Enrichment System in a brief and simplified schematic illustration. Step 1: DNA digestion, Step 2: Probe hybridization, Step 3: Ligation of fragments and wash, Step 4: Amplification with PCR. Published with permission from Agilent Technologies.

Molecular barcodes

Molecular barcode is the idea of tagging by molecule in contrast to index barcode that is tagging by sample. Molecules are tagged with a barcode sequence unique for the molecule making it possible to identify from which molecule the sequenced fragment is derived from. Molecular barcodes consist of a library of randomly generated nucleotide sequences of a certain length, in our case with the length of ten bases. Since there are four kinds of nucleotides each position may obtain, the total number of barcode combinations will theoretically be 4^{10}= 1 048 576 combinations. Errors can occur during sequencing of the molecular barcode sequence and thus transforming the barcode to another with a different sequence. This problem can be overcome by using error correction barcodes instead. Error correcting barcodes are designed in such a way that if sequencing errors occur in the barcode sequence, the probability that the false generated barcode sequence exist in predefined molecular barcode sequence pool is very low making it possible to filter false barcode sequences or in best scenario track it back to the correct barcode (CASBON *et al.* 2011; FU *et al.* 2011; MINER *et al.* 2004). Another thing to keep in mind is the probability of assigning the same barcode sequence to different molecules, which we are referring as collision of barcodes. The probability that a collision occur can be related to a known problem in probability theory known as birthday paradox (GORT *et al.* 2006). The birthday paradox concerns the probability that, given n randomly chosen people from a population, at least two peoples having the same birthday. This probability will increase with increasing number of people and decrease with increasing number of days in a year (SAPERSTEIN 1972). Relating this to our case where molecules are assigned a barcode, the probability of collision increases for larger number of molecules and decreases for larger number of molecular barcode sequences (KLAMKIN and NEWMAN 1967; NAUS 1974; SAPERSTEIN 1972; WAGNER 2002).

The probability for no collision of molecular barcodes for n molecules and c different barcode combinations can be calculated using equation.1:

(Eq.1) $$\mathrm{P}(n,c) = 1 \times \left(1-\frac{1}{c}\right) \times \left(1-\frac{2}{c}\right) \cdots \left(1-\frac{n-1}{c}\right) = \frac{c \times (c-1) \cdots (c-n+1)}{c^n} = \frac{c!}{c^n (c-n)!}$$

In the case of ten nucleotides with 4^{10} different barcodes, the probability for n molecules can be calculated using the formula:

(Eq.2) $$\mathrm{P}\,(n) = 1 \times \left(1-\frac{1}{4^{10}}\right) \times \left(1-\frac{2}{4^{10}}\right) \cdots \left(1-\frac{n-1}{4^{10}}\right) = \frac{4^{10} \times (4^{10}-1) \cdots (4^{10}-n+1)}{4^{10n}} = \frac{(4^{10})!}{4^{10n}(4^{10}-n)!}$$

Project aim

A great future application for targeted resequencing is diagnostics for which the amount of isolated DNA are typically very low, such as after a biopsy. Another difficulty is that cancer cells are mixed with normal cells so the frequency of mutation by sequencing is often low. Therefore, a more sensitive method is needed to avoid missing any information and to be able to distinguish sequencing errors from actual mutations. The project goal was to develop and optimise a new version of the HaloPlex Target enrichment technique that allows the implementation of molecular barcodes to increase the sensitivity for rare alleles, reduce errors in sequencing data and at the same time make the protocol faster. The new proposed version of the HaloPlex enrichment is illustrated in figure 2. Another important part of the project was to demonstrate that molecular barcodes can be used for the intended purpose.

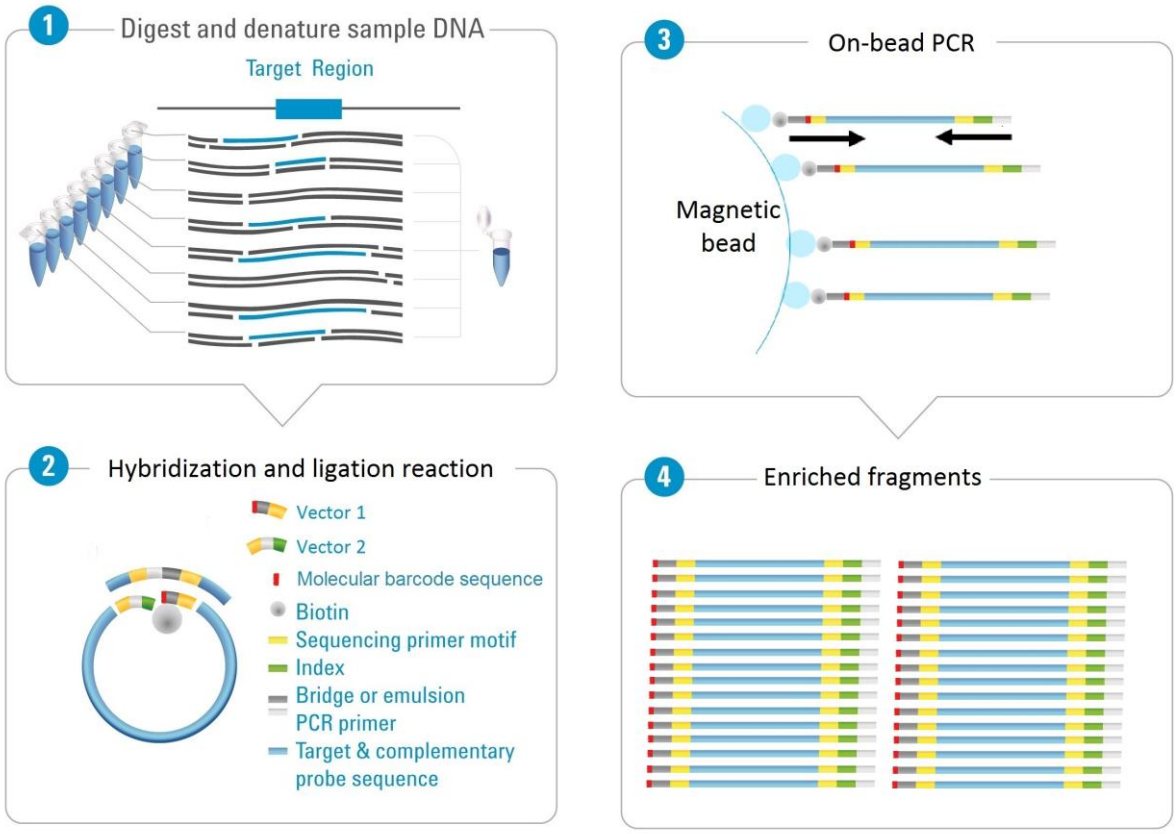

Figure 2. Illustration of the new protocol. Vector 1 fragments are with Molecular Barcodes (shown in red) while vector 2 is with Sample Barcode (shown in green). Biotin is located on vector 1. The light blue circles in step 3 illustrate the streptavidin molecules located on the magnetic beads. Step 1: DNA digestion. Step 2: Hybridizaton and ligation of vector 1 and vector 2 to fragments. Step 3: Binding of biotinylated fragment to streptavidin coated beads followed by on-bead PCR. Step 4: Binding beads to magnets and collect enriched fragments in supernatant.

Materials and methods

Various evaluation and optimisation steps have been made during the project process to evaluate the new protocol. The final protocol with the introduction of molecular barcodes is presented last in section "HaloPlex 2.0 Alpha Target Enrichment protocol". Evaluations and optimisations leading to the new protocol are presented in section "Experiment 1-5". The current official HaloPlex protocol is presented in the section below and changes in the protocol were made gradually during the development of the new HaloPlex 2.0 protocol. Genomic DNA that was used in this study was HapMap NA18507 (Coriell Institute for Medical Research)

HaloPlex Target Enrichment

The standard protocol for HaloPlex is described in this section.

Restriction enzyme digestion of genomic DNA: The digestion of genomic DNA took place in eight reactions each containing two restriction enzymes, see table 1. All restriction enzymes were obtained from New England Biolabs (NEB). Each digestion reaction contained 100 ng DNA, 0.1 U/µl of each restriction enzyme and NEB-buffers in a total volume of 10 µl. The RE reactions were incubated at 37°C for 30 min. Restriction enzyme digestion reactions were analyzed by PAGE.

Table 1. Restriction enzymes used for the different reactions.

Well orientation	Enzyme 1	Enzyme 2
A	AluI	MSII
B	BccI	MlyI
C	Bsp1286I	MluCI
D	BtgI	DdeI
E	DraI	MnlI
F	HaeIII	MseI
G	HpyCH4III	SfcI
H	HpyCH4V	StyI-HF

Hybridization of DNA fragment to HaloPlex probes and vectors: The hybridization reaction contained 200 ng digested genomic DNA, 625 nM of each vector (Agilent technologies), 10 pM of each biotinylated probe (Agilent Technologies), 0.07% Tween-20, 0.7 M NaCl, 3.5 mM EDTA, 10% formamide (Sigma) and 7 mM Tris-Hcl (pH 7.5). For the hybridization reaction, samples were first incubated at 95°C for 10 minutes followed by 3 hours incubation at 54°C.

Capture and wash: 0.4 mg of NanoLink streptavidin coated magnetic beads (Solulink) were re-suspended in 7 mM Tris-HCL (pH 7.5), 0.7 M NaCl, 3.5 mM EDTA and 0.07% Tween-20 to 10 mg/ml. 0.4 mg NanoLink beads were incubated with 160 µl Hybridized samples for 15 minutes at room temperature.
The supernatant was removed using a magnetic plate and beads were re-suspended in 100 µl wash solution (10 mM Tris-HCl (pH 7.5), 1 M NaCl, 5 mM EDTA, 0.1% Tween and 20% formamide) and incubated in 46°C for 10 minutes. Supernatant were removed using a magnetic plate.

Ligation of vectors to DNA fragments: After removal of supernatant, 50 µl of Ligation solution (1.2 U thermo stable DNA ligase (Epicentre Biotechnologies), 20 mM Tris-HCl (pH 7.5), 25 mM KCL, 10 mM MgCl2, 0.5 mM

NAD and 0.01% Triton X-100) were incubated with streptavidin beads with captured product in 55 °C for 10 minutes followed by a hold step at 4°C. The supernatant was removed using a magnetic plate.

NaOH elution and multiplex PCR: Supernatant was removed using a magnetic plate when the Ligation reaction was completed. While keeping the tubes on magnetic plate, 100 µl of SSC buffer (Agilent Technologies) was added. Supernatant was removed and 25 µl of 50 mM NaOH was added to each tube. Beads were re-suspended in solution and incubated in 1 minute at room temperature.
20 µl supernatant was collected by using magnetic plate and pooled with 30 µl PCR Master mix for the final concentrations of 1x Herculase II reaction buffer (Agilent), 0.2 mM Herc II supplied dNTP (Agilent), 0.5 µM Primer 1 (Agilent), 0.5 µM Primer 2 (Agilent), 20 mM Acetic acid (Sigma) and 0.4 U/µl Herculase II enzyme (Agilent). The PRC cycling was performed as follows: 98°C for 2 min followed by X cycles of 98°C for 30 s, 60°C for 30 s and 72°C for 1 min. The cycling was followed with a final elongation step for 10 minutes at 72°C. The cycle number for the multiplex PCR differs depending on the probe library design. The optimal cycling number for each probe library design used is presented in table 5.

Vector oligonucleotides: Oligonucleotide vectors were designed to target the general motif in HaloPlex Probes. The vectors composed of sequences for read primer annealing in Illumina sequencing. Vectors were composed of two different vectors called vector 1 and vector 2. Vector 2 contained index barcode sequence which allowed pooled samples to be sequenced in the same run. The Vector oligonucleotides used in HaloPlex 2.0 Alpha protocol were different from the ones used for this protocol. Vector 1 for HaloPlex 2.0 contained biotin in 5´end and a molecular barcode sequence. The molecular barcode sequence was composed of ten randomized nucleotides used to tag each specific molecule in the hybridization.

Experiment 1: Production of non-biotinylated HaloPlex probes

Probe PCR: HaloPlex library probes were designed and synthesized by Agilent technologies on microarrays. PCR amplification of the HaloPlex library probes were performed in a 100 ml reaction bag (Life technologies) containing 0.02 U/µl Platinum Taq DNA Polymerase (Invitrogen), 1X PCR Buffer (Invitrogen), 2 mM $MgCl_2$ (Sigma-Aldrich) , 0.2 mM dNTP (Enzymatics), 0.5 µM forward primer (IDT), 0.5 µM reverse primer (IDT). The PCR cycling was done in SOLid EZ Bead Amplifier thermal cycler (Life technologies) and was performed as follows: 95°C for 5 min followed by 18 cycles of 95°C for 15 s, 55°C for 15 s and 70°C for 1 min. The cycling was followed with a final elongation step for 5 minutes at 70°C. The optimal cycle number was evaluated by a cycle titration with the cycle numbers 12, 14, 16 and 18. An optimal amplification is when maximal amount of product have been produced without signs of non-specific products being formed. The forward primers did not contain a 5'-biotin molecule in comparison with production of biotinylated HaloPlex probes. The reverse primers were 5´-phosphorylated.

Probe purification and concentration: HaloPlex Probes were first concentrated with Amicon Ultra 10K 15ml centrifugal filters (Millipore), followed by PCR purification with Agencourt AMPure XP magnetic beads (Backman Coulter) according to the manufacturer's instructions with the exception of the amount of beads used (1.8X). The probes were concentrated a second time with Amicon Ultra 10K 0.5 ml centrifugal filters (Millipore).

Lambda exonuclease reaction: The Haloplex probes was made single stranded by treatment with 1U/µl lambda exonuclease (NEB), 1X lambda buffer and 0,8 µM HaloPlex probes in a reaction volume of 60 µl (1U lambda exonuclease was used per 10 nmol nucleotides). Reactions were incubated at 37°C or 30 minutes followed by

inactivation at 75°C for 10 minutes. The single stranded Haloplex probes was diluted to the final concentration of 80 pM / probe.

Quality Control and analysis: Aliquots from the different probe production steps was collected and run on a 6% Novex TBE gel (Invitrogen) as a quality control for the probe production. Qubit Fluorometer with Quant-iT dsDNA BR Assay Kit was used according to manufacturer's instruction to quantify HaloPlex 2.0 Probes. Aliquots from the probe production were quantified and the recovery for each step was calculated.

Experiment 2: PCR on streptavidin coupled beads (On-bead PCR)

For the new HaloPlex protocol, targeted fragments were bound to beads during PCR amplification and therefore it was important that the beads presence did not inhibit the PCR, see step 3 in figure. On-bead PCR was evaluated for different polymerases in combination with different streptavidin coated beads, (table 2). Two test experiments were performed. In the first on-bead PCR experiment, PCR was performed in the presence of streptavidin coated beads without fragments actually being attached to beads. In the second experiment, PCR was performed on biotinylated DNA fragments bound to streptavidin coated beads with added biotinylated vector 1 fragments. The addition of vector 1 was made to investigate if the biotinylated vectors would affect the binding of the fragments to the beads.

Table 2. Beads present in PCR reactions

Beads	**Polymerase**
Nanolink(Solulink)	Herculase II fusion
Nanolink(Solulink)	KAPA HiFi Hot Start
Nanolink(Solulink)	AccuPrime Pfx
Dynabeads MyOne C1	Herculase II fusion
Dynabeads MyOne T1	Herculase II fusion
Dynabeads M-280	Herculase II fusion
Agilent LodeStars 2.7	Herculase II fusion

For the first experiment, HaloPlex PCR enrichment was diluted 1:1000, 0.4 mg and 0.2 mg beads were added to samples and samples were re-amplified with 12 PCR cycles. Samples were purified with Ampure XP beads according to the description in section "Post PCR cleaning" below and run on Bioanalyzer.

For the second on-bead PCR experiment, HaloPlex PCR enrichment was diluted 1:1000 and PCR amplified with 12 cycles with biotinylated forward primers (IDT) to obtain the final product of biotinylated fragments. The biotinylated amplicons were diluted 1:1000 and 0.3125 µM biotinylated vector 1 was added. The fragments were bound to streptavidin coated beads, and re-amplified after addition of PCR master mix and primers, see figure 3.

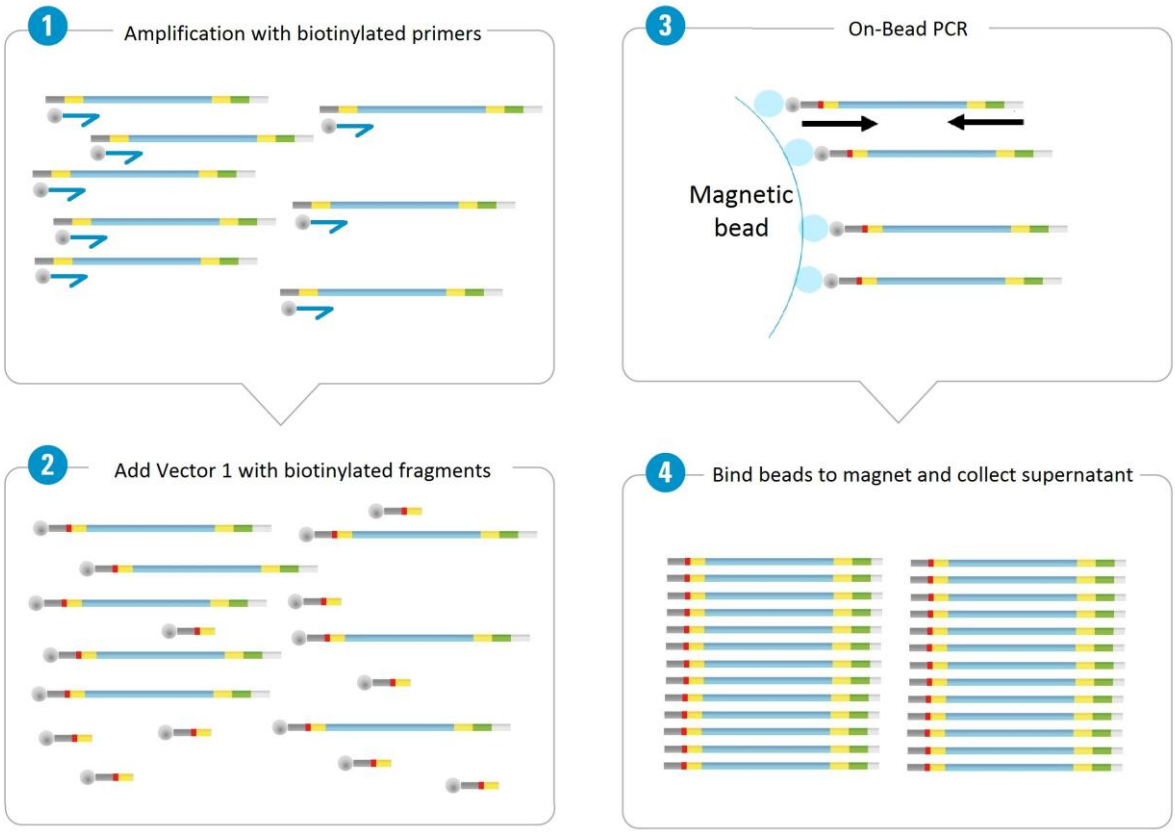

Figure 3. Schematic illustration for the On-bead PCR experiment. Biotin are illustrated with red dots, and biotinylated primers are shown in step 1 (blue fragment with red dots). Streptavidin molecules in light blue coated on magnetic beads are shown in step 3. Step 1: Final HaloPlex Target enrichment product was diluted and PCR amplified with biotinylated primers. Step 2: Diluted biotinylated fragments was mixed with biotinylated vector 1. (Step 3) Fragments and vector 1 was bound to beads and on-bead PCR was performed. Step 4: Beads were bound to magnet and the PCR products were collected for analysis on the Bioanalyzer.

The PCR cycling was done in SureCycler 8800 thermal cycler (Agilent Technologies) and was performed with the concentrations summarized in table 3 for the different enzymes.

Table 3. PCR master mix concentrations for the different enzymes tested.

Herculase II		**KAPA HiFi Hot**		**AccuPrime Pfx**	
	Reaction Conc.		**Reaction Conc.**		**Reaction Conc.**
Herculase II rxn buffer	1X	KAPA HiFi GC rxn buff.	1X	Accu Buff (+dNTP)	1X
HercII supplied dNTP	0.2 mM	dNTP Mix	0.3 mM	dNTP	0 M
Primer 1	0.5 µM	Primer 1	0.5 µM	Primer 1	0.5 µM
Primer 2	0.5 µM	Primer 2	0.5 µM	Primer 2	0.5 µM
Herculase II Enzymer	20 U	KAPA HiFi Enzyme	1 U	AccuPrime Enzyme	1 U
Total Volume:	**30 µl**		**30 µl**		**30 µl**

The PCR cycling was performed as follows:
For Herculase II polymerase (Agilent): 98°C for 2 min followed by 12 cycles of 98°C for 30 s, 60°C for 30 s and 72°C for 1 min. The cycling was followed with a final elongation step for 10 minutes at 72°C.
For KAPA HiFi Hot Start polymerase (Kapa Biosystems): 98°C for 30 s followed by 12 cycles of 98°C for 10 s, 65°C for 30 s and 72°C for 30 s. The cycling was followed with a final elongation step for 5 minutes at 72°C.
For AccuPrime Pfx polymerase (Invitrogen): 95°C for 2 min followed by 12 cycles of 98°C for 15 s, 65°C for 30 s and 68°C for 1 min. The cycling was followed with a final elongation step for 5 minutes at 68°C.

Experiment 3: Beads binding capacity

10 pM/probe (10 pM/probe in 160 µl with 12 655 different probes which gives a total of 20.2 pmol) of biotinylated probes were mixed with streptavidin coated magnetic beads of different kinds and amounts. Biotinylated probes were incubated with beads for 15 minutes at room temperature and the supernatant was collected using a magnetic plate. A dilution series of probes was created with the concentrations 10 pM, 5 pM, 2.5 pM, 1.25 pM, 0.625 pM and 0 pM.
Supernatants were amplified together with standard curve dilutions to strengthen the signal and thus make the analysis possible. The amount of beads used was 25 µg, 50 µg, 100 µg, 200 µg and 400 µg with the incubation volume of 160 µl, see table 8. Probes in the supernatant represent the proportion of unbound probes which may be used to calculate binding capacity for the different beads.

Table 4. PCR Master Mix concentrations for the amplification of probes in experiment 3.

	Article no.	**Stock conc.**	**Reaction volyme.**	**Final conc.**
Template			5 µl	0.000025X
PCR buffer	Y02028 (Invitrogen)	10X	10 µl	1X
MgCl2	Y0216(Invitrogen)	50 mM	4 µl	2 mM
dNTP	21414(Enzymatics)	25 mM	0.8 µl	0.2 mM
rev primer	20204(IDT)	20 µM	2.5 µl	0.5 µM
fwd primer	20205(IDT)	20 µM	2.5 µl	0.5 µM
Platinum Taq Pol.	10966-034(Invitrogen)	5 U/µl	0.4 µl	0.02 U/ µl
H2O			74.8 µl	
Total volume:			**100 µl**	

The PCR cycling was performed as follows: 95°C for 5 min followed by 5 cycles of 95°C for 15 s, 55°C for 15 s and 70°C for 1 min. The cycling was followed with a final elongation step for 5 minutes at 70°C. For the analysis, 5 µl samples were loaded together with 1.25 µl loading dye (per well) on a 6% PAGE gel (Invitrogen).

Experiment 4: Hybridization and ligation

In the current protocol, Hybridization of DNA fragments to probes and vectors are performed first followed by ligation of vectors to fragments in another step. The new protocol aims to perform these two processes simultaneously. A buffer capable of both hybridizing and ligating is therefore needed.

Hybridization in new buffers: Hybridization was performed acording to current HaloPlex protocol with the exception of the buffer used and the approach of inactivating the restriction enzyme. The two tested alternative buffers were the following: 1xPCR Buffer (Invitrogen), 10 mM $MgCl_2$, 1 mM NAD and for the other buffer 1x Ampligase buffer, 2 mM $MgCl_2$, 1 mM NAD. These buffers will in future be referred to as Buffer A and Buffer B respectively. Standard HaloPlex buffer were used as a positive control. Restriction enzymes were heat inactivated in 95°C for 10 minutes prior to adding the DNA to the hybridization mix. The hybridization was performed in different temperatures shown in figure 12 and 13. Probe libraries used were AACP (Agilent Technologies) consisting of 6614 different probes. Samples were incubated at 95°C for 5 min followed 16 hours incubation in respective hybridization temperature (figure 12 and 13). The rest of the target enrichment was performed according to HaloPlex protocol.

Hybridization and ligation in new buffers: Target enrichment was performed according to current HaloPlex protocol with some exceptions. The two alternative buffers that were used was Buffer A and Buffer B supplemented with 25 U Ampligase (Epicenter). Restriction enzymes were heat inactivated at 95°C for 10 minutes prior to addition of digested DNA into the reaction. Reactions were incubated at 95°C for 5 min, followed by overnight (~16h) incubation in at temperatures presented in table 6 below. The rest of the target enrichment was performed according to current HaloPlex protocol except for the ligation step that was skipped.
Different hybridization and ligation temperatures that were chosen for this step is presented in table 5.

Table 5. Temperatures and buffers used for the Hybridization experiment.

Hybridization Temperature	Buffer	Probe design	PCR cycles
60°C	Buffer A	AACP_2	23
60°C	Buffer B	AACP_2	23
54°C	Buffer A	AACP_2	23
54°C	Buffer B	AACP_2	23
60°C	Buffer A	Bayler	19
60°C	Buffer A	AACK_0	21

Experiment 5: Hybridization and ligation with formamide

Target enrichment was performed according to HaloPlex protocol with the exception of adding formamide to the new hybridization and ligation buffer (1xPCR Buffer (Invitrogen), 10 mM MgCl2, 1 mM NAD, 25 U Ampligase and X% formamide). Formamide was added to hybridization buffer in different concentrations in order to eliminate non-specific binding of probes to DNA fragments. The concentration formamide used was 10%, 7.5%, 5%, 2.5% and as a control 0%.

HaloPlex 2.0 Target Enrichment Alpha protocol

In this section, the new final protocol for HaloPlex 2.0 Alpha Target Enrichment with molecular barcodes is presented.

Restriction enzyme digestion: Digestion of genomic DNA was performed as described earlier in current HaloPlex protocol and no changes of this part were done.

Hybridization and ligation of fragments to HaloPlex Probes: Digested DNA was denatured and enzymes were heat inactivated for 10 min at 95°C. The hybridization reaction contained 200 ng digested genomic DNA, 625 nM of each vector one of which was biotinylated (IDT), 0.5 pM of each probe (Agilent Technologies), 1xPCR Buffer (Invitrogen), 10 mM MgCl2 (Sigma), 1 mM NAD (Sigma), 25 U Ampligase. For the hybridization reaction, samples were first incubated at 95 °C for 5 minutes followed by overnight incubation at 60 °C.

Fragment capture, NaOH eluation and multiplex PCR: 0.4 mg of Dynabeads MyOne T1 streptavidin coated magnetic beads (Life technologies) were resuspended in 7mM Tris-HCL (pH 7.5), 0.7 M NaCl, 3.5 mM EDTA and 0.07% Tween-20 in 10 mg/ml. Beads were incubated with 160 µl Hybridized samples for 15 minutes at room temperature. The supernatant was removed using a magnetic plate and beads were re-suspended in 50 µl fresh prepared 35 mM NaOH for 1 minute. Supernatant was removed using a magnetic plate. The re-suspension in NaOH was repeated for a total of two times. Beads were washed in 100 µl elution buffer (Qiagen)

for a total of three washes. After the last wash supernatant was discarding and beads was re-suspended with 50 µl PCR master mix (1x Herculase II reaction buffer (Agilent Technologies), 0.2 mM Herc II supplied dNTP (Agilent Technologies), 0.5 mM Primer 1 (Agilent Technologies), 0.5 mM Primer 2 (Agilent Technologies) and 0.4 U/µl Herculase II enzyme. The PRC cycling was performed as follows: 98°C for 2 min followed by X cycles of 98°C for 30 s, 60°C for 30 s and 72°C for 1 min. The cycling was followed with a final elongation step for 10 minutes at 72°C.

Post PCR cleaning

PCR products were purified from unwanted residue such as PCR primers, DNA polymerase and dsDNA artifacts less than 150 bp by using Ampure XP beads (Beckman Coulter). 2.5X Ampure XP beads of total PCR volume mixed with 1X purified Water (Acros Organics) was incubated with the PCR product for 5 minutes in room temperature with continuous shaking. Supernatant was removed by using magnetic plates. 200µl 70% ethanol was used to wash the beads for one minute for a total of two washes. Double stranded DNA was eluted from beads by incubating in 2 min with 1X Elution buffer (Qiagen).

DNA sample quantification and analysis

For the analysis of DNA, samples were mixed with 5X Blue Juice (Invitrogen) to final concentration of 1X. 5 µl of DNA sample mixed with Blue Juice was loaded into 6% Novex TBE gels (Invitrogen). Gels were run in 210 V for 15 minutes in an XCell SureLock Mini-cell Electrophoresis System (Invitrogen) with 1X TBE buffer. Gels were stained for 15 minutes in 3X Gel Red (Biotium) and pictures was taken in Red imager (Alpha Innotech). 3 µl of 25 bp and 50 bp ladders were loaded into gels according to manufacturer's protocol.

Broad range dsDNA-assay kit for Qubit (Invitrogen) was used to quantify the dsDNA for the HaloPlex Probes production. The concentration for the HaloPlex probe production QC aliquots were measured for the probe recovery calculations. The quantification was performed according to manufacturer's protocol for Broad range dsDNA-assay kit.

Purified PCR products were analyzed with High Sensitivity DNA Kit and run on 2100 Bioanalyzer (Agilent Technologies) according to manufacturer’s instructions.

Calculations and simulations for Birthday problem

Theoretical calculations, simulations and plots for the probability that different DNA molecules are assigned the same barcode were made in R statistics software and data along with R-scripts are presented in appendix 2 and 7.

Calculations and simulations for analysis of molecules with molecular barcodes (Coupons collectors’ problem)

The simulations, calculations and plots were made in R statistics software and data along with R scripts are presented in appendix 3, 4, 5 and 7.

Results and Discussion

The implementation of molecular barcodes to the current protocol has earlier encountered problems, and the current protocol needs to be changed to successfully implement molecular barcodes to the HaloPlex technology. In current HaloPlex enrichment, remaining vector oligos can be transferred into the PCR reaction and act as primers. If vector 1, with the molecular barcode “A", is used as a primer to replicate a molecule with molecular barcode “B”, a new “unique” molecule will be created. For the new protocol, vector 1 and 2 are removed or made inactive before PCR thus eliminating the possibility for them to act as primers. For the removal of remaining vector 1 before the PCR, vector 1 is biotinylated and captured with streptavidin magnetic beads (step 3 figure 2). We investigated to what extent this streptavidin bound vector interferes with PCR. As vector 2 is not biotinylated it can effectively be removed with the supernatant. A prerequisite for the new protocol is that ligation of vector to fragments should be done before the capture step, see step 2 in figure 2.

The proof of concept for the new protocol was fist done in different sub steps for better understanding of the possible difficulties that may arise during the development.

During the development of the new protocol, target enrichment was mainly analyzed by bioanalyzer electropherograms. As it is important to know how to interpret these to understand the results two typical bioanalyzer electropherogram are shown in figure 4 where (A) show a successful enrichment and (B) an unsuccessful one. Basically, the graph shows the amount of product (y-axis) for different fragment lengths (x-axis). The peaks in figure 4 (A) correlate with what is expected from this HaloPlex capture with amplicons ranging between 150-650 bp. The bioanalyzer electropherogram shown in figure 4 (B) shows random peaks with no clear pattern in the area between 150-650 bp indicating no enrichment of targeted fragments.

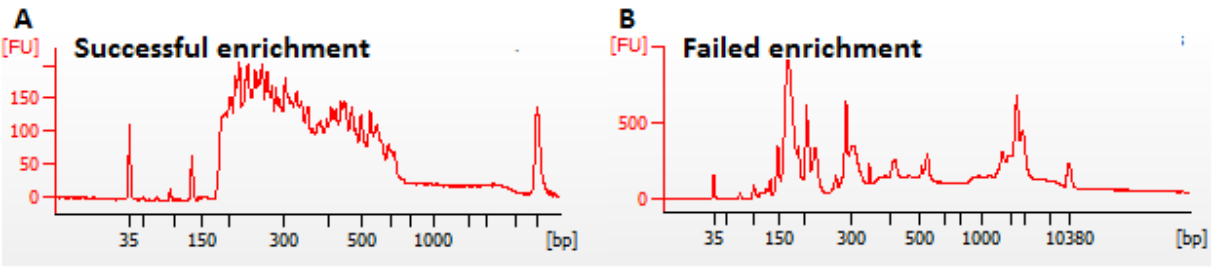

Figure 4. Bioanalyzer electropherograms for a successful enrichment (A) in comparison with an example of a failed enrichment(B). The first and last peaks in the bioanalyzer electropherograms represents the markers.

Molecular barcode vectors quality control

To test the new vectors containing the molecular barcodes, they were used in a standard HaloPlex target enrichment and the result was compared with standard HaloPlex Vectors, see table 6. Negative control, using no vector resulted in no product, which was expected. Target enrichment with molecular barcodes vectors (sample 3-5, table 6) resulted in lower product yield compared to the vectors without molecular barcodes (sample 1-2, table 6). There is also a shift in amplicon size with approximately 10bp which is the length of the additional 10 bases of the molecular barcode sequence. Despite the lower yield with molecular barcode vectors, this experiment showed that the vectors are functional.

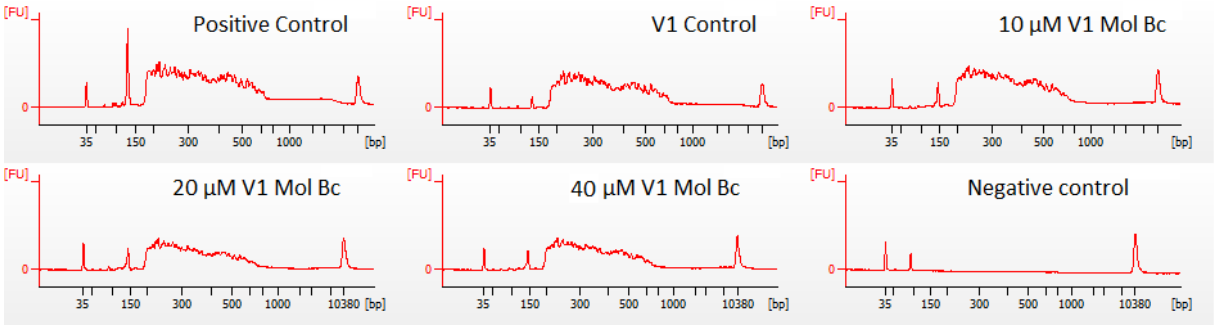

Figure 5. Bioanalyzer electropherograms for the quality control of the molecular barcode vectors (IDT). Sample 2 are IDT synthesizes standard vector 1 sequence, and was used as a control for the IDT vectors. For the positive control (sample 1), standard HaloPlex vectors were used and for the negative control (sample 6), no vector 1 was added to the samples. In sample 3, 4 and 5 the concentration Vector 1 with molecular barcodes that was used was 10 µM, 20 µM and 40 µM respectively.

Table 6. Concentrations and yield for the QC of new Molecular barcode vectors.

Sample	Description	Concentration	Product Yield
1	Pos Ctrl	20 µM	139.5 nM
2	V1 ILM	20 µM	135.0 nM
3	V1 Mol Bc	20 µM	81.5 nM
4	V1 Mol Bc	40 µM	64.0 nM
5	V1 Mol Bc	80 µM	66.0 nM
6	Neg Ctrl	0 µM	0 nM

Experiment 1: Production of non-biotinylated HaloPlex probes

For the probe production, the optimal cycle number for amplification was evaluated by a cycle titration presented in figure 6. The optimal cycle number is the highest cycle number with no non-specific PCR product (smear above the true product). Cycle number 16 showed high intensity at the expected size of ~145 bp with little or no non-specific PCR product and was therefore chosen as the optimal cycle number. The cycle titration PCR was performed in SureCycler 8800 thermal cycler (Agilent Technologies) while the large scale probe production was performed on SOLid EZ Bead Amplifier thermal cycler (Life technologies). Past experience has shown that the heat transfer has not been as effective on SOLid EZ Bead Amplifier thermal cycler compared to SureCycler 8800 thermal cycler and therefore this was compensated for with two extra cycles. Thus 18 cycles was used for the probe production in the SOLid EZ Bead Amplifier thermal cycler.

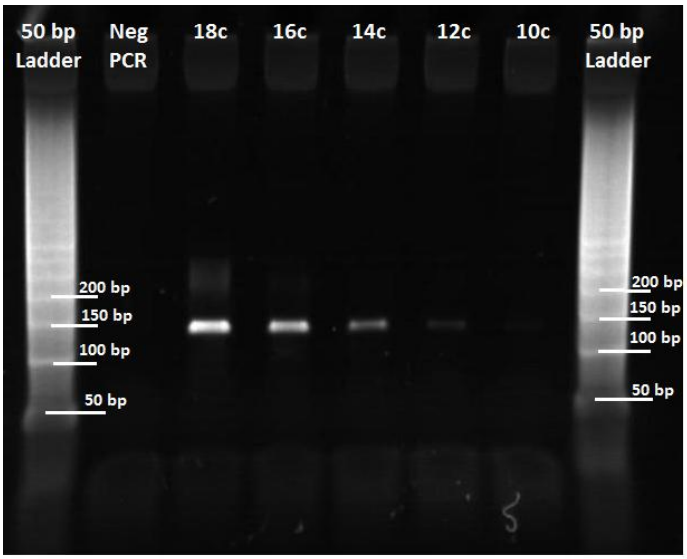

Figure 6. PAGE gel image for the cycle titration of AACP-2 probe production. First and last well are 50bp ladders. The negative PCR are without the addition of diluted probe library. The remaining wells consist of the cycle titration from the cycle number 10 to 18 PCR cycles.

Quality control for production of probes: Aliquots from the different probe production steps that were run on a 6% Novex TBE gels are presented in figure 7. In the negative control no product could be observed while a clear band around 150 bp was observed for the positive PCR reaction (figure 7). The length of the probe pool was 145 bp which corresponded well with the bands observed in figure 7. For the aliquots taken after the first Amicon concentration step (Amic. 15), one could observe a stronger signal which is what is expected after concentration of the PCR product. The aliquot that was taken after purification with Ampure XP beads resulted in significantly less PCR rest product, such as primers, suggesting a successful cleanup of the reaction. After Lambda Exonuclease treatment, no clear band could be observed at 145 bp anymore but instead the product looked more smudged on a larger area which is expected for single stranded DNA that migrates differently compared to double stranded DNA when run on a PAGE gel. The quality control was approved and the data presented in figure 7 indicates a successfully performed probe production.

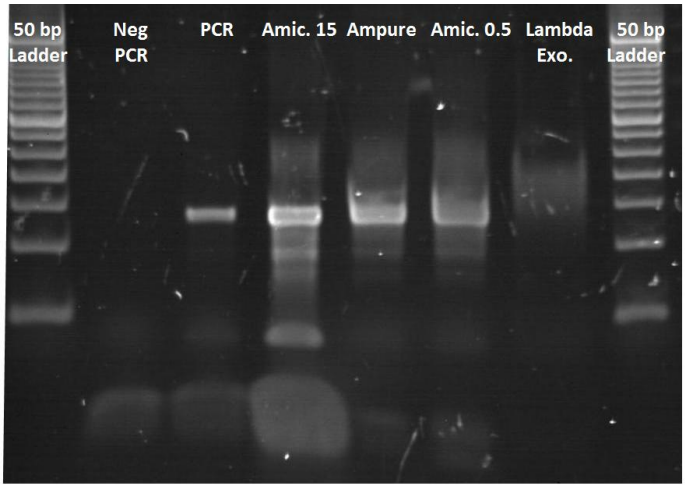

Figure 7. PAGE gel image for the quality control of the probe production. Aliquots for the different steps in the probe production were run on gel. First and last well are 50bp DNA ladders. The negative PCR is without the addition of diluted probe library. Well 3 is the post-PCR aliquot followed by well 4 that is after concentration with Amicon 15ml, followed by well 5 that is after Ampure XP purification, followed by well 6 that is the aliquot taken after the second concentration with Amicon 0.5ml. The second last well from the right is the aliquot taken after the Lambda Exonuclease treatment.

The concentrations of the aliquots were measured with Qubit and the recovery after purification and concentration was calculated, see table 7. The final recovery after purification and concentration was 41% of the post-PCR product which is considered to be a good recovery value. With values obtained by Qubit, the probe concentration could be estimated and dilution to final concentration could be made. The final concentration of the diluted probes was 80 pM/probe in a total volume of 1640 µl which is enough probes for 82 HaloPlex reactions (20 µl per reaction).

Table 7. Qubit values, concentration and calculated recovery after each step in the probe production.

								Recovery	
Step	Vol. for Qubit	Qubit	Dilution	Concentration		Vol. µl	Amount	Step	Total
NEG PCR	2	0	100	0	ng/µl	100	0	-	-
PCR	2	0,0489	100	4,89	ng/µl	65000	317850		
Amicon 15ml	2	1,15	100	115	ng/µl	1700	195500	62%	62%
Ampure	2	0,706	100	70,6	ng/µl	1700	120020	61%	38%
Amicon 0.5ml	10	7,7	200	1540	ng/µl	85	130900	107%	41%

Experiment 2: PCR on streptavidin coupled beads (On-bead PCR)

Evaluation of NanoLink Streptavidin beads presence in PCR reactions:

In this experiment, we evaluated whether the presence of streptavidin beads in PCR affects amplification performance. Four different commercially available beads were tested (NanoLink, MyOne C1, Dynabeads M-280 and agilent Lodestars), using three different polymerases (Herculase II, Kapa HiFi and AccuPrime Pfx).

The presence of NanoLink streptavidin magnetic beads (Solulink) inhibited all three tested polymerases during PCR resulting in no amplification. For the control where no beads were added, PCR was successfully performed with amplification of fragments, see figure 8. For the new protocol, on-bead PCR was a requirement and therefore different beads needed to be evaluated. The fact that the presence of NanoLink beads inhibited all three tested enzymes with no product at all indicated that on-bead PCR with NanoLink would not be an option.

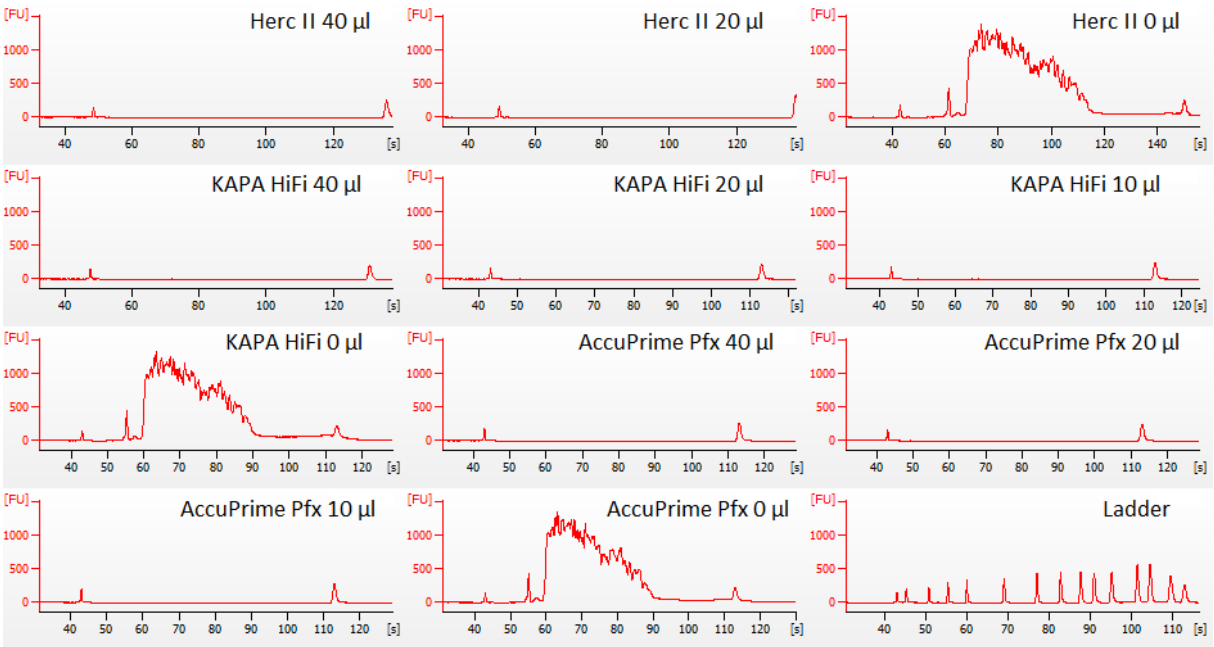

Figure 8. Bioanalyzer electropherograms for the PCR reactions containing NanoLink Streptavidin magnetic beads. The samples marked with Herc II are amplification with Herculase II (Agilent) polymerase, samples marked with KAPA HiFi are amplified with KAPA HiFi Hot Start polymerase (Kapa Biosystems) and samples marked with Accuprime Pfx are amplified with AccuPrime Pfx polymerase (Invitrogen). 40 µl, 20 µl and 0 µl represent the volumes of beads used. Since all beads were in 10 mg/ml solutions the volumes corresponds to 0.4 mg, 0.2 mg and 0 mg beads respectively.

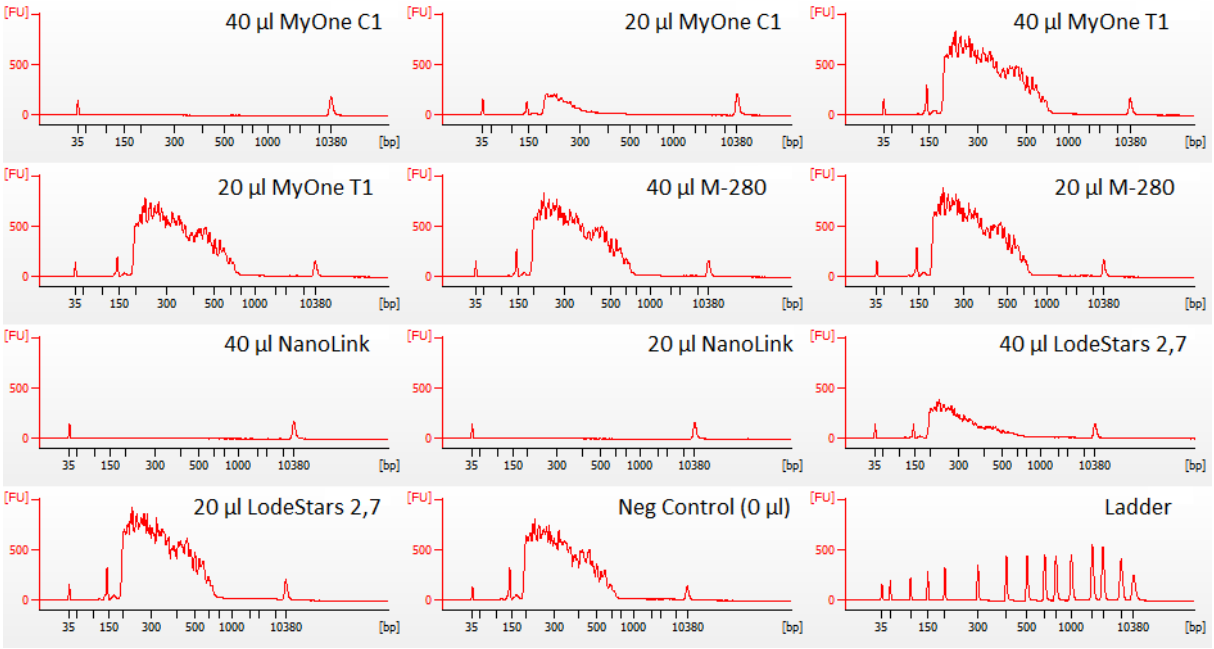

Figure 9. Bioanalyzer electropherograms for the PCR reactions with the present of different streptavidin coated beads. 40 µl, 20 µl and 0 µl represent the volumes of beads used. Since all beads were in 10 mg/ml solutions the volumes corresponds to 0.4 mg, 0.2 mg and 0 mg beads respectively. The beads used were Dynabeads MyOne C1 (Life technologies), Dynabeads MyOne T1, Dynabeads M-280 (Life technologies), NanoLink (Solulink) and Lodestars 2.7 (Agilent). The sample marked 0 µl was the negative control with no beads present.

When PCR was performed in the presence of Dynabeads MyOne T1 and Dynabeads M-280, no significant effect on amplification was observed compared to the negative control; see figure 9 and table 8. LodeStars 2.7 showed some degree of inhibition but did not inhibit the PCR completely. Dynabeads MyOne T1 and Dynabeads M-280 showed strong on-bead PCR capacity and was therefore good candidates for use in the new protocol.

Table 8. Summarized data of the PCR reactions containing streptavidin magnetic beads from different manufacturers.

Beads	Manufacturer	Polymerase	Amount beads	Yield	Inhibition [%]
Nanolink	Solulink	Herculase II fusion	0,4 mg	0 nM	100 %
Nanolink	Solulink	KAPA HiFi Hot Start	0,4 mg	0 nM	100 %
Nanolink	Solulink	AccuPrime Pfx	0,4 mg	0 nM	100 %
Dynabeads MyOne C1	Life technologies	Herculase II fusion	0,4 mg	0.0455 nM	100 %
Dynabeads MyOne C1	Life technologies	Herculase II fusion	0,2 mg	9.56 nM	87.9 %
Dynabeads MyOne T1	Life technologies	Herculase II fusion	0,4 mg	72.9 nM	7.4 %
Dynabeads MyOne T1	Life technologies	Herculase II fusion	0,2 mg	70.8 nM	10 %
Dynabeads M-280	Life technologies	Herculase II fusion	0,4 mg	73.0 nM	7.2 %
Dynabeads M-280	Life technologies	Herculase II fusion	0,2 mg	73.1 nM	7.1%
LoadStars 2.7	Agilent	Herculase II fusion	0,4 mg	27.4 nM	65.2 %
LoadStars 2.7	Agilent	Herculase II fusion	0,2 mg	61.6 nM	21.7 %
Positive Control (no beads)	-	Herculase II fusion	0 mg	78.7 nM	-

PCR on biotinylated fragments bound to streptavidin coated magnetic beads (On-bead PCR):

On-bead PCR was evaluated using Dynabead MyOne T1 and Dynabead M-280 beads as they did not show any significant inhibitory effect when present in PCR. Bioanalyzer electropherograms for On-Bead PCR are shown in figure 10. For the result presented in figure 10 and table 9, Herculase II polymerase and 0.4 mg beads was used in all samples. The effect of adding free biotinylated vector 1 is summarized in table 9. The yield was around 50% lower for samples with added biotinylated vector 1 in comparison with samples without this vector. This is probably due to the fact that added biotinylated vector 1 competes with biotinylated fragment for the binding

to the streptavidin magnetic beads. The beads will therefore be saturated and unbound targeted fragments will be washed away and not captured.

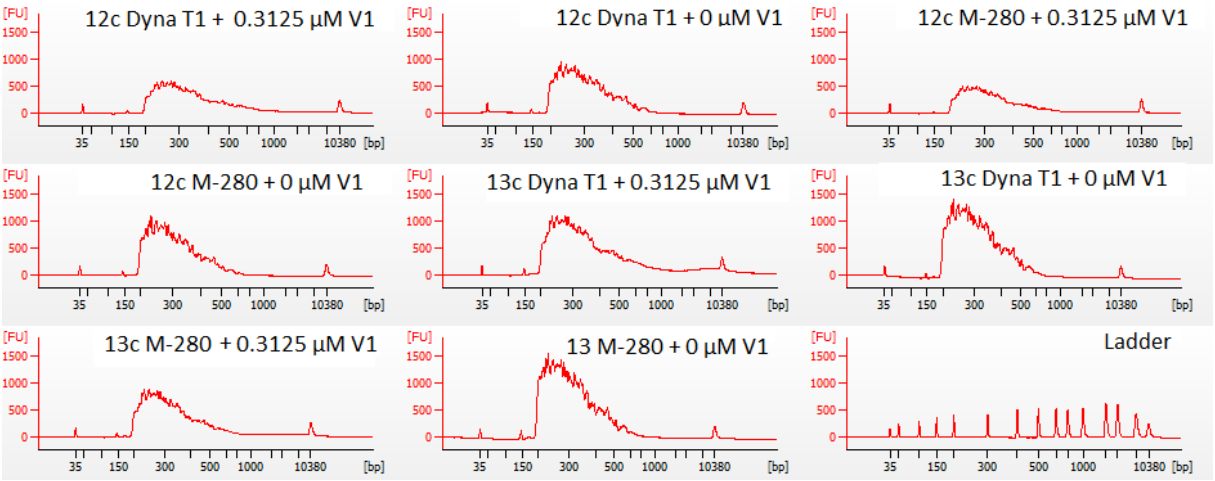

Figure 10. Bioanalyzer electropherograms for the on-bead PCR samples with added vector 1. 0.4 mg Dynabeads MyOne T1 and Dynabeads M-280 was used for the different samples with 0.3125 µM and 0 µM (control) added Vector 1. 12 cycles and 13 cycles were used for the PCR amplification. Herculase II polymerase and 0.4 mg beads was used in all samples.

Table 9. Summerized data for experiment 2

Beads	Cycles	Amount Vector 1	Yield	Yield [%]
Dynabeads MyOne T1	12	0.3125 µM	36.6 nM	57.0 %
Dynabeads MyOne T1	12	0 µM	64,2 nM	-
Dynabeads M-280	12	0.3125 µM	29.5 nM	44.3 %
Dynabeads M-280	12	0 µM	66,6 nM	-
Dynabeads MyOne T1	13	0.3125 µM	36.5 nM	42.9 %
Dynabeads MyOne T1	13	0 µM	85.1 nM	-
Dynabeads M-280	13	0.3125 µM	40.0 nM	48.5 %
Dynabeads M-280	13	0 µM	82.5 nM	-

a **The Yield is in percentage against the proportion of the sample without vector 1**

Experiment 3: Streptavidin magnetic beads binding capacity evaluation

The binding capacities for NanoLink, Dynabeads MyOne T1, DynaBeads MyOne C1 and Dynabeads M-280 were studied and are shown in figure 11 and table 10. An estimation of the beads binding capacities can be made by comparing the intensity of the standard curve (lanes 1-6, figure 11) bands with product generated from the supernatants obtained after binding of probes to beads (lanes 7-11, figure 11). NanoLink showed clearly the best binding capacity of the four tested beads. For NanoLink beads (figure 11, A), wells 7-10 (40µl-5µl) shows a band corresponding to 0.625 pM in the standard curve (figure 11, A, well 5). The experimentally estimated binding capacities are presented in table 10 together with the binding capacities given by the manufacturers. The manufacturers' values are based on 23 nucleotide long ssDNA while the experimental are based on 145 bp long probes which could explain the quite large difference

Despite NanoLink beads having the best binding capacity, On-bead capability was not possible and thus NanoLink beads were not used for the new protocol. The two bead types that showed On-bead PCR compatibility was Dynabeads MyOne T1 and Dynabeads M-280. Due to the fact that Dynabeads MyOne T1 had higher binding capacity of the two, it was chosen for use in new protocol. An interesting observation is that a weak band could be observed, with the band intensity corresponding to approximately 0.625 pM, even with the highest amount of beads tested (most obvious in the NanoLink data). This indicated that there were

unbound probes in the collected supernatants despite the fact that the beads were not saturated. The reason is most likely that a small fraction of the probes did not contain biotin and thus binding to beads was not possible.

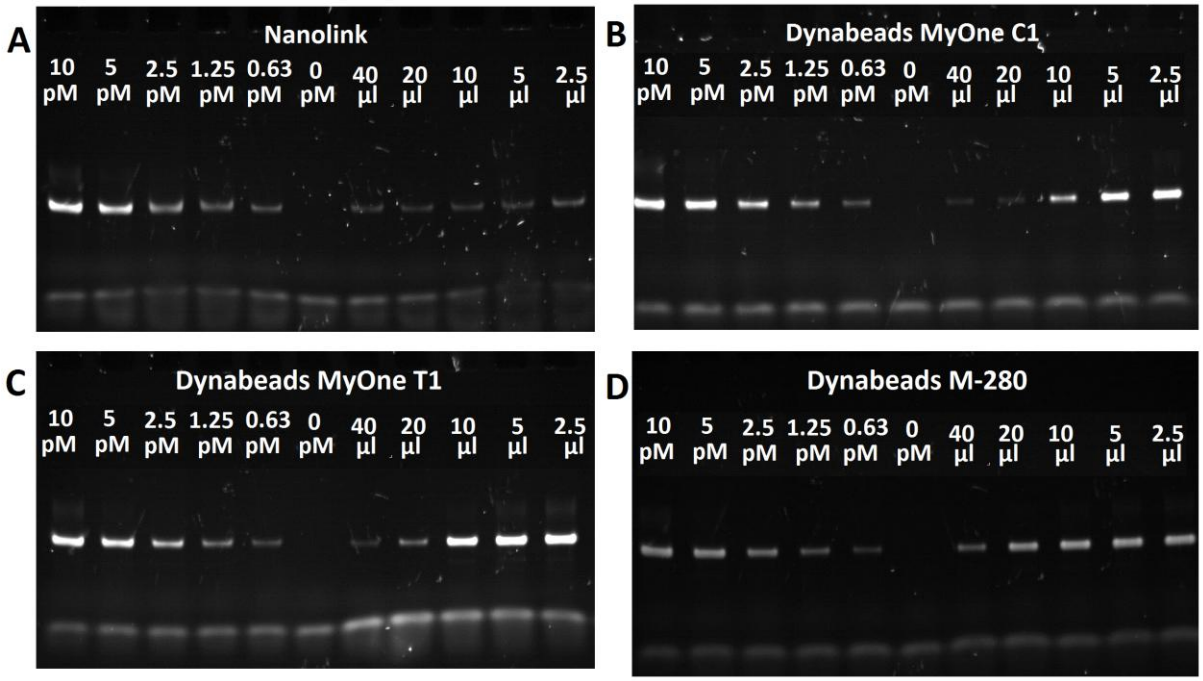

Figure 11. PAGE gel images for the bindig capacity evaluation. Wells 1-6 consisted of the standard curve while wells 7-11 represent the unbound probes collected in supernatant for 40 µl, 20 µl, 10 µl 5 µl and 2.5 µl beads used (10 mg/ml) respectively. A) For Nanolink streptavidin magnetic beads, B) For Dynabeads MyOne C1 streptavidin magnetic beads, C) For Dynabeads MyOne T1 streptavidin magnetic beads, D) For Dynabeads M-280 streptavidin magnetic beads.

Table 10. Summarized data for the binding capacity evaluation

Beads	Manufacturer	Teor. Binding capacity	Exp. Binding capacity
Nanolink	Solulink	2.5 nmol/mg	0.4 nmol/mg
Dynabeads MyOne T1	Life Technologies	0.4 nmol/mg	0.1 nmol/mg
Dynabeads MyOne C1	Life Technologies	0.5 nmol/mg	0.05 nmol/mg
Dynabeads M-280	Life Technologies	0.2 nmol/mg	-

Experiment 4: Ligation in hybridization reaction

In this experiment, we wanted to test if ligation could be performed during hybridization for the different buffers. We also wanted to find the optimal temperature for the reaction. The optimal temperature would be the one with most product yield and with as low artifact as possible.

Hybridization in PCR based buffer (buffer A) and Ampligase based buffer (buffer B):

Bioanalyzer electropherograms for HaloPlex target enrichment for different hybridization temperatures performed in Buffer A and in Buffer B along with standard buffer as control are shown in figure 12 and 13. Differences in yield were observed between the buffers used, but no clear pattern could be observed, see figure 14. Hybridization in higher temperature resulted in lower product yield while lower temperature resulted in higher 185bp peaks. This peak has previously been shown to be non-specific hybridization and amplification of Alu repeats, a region highly abundant in the human genome. Product yield in nM for each

sample is summarized in figure 14. For higher temperatures more specific hybridization was obtained, while for a lower hybridization temperature more non-specific binding was obtained.

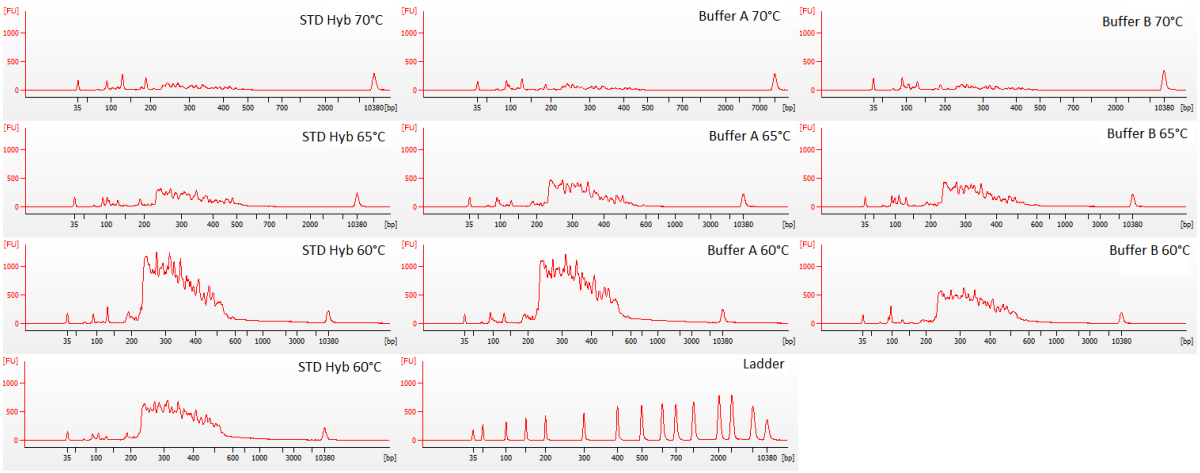

Figure 12. Bioanalyzer electropherograms for hybridization in Buffer A, Buffer B and in standard HaloPlex Hybridization solution. The hybridization was performed in different temperatures between 45 °C to 70 °C, all not shown in this figure. The rest of the samples are shown in figure 13.

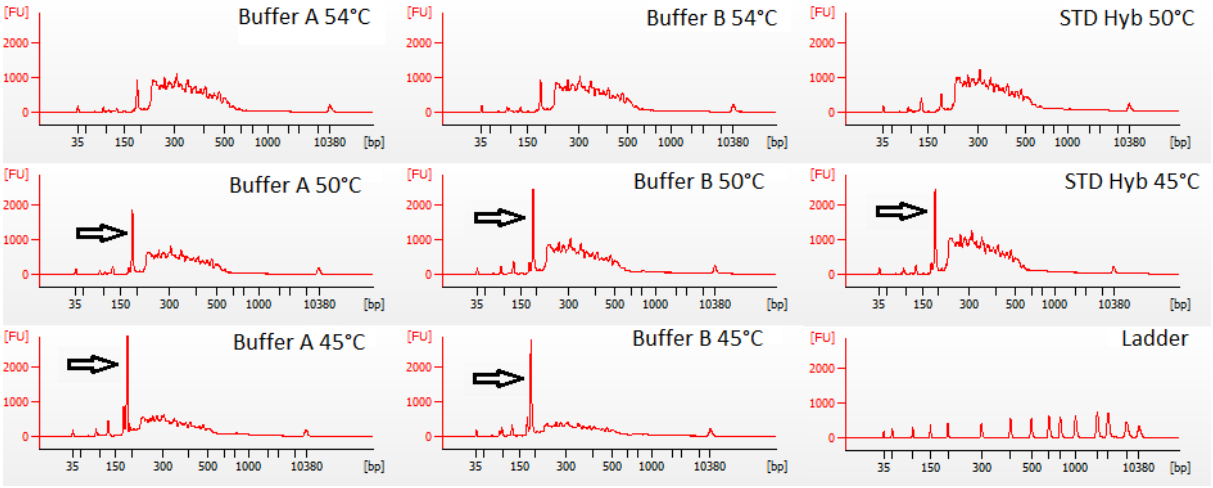

Figure 13. . Bioanalyzer electropherograms for hybridization in Buffer A, Buffer B and in standard HaloPlex Hybridization solution. The hybridization was performed in different temperatures between 45 °C to 70 °C, all not shown in this figure. The rest of the samples are shown in figure 12.

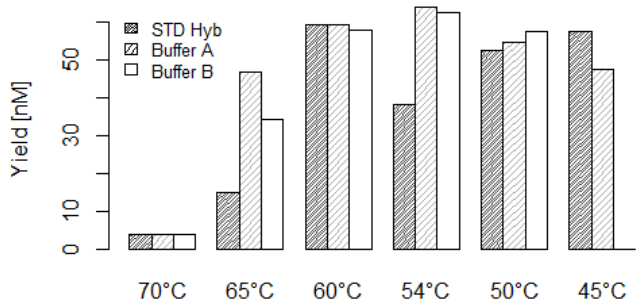

Figure 14. Bar chart of the target enrichment product yield obtained from bioanalyzer electropherograms. The product Yield in nM for the different samples shown in figure 12 and figure 13.

Ligation in hybridization reaction: Bioanalyzer electropherograms for the ligation in hybridization reactions (using the AABL probe library) are shown in figure 15. Both duplicates for 60°C Buffer A resulted in HaloPlex product of expected sizes and with sufficient yield. For the Buffer B reactions the product does not have the expected sizes which could be a sign of non-specific ligation and subsequent amplification. Buffer A using 60°C incubation was therefore chosen as the best candidate to continue development.

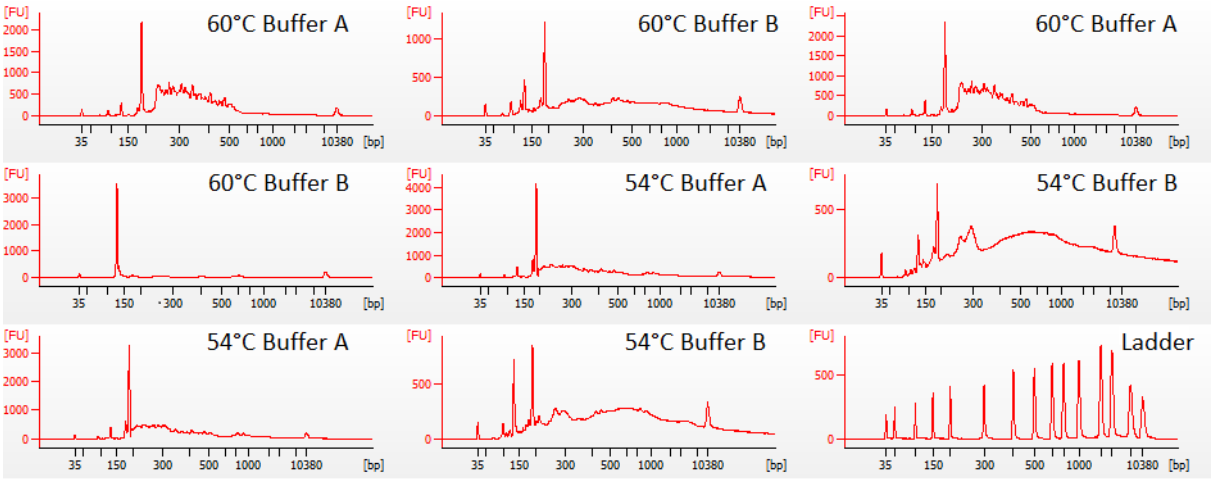

Figure 15. Bioanalyzer electropherograms for the ligation in hybridization reactions. The Hybridization temperatures that were used were 60 °C and 54 °C in Buffer A and in Buffer B. Note the different scales in y-axis.

Bioanalyzer electropherograms for ligation in hybridization reactions for different HaloPlex probe library designs in buffer A using 60 °C incubation temperature are shown in figure 16. The ligation in hybridization reaction resulted in product for each of the tested probe library design with sufficient yield. Unspecific binding (represented by the 185 bp peak) were found in all probe library designs.

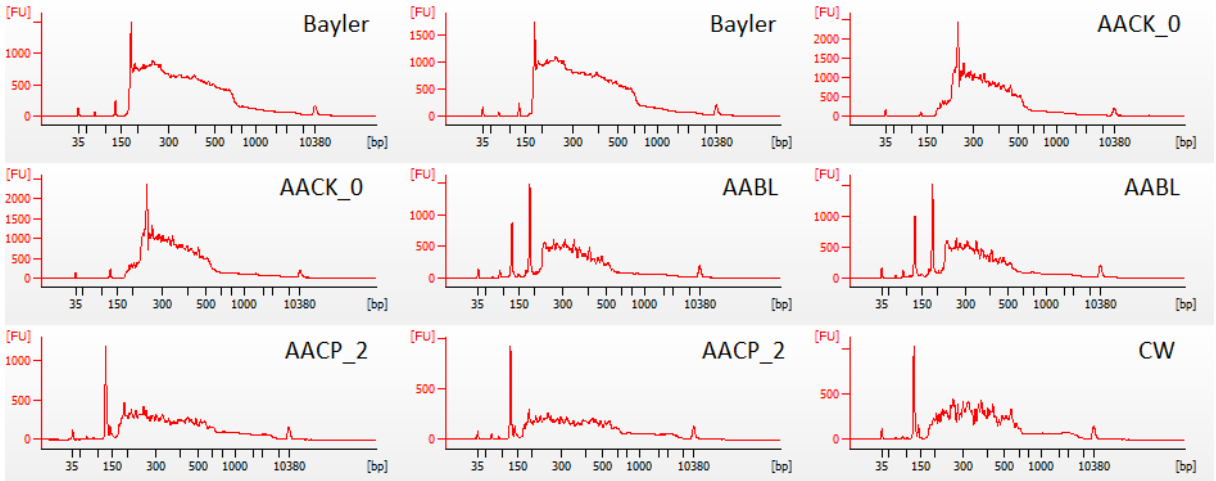

Figure 16. Bioanalyzer electropherograms for the ligation in hybridization reactions in Buffer A for different probe designs. The Hybridization temperature was 60 °C and the probe designs used was: Bayler, AACK-0, AABL, AACP-2 and CW. Note the different scales in y-axis.

Experiment 5: Ligation in hybridization reaction with formamide.

Bioanalyzer electropherograms for ligation in hybridization reactions (AABL probe library) with added formamide in different concentrations are shown in figure 17. Non-specific binding could previously been reduced by adding formamide to the hybridization buffer. Formamide lowers the melting point of DNA and is commonly used as denaturing agent for DNA (FUCHS *et al.* 2010; KE and WARTELL 1996). The amount of the 185bp artifact decreases together with the total product yield with increasing formamide concentrations (figure 18, A and B). The ratio artifact/yield versus the concentration of formamide in the buffer is shown in plot C in figure 18. Bioanalyzer electropherograms shows that the 185 bp artifacts was completely gone at higher concentrations than 2.5% formamide suggesting that a formamide concentration between 2.5 and 5% is sufficient for the elimination of unspecific binding. Experiments with formamide in hybridization shows positive result but are still not robust. Further attempts should be made to find the optimal concentration of formamide allowing as much yield and as little as possible artifact thus obtaining a low artifact/yield ratio.

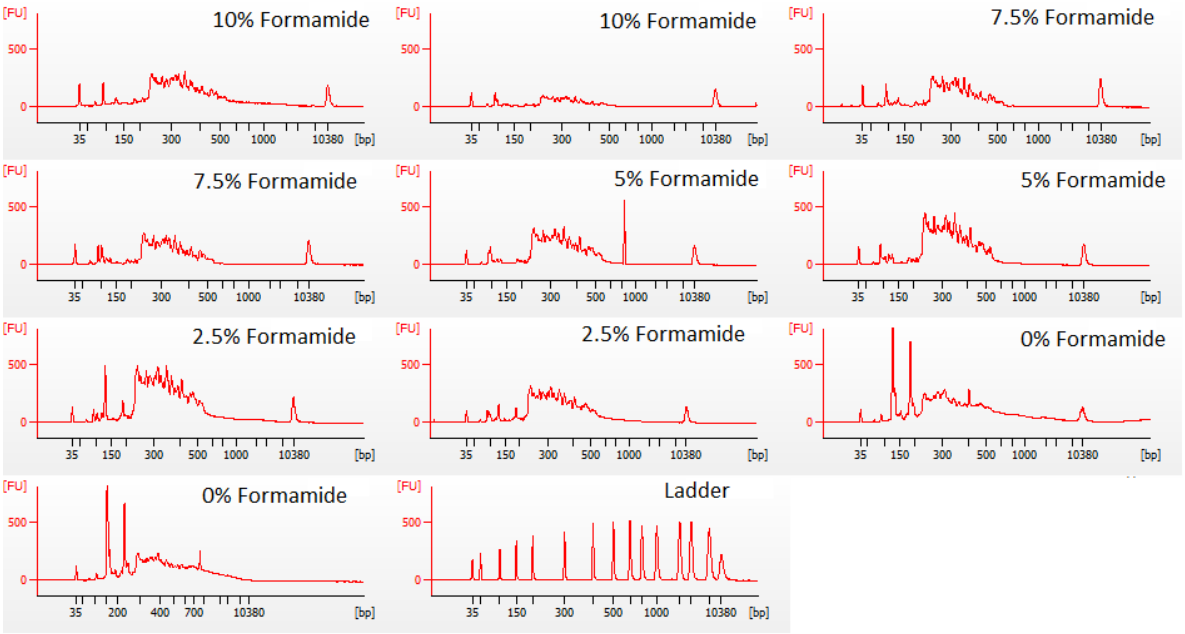

Figure 17. Bioanalyzer electropherograms for the ligation in hybridization reactions in Buffer A with added formamide. The hybridization temperature was 60 °C and the probe library design that was used was AABL.

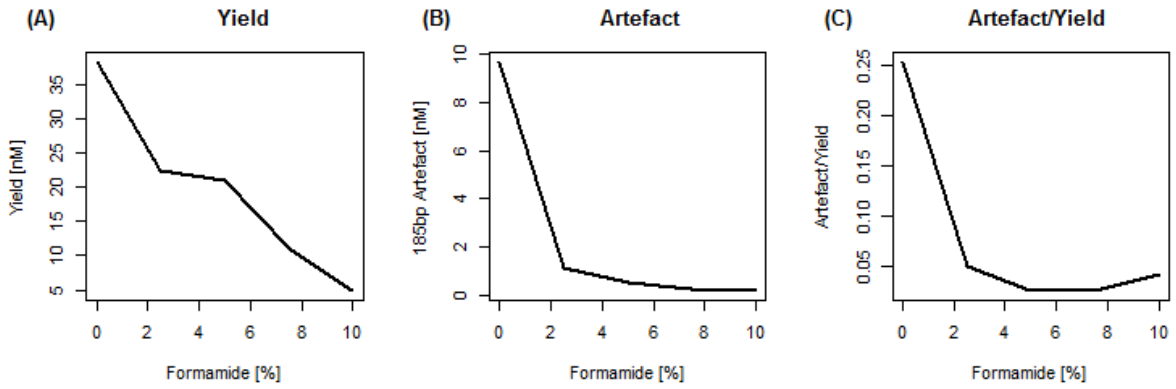

Figure 18. Line chart showing the change in product yield and artifact for different formamide concentrations in hybridization reactions A) Percentage fomamide concentration in reaction versus product yield in nM, B) Percentage formamide concentration versus 185 bp artifact in nM, C) Percentage formamide concentration versus the ratio artifact/yield.

HaloPlex 2.0 Alpha Target enrichment proof of concept

Bioanalyzer electropherograms for HaloPlex 2.0 target enrichment with different probe and vector dilutions are shown in figure 19. Concentrations for the probes and vectors are shown in table 11. Samples with the probe concentration 10 pM/probe resulted in no HaloPlex product in the expected size range of 175-625bp while 0.5 pM and 0.125 pM samples resulted in the expected trace profile and with sufficient product yield for sequencing. Two samples, one 0.5 pM and one 0.125 pM, were pooled and sequenced on the MiSeq using the dual index sequencing workflow. After demultiplexing the reads per sample barcode and aligning to the genome, the standard enrichment performance metrics such as uniformity of read depth and specificity was calculated (table 11). Comparing these data to the same HaloPlex panel enriched with the standard protocol indicate no significant difference in performance (table 11). The new protocol performs at least as good as the previous one in terms of specificity, uniformity and coverage. The large peak seen at 300 bp in the sample 2 did not affect the specificity or uniformity of the enrichment.

Table 11. Concentrations for probes and vectors for the different samples.

Sample	Probe and vector dilutions	Probe conc.	Vector conc.	Mol. Barcodes	Total reads
1	1X	10 pM/probe	20 µM	-	-
2	20X	0.5 pM/probe	1 µM	974 049	6 396 796
3	80X	0.125 pM/probe	0.25 µM	584 526	4 126 548

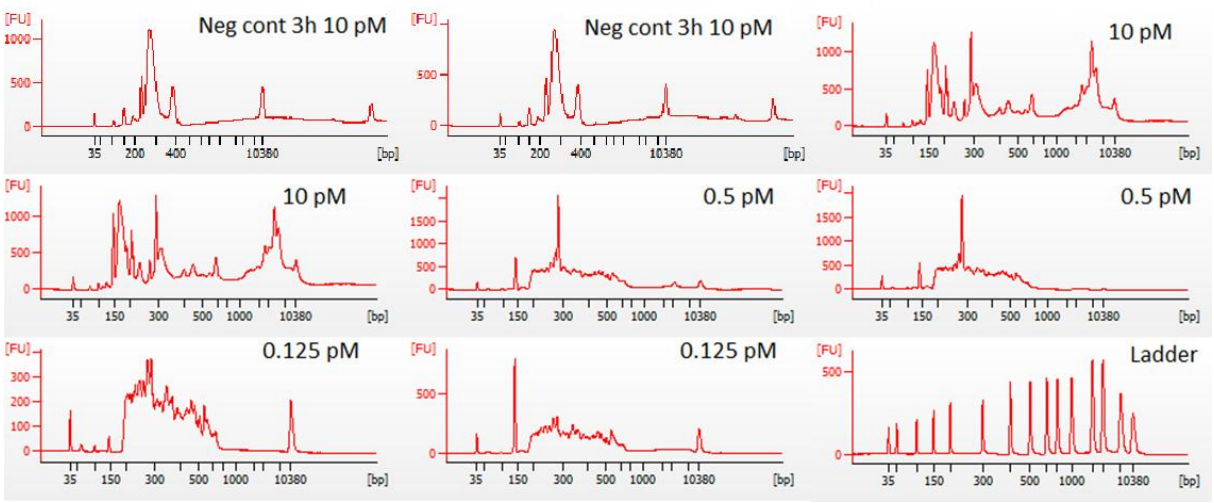

Figure 19. Bioanalyzer electropherograms for the HaloPlex 2.0 target enrichment for different probe and vector dilutions. The negative controls with no DNA were incubated in three hours for the hybridization and ligation reactions, while the other samples were incubated over night. Note the different scales in y-axis.

Table 12. Summary for the sequencing result of the new protocol in comparison with the current one.

Sample	Raw reads	Aligned to Hg19	Specificity	Depth in Covered Region	Depth in Target Region	Coverage at 0.1x of average	Coverage at 0.2x of average	Coverage >=1x	Coverage >=10x	Coverage >=20x	Coverage >=100x
Standard protocol	619493	99%	99,2%	207	347	96.5%	93.3%	99.6%	98.0%	96.7%	83.9%
20X sample	7705819	98%	98.0%	8760.67	17648.7	98.7%	97.5%	100.0%	100.0%	100.0%	98.97%
80X sample	4202999	99%	98.9%	5845.39	11802.8	98,4%	97.2%	100.0%	100.0%	100.0%	99.82

Analysis of Molecular Barcode data:

To analyze the molecular barcode information, custom software developed by Agilent informatics department was used to associate all reads with one target fragment before the number of unique molecular barcodes was counted, giving the total number of reads per amplicon as well as the number of unique molecules observed. In figure 21, the number of reads per amplicon (x-axis) is plotted against the number of unique molecules per

amplicon (y-axis). One can observe that the average y/x ratio for each amplicon was 0.75 and 0.26 for the 20x and 80x sample respectively. This means that for the 20x sample an average of 75% of the reads had unique barcodes whereas 26% were unique for the 80x sample.

The frequencies of the distribution of the molecular barcodes are shown in figure 20. The total number of barcodes found was 974 049 for the 20x sample and 584 526 for the 80x sample. The maximum number of theoretically possible molecular barcodes is 4^{10}=1 048 576.

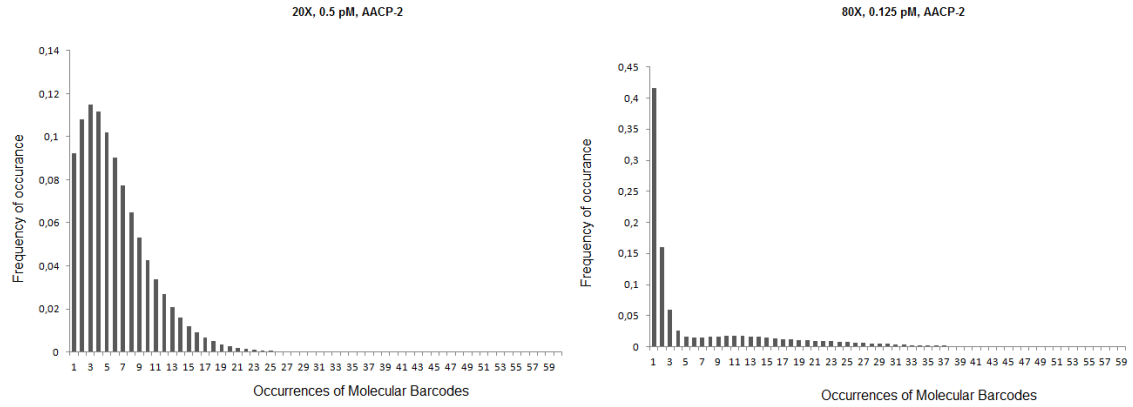

Figure 20. The frequencies for the distribution of the molecular barcodes. (A) sample 2 with 20x dilution (B) sample 3 with 80x dilution.

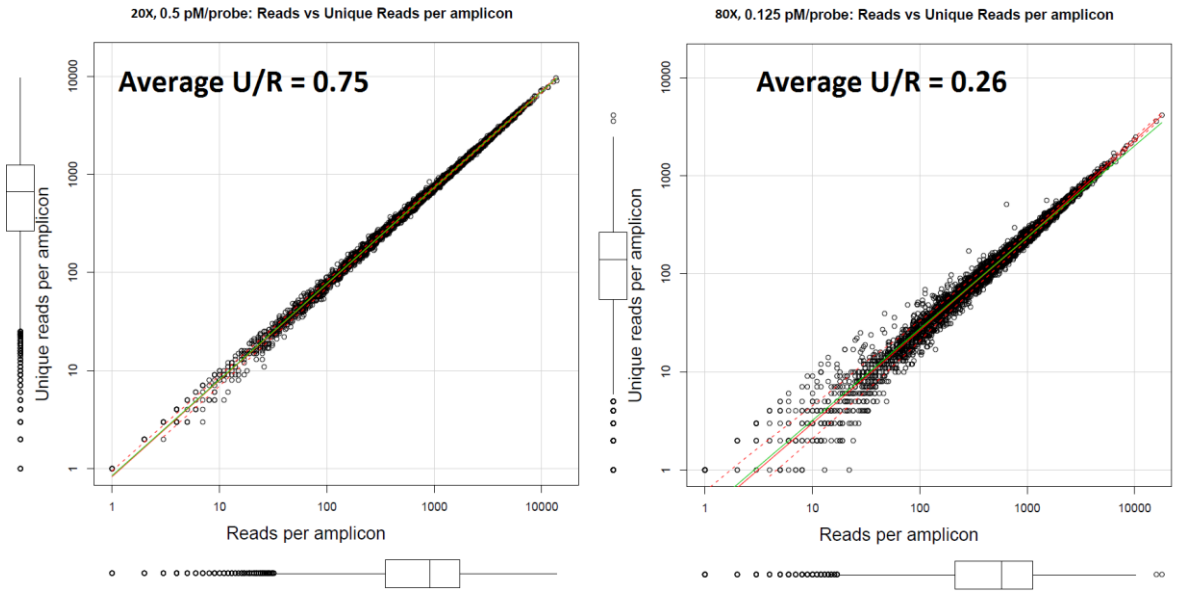

Figure 21. Number of reads with unique barcode in y-axis versus total reads per amplicon in x-axis for each amplicon. Axes are in logarithmic scale.

200 ng of input DNA split into eight restriction digestion reactions result in approximately 7500 available molecules of each restriction fragment. The fraction of these 7500 molecules that is captured by the HaloPlex procedure can be used as a quality measurement of the enrichment technology. In standard HaloPlex the captured fragments are PCR amplified to millions of copies of each molecule. In the subsequent sequencing step, copies are sampled (reads) from this pool and there is no way of knowing if two reads are from two unique molecules or two copies of one molecule. With the additional information the molecular barcode brings, this distinction can now be done. In the following section the statistics of sampling reads from a pool of

molecules will be discussed to show how the number of unique molecules available after PCR can be predicted directly from the ratio between reads and observed unique molecules.

For HaloPlex, captured fragments are PCR amplified to millions of copies of each molecule. Having a PCR pool with a limited number of unique molecules but with millions of copies of each molecule, sequencing of PCR duplicates can occur. The probability for this happening will increase with the amount of sequenced fragments and decrease if the number of unique molecules increases. As the number of samplings (reads) is much lower than the number of available copies to sample from, the probability of sampling a copy of a certain unique molecule does not change during the course of sampling. Because of this one can assume that there are an infinite number of copies for each molecule, which makes it possible to relate the sequencing of PCR fragments with a phenomenon in combinatorics and probability theory called "sampling with replacement". Additionally the sequencing of fragments can also be related the "coupons collectors' problem" in probability theory, that states: Suppose that there are n different coupons, equally likely, from which coupons are being collected with replacement. What is the probability that more than t sample trials are needed to collect all n coupons? An alternative statement would be: Given n coupons, how many coupons do you expect you need to draw with replacement before having drawn each coupon at least once? Relating this to our case, having sequenced R reads from PCR population of N unique molecules with replacement. Recording the number of reads R and the number of observed unique molecules, U, how can we with this information estimate the number of available molecules N in PRC population? In other words, having the number of molecules sequenced, and the total number of sequenced fragments (reads), we want to estimate the number of available unique molecules in the PCR population.

To find the relationship between R, N and U computer simulations were made using the open source software "R" (TEAM 2013). In addition, an approximation method was used to confirm the results from the simulations. The simulations were made in three steps:

- Sample R reads from PCR population of N unique molecules with replacement
- Do this for varying R and N using 100 independent simulations per value.
- Record the average number of unique observed molecules (U) for every combination of R and N

In Figure 22, the fraction of the number of observed unique molecules (U) per available molecules (N) is plotted in function of the number of unique molecules (U) per number of reads (R). With a given U and R, this plot can be used to estimate N (number of available molecules after capture). As calculated above (figure 21), the average value of the quotient "Observed unique molecules/Total read" per amplicon is 0.75 for the 20X sample and 0.26 for the 80X sample. By looking up these values in the x-axis in figure 22(or table 13 in appendix 4) and relating them to the corresponding y-value, 0.455 and 0.977 are obtained. This means that we have observed 45.5% and 97.7% of all available molecules in the PCR population for the 20x sample and 80x respectively. By dividing the number of observed unique molecules by this value (unique molecules / available molecules), an estimate of the number of unique available molecules in the PCR pool (N) can be obtained. Then, an approximate average capture efficiency can be calculated by dividing N by 7500, which is the expected number of unique molecules in 200 ng DNA split in eight reactions. The average capture efficiency for the 20X sample was 23% and for the 80X sample it was 5.3% (figure 23). The estimation was based on an average value and is not true for all molecules, but as there always will be a distribution around the mean and the number of points (the number of amplicons) is high the average value can be used for the comparison of two samples .

This estimation of the capture efficiency can be used as a new metric for the performance of the enrichment, in combination with current metrics. By only looking at the sequencing data presented in table 12, one might

think that enrichment in sample 2 and in samples 3 are of the same quality as the specificity, uniformity and coverage are the same. But by comparing the capture efficiency, it is evident that more information about the sample can be obtained by using the sample 2 protocol compared to the sample 3 protocol as more unique molecules can be sequenced. The possibility of interrogating more molecules means that rare variants in the sample can be detected more easily. This new metric of capture efficiency gives new opportunities for better optimisation of the HaloPlex target enrichment protocol.

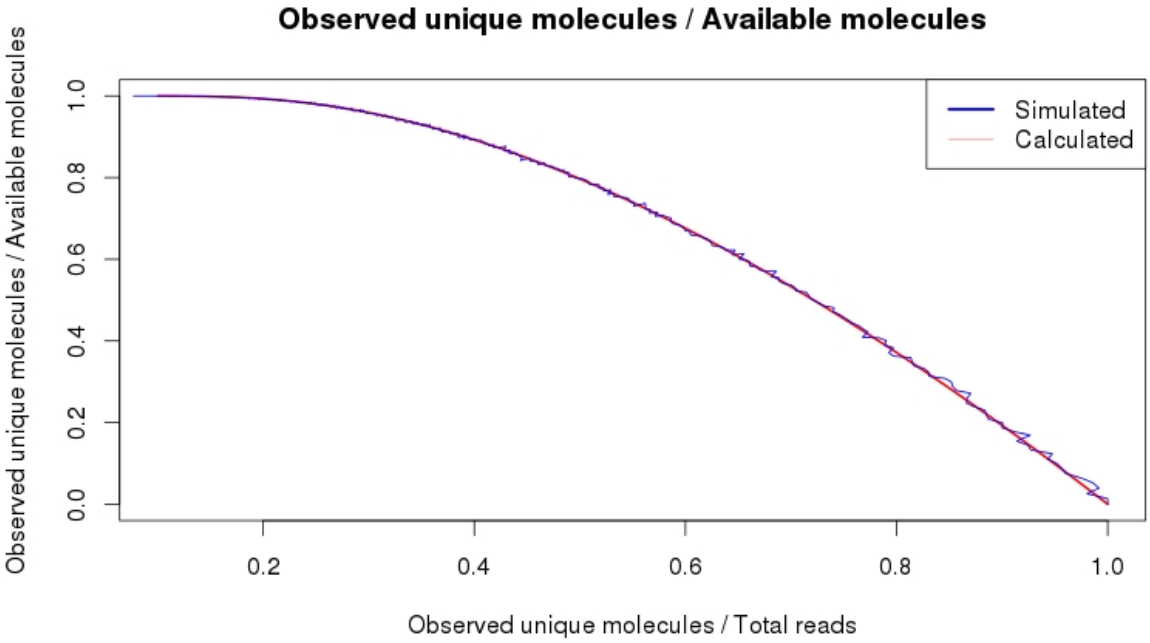

Figure 22. Simulations and calculations for "observed unique molecules/Available molecules". Simulations and calculations to relate the ratio of "observed unique molecules/total reads" to the ratio of "unique observed molecules/available fragments". Using this information, one can estimate how many different molecules there are in the population.

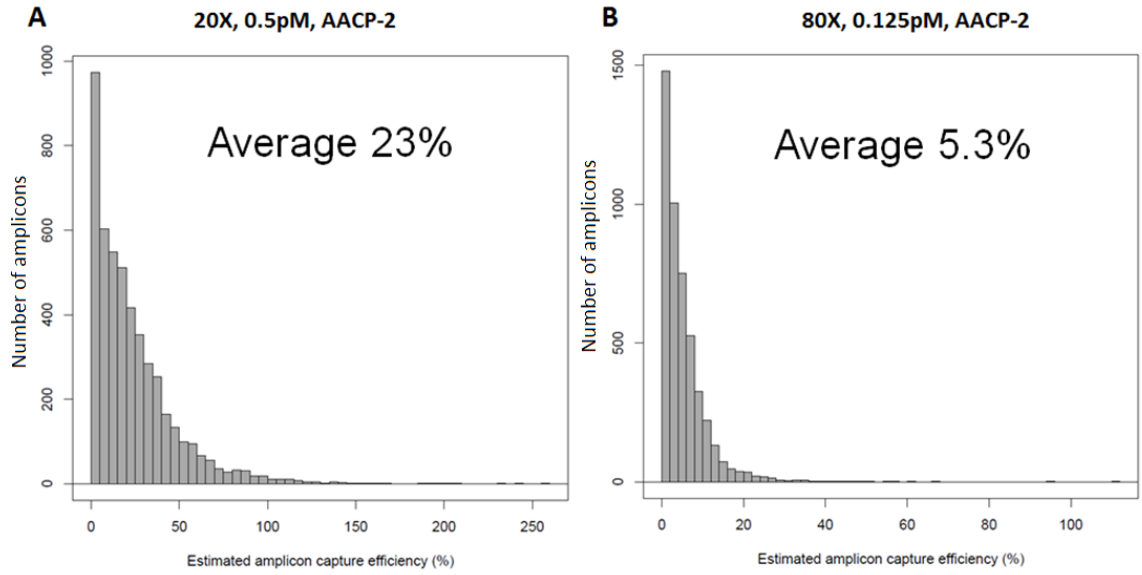

Figure 23. The distribution of the estimated capture efficiency for the different amplicons. (A) Sample 1 with 20x dilutions and 0.5pM probe concentration, (B) Sample 1 with 80x dilutions and 0.125pM probe concentration. The average value for the capture efficiency is shown for each sample

Future applications

In addition to giving better quality control (as described above), implementation of molecular barcodes can bring other advantages and improvements to the HaloPlex enrichment. When different molecules of the same amplicon are amplified, the molecules can be amplified with different efficiency leading to incorrect allele frequencies. The problem with skewed allele frequencies will be greater for lower DNA input as the effect of statistical sampling will have more impact. Using molecular barcodes, one can count the number of unique molecules observed with each allele and thus obtain a more accurate representation of the true allele frequency. Figure 24 is a very simplified illustration showing how one can obtain more accurate representation of allele frequencies with the help of molecular barcodes. A more detailed illustration is presented in Appendix 6.

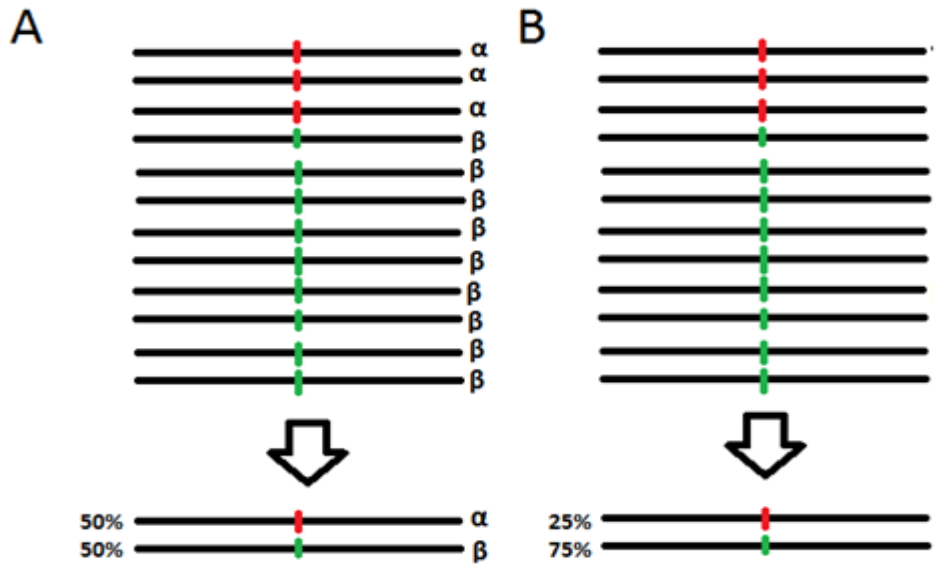

Figure 24. Illustration for the correction of skewed allele frequencies with the help of molecular barcodes. A) With molecular barcodes, B) Without molecular barcodes. The black lines represents reads of the same sequence. The Greek letters (α and β) represent molecular barcodes. The green and red marks represent two different types of nucleotides.

Another possible application of molecular barcodes is the detection of rare mutations where often the problem is the formation of false positives mutations (mainly polymerase errors) during the enrichment and sequencing steps. If a variant is discovered in next-generation sequencing data it can be difficult to say whether it is due to an actual mutation in the genome or whether it is due to a technical mutation introduced during library preparation or sequencing. Through the use of molecular barcodes, one can examine if the variation can be seen in other molecules or if they only occurred in duplicates of the same molecule. In the latter case, it may be because there has been a mutation early in the amplification step during library preparation or during the clustering step for Illumina sequencing. Finding the same variation in two different molecules is very unlikely when two random errors must occur in exactly the same position in two different molecules. In figure 31 in Appendix 6 an illustration is presented showing how technical mutations can be eliminated using molecular barcodes.

Acknowledgements

First of all I would like to thank my supervisor Fredrik Roos for his excellent supervision and for inspiring me during the project. Secondly, I would like to thank Marcus Danielsson for sharing his knowledge and helping me during the project. Additionally I would like to thank everybody else at Agilent Technologies for their help and also the bioinformatics team, for helping with the bioinformatic part of the project. Also a big thanks to my scientific reviewer Lotte Moens and examiner Lars-Göran Josefsson for their excellent feedback on the report.

References

ADLER, M., P. VAN MOERBEKE and P. VANHAECKE, 2008 Singularity confinement for a class of m-th order difference equations of combinatorics. Philos Trans A Math Phys Eng Sci **366:** 877-922.

ALBERT, F. W., E. HODGES, J. D. JENSEN, F. BESNIER, Z. XUAN *et al.*, 2011 Targeted resequencing of a genomic region influencing tameness and aggression reveals multiple signals of positive selection. Heredity (Edinb) **107:** 205-214.

ANSORGE, W. J., 2009 Next-generation DNA sequencing techniques. N Biotechnol **25:** 195-203.

CASBON, J. A., R. J. OSBORNE, S. BRENNER and C. P. LICHTENSTEIN, 2011 A method for counting PCR template molecules with application to next-generation sequencing. Nucleic Acids Res **39:** e81.

FU, G. K., J. HU, P. H. WANG and S. P. FODOR, 2011 Counting individual DNA molecules by the stochastic attachment of diverse labels. Proc Natl Acad Sci U S A **108:** 9026-9031.

FUCHS, J., D. DELL'ATTI, A. BUHOT, R. CALEMCZUK, M. MASCINI *et al.*, 2010 Effects of formamide on the thermal stability of DNA duplexes on biochips. Anal Biochem **397:** 132-134.

GLENN, T. C., 2011 Field guide to next-generation DNA sequencers. Mol Ecol Resour **11:** 759-769.

GORT, G., W. J. KOOPMAN and A. STEIN, 2006 Fragment length distributions and collision probabilities for AFLP markers. Biometrics **62:** 1107-1115.

HATTORI, M., 2005 [Finishing the euchromatic sequence of the human genome]. Tanpakushitsu Kakusan Koso **50:** 162-168.

HODGES, E., Z. XUAN, V. BALIJA, M. KRAMER, M. N. MOLLA *et al.*, 2007 Genome-wide in situ exon capture for selective resequencing. Nat Genet **39:** 1522-1527.

HOLLELEY, C. E., and P. G. GEERTS, 2009 Multiplex Manager 1.0: a cross-platform computer program that plans and optimizes multiplex PCR. Biotechniques **46:** 511-517.

JOHANSSON, H., M. ISAKSSON, E. F. SORQVIST, F. ROOS, J. STENBERG *et al.*, 2011 Targeted resequencing of candidate genes using selector probes. Nucleic Acids Res **39:** e8.

KE, S. H., and R. M. WARTELL, 1996 The thermal stability of DNA fragments with tandem mismatches at a d(CXYG).d(CY'X'G) site. Nucleic Acids Res **24:** 707-712.

KLAMKIN, M. S., and D. J. NEWMAN, 1967 Extensions of the birthday surprise. Journal of Combinatorial Theory **3:** 279-282.

LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860-921.

LI, J., R. LUPAT, K. C. AMARASINGHE, E. R. THOMPSON, M. A. DOYLE *et al.*, 2012 CONTRA: copy number analysis for targeted resequencing. Bioinformatics **28:** 1307-1313.

MAINLAND, J. D., J. R. WILLER, H. MATSUNAMI and N. KATSANIS, 2013 Next-generation sequencing of the human olfactory receptors. Methods Mol Biol **1003:** 133-147.

MAMANOVA, L., A. J. COFFEY, C. E. SCOTT, I. KOZAREWA, E. H. TURNER *et al.*, 2010 Target-enrichment strategies for next-generation sequencing. Nat Methods **7:** 111-118.

MARGULIES, M., M. EGHOLM, W. E. ALTMAN, S. ATTIYA, J. S. BADER *et al.*, 2005 Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:** 376-380.

MEUZELAAR, L. S., O. LANCASTER, J. P. PASCHE, G. KOPAL and A. J. BROOKES, 2007 MegaPlex PCR: a strategy for multiplex amplification. Nat Methods **4:** 835-837.

MINER, B. E., R. J. STOGER, A. F. BURDEN, C. D. LAIRD and R. S. HANSEN, 2004 Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. Nucleic Acids Res **32:** e135.

NAUS, J., 1974 Probabilities for a Generalized Birthday Problem. Journal of the American Statistical Association **69:** 810-815.

SANGER, F., S. NICKLEN and A. R. COULSON, 1977 DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A **74:** 5463-5467.

SAPERSTEIN, B., 1972 The Generalized Birthday Problem. Journal of the American Statistical Association **67:** 425-428.

SHENDURE, J., 2011 Next-generation human genetics. Genome Biol **12:** 408.

SHENDURE, J., and H. JI, 2008 Next-generation DNA sequencing. Nat Biotechnol **26:** 1135-1145.

SHENDURE, J., and E. LIEBERMAN AIDEN, 2012 The expanding scope of DNA sequencing. Nat Biotechnol **30:** 1084-1094.

TEAM, R. C., 2013 R: A Language and Environment for Statistical Computing, pp. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.

WAGNER, D., 2002 A Generalized Birthday Problem, pp. 288-304 in *Advances in Cryptology — CRYPTO 2002*, edited by M. YUNG. Springer Berlin Heidelberg.

VASTA, V., S. B. NG, E. H. TURNER, J. SHENDURE and S. H. HAHN, 2009 Next generation sequence analysis for mitochondrial disorders. Genome Med **1:** 100.

VENTER, J. C., M. D. ADAMS, E. W. MYERS, P. W. LI, R. J. MURAL *et al.*, 2001 The sequence of the human genome. Science **291:** 1304-1351.

Index of appendices

Appendix 1

Comparison of the enrichment steps for current protocol with the new protocol.

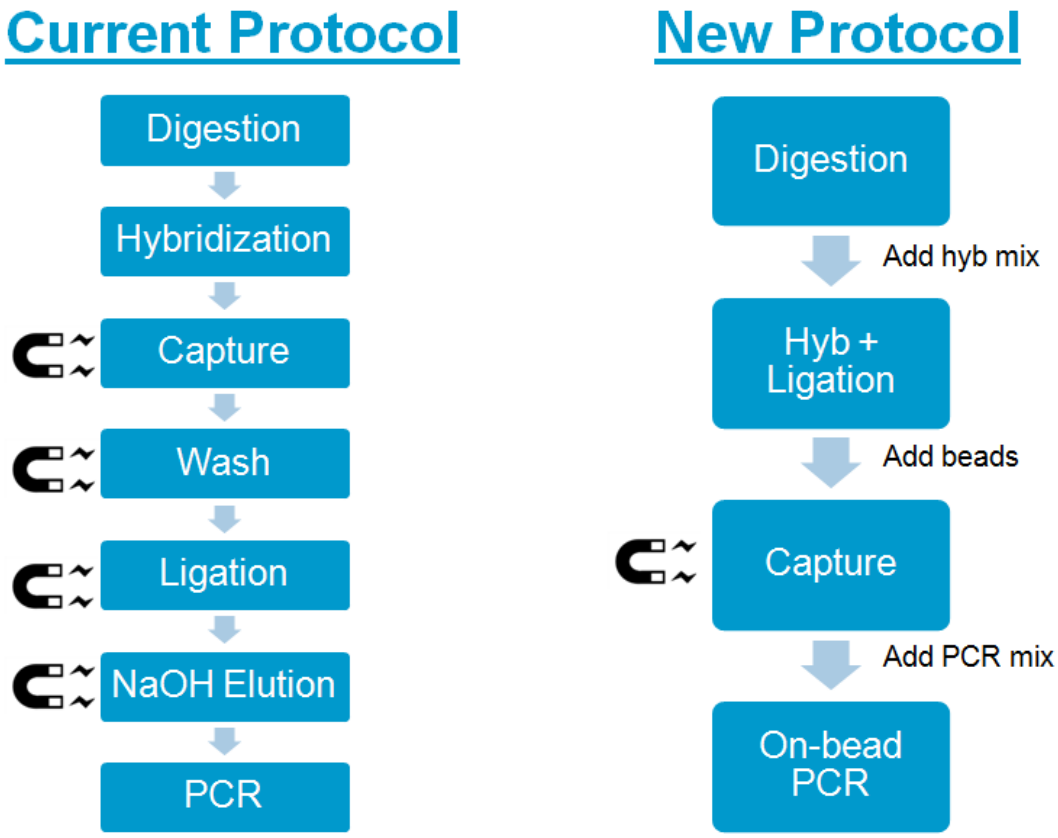

Figure 25. Schematic illustration of the steps for the comparison of current protocol with the new one developed during this project.

Appendix 2

Calculations and simulations for Molecular Barcodes

Theoretical calculations, simulations and plots for the probability that different DNA molecules are assigned the same barcode are presented in figure 26, A. Simulations were made for the frequencies (probabilities) that a certain number of molecules are assigned non-unique/overlapping barcodes, se plot B in figure 26. The probability that all 7500 molecules (200 ng input DNA split into 8 digest reactions) are assigned a unique barcode is 0 and in average 26.8 molecules is not assigned unique barcodes.

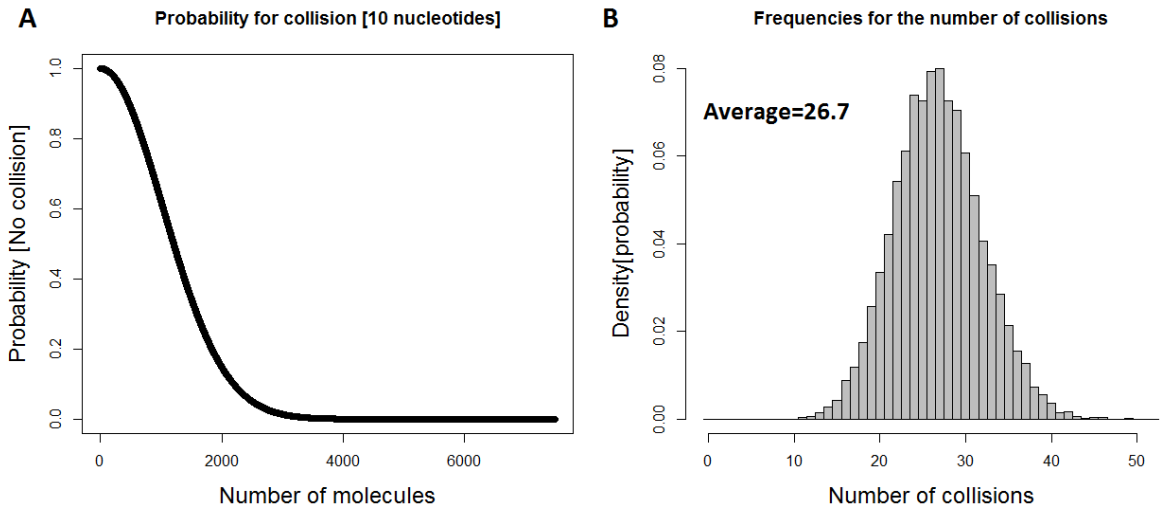

Figure 26. Calculated and simulated probabilities that identical barcodes are assigned to different DNA molecules. The calculations and simulations are done based on a 10 nucleotide long molecular barcode sequence (resulting in 4^{10} unique barcodes). A) Calculated probability of assigning unique molecular barcode to all DNA molecules for different number of molecules. B) Simulated frequencies for the number of molecules with overlapping molecular barcodes, simulations were performed for 7500 molecules which corresponds to 200 ng input DNA.

For longer molecular barcode sequence the probability that all molecules are assigned unique barcodes is increasing and for a 14 nucleotide long barcode sequence this probability is high (higher than 0.85), see figure 27, A. The average number of molecules with overlapping barcode sequence is simulated for different number of molecules, see figure 27, B.

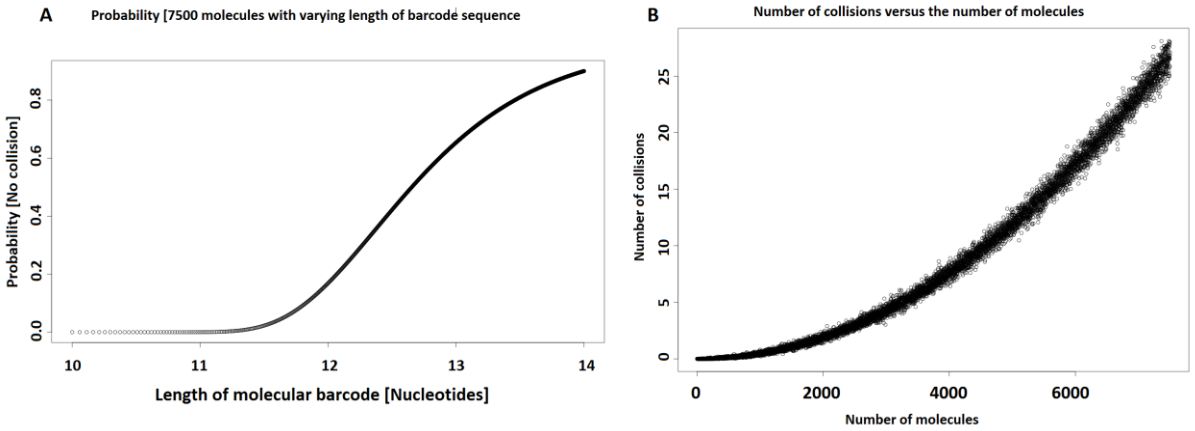

Figure 27. A) Calculated and simulated probabilities that identical barcodes are assigned to different DNA molecules (for 7500 molecules), based on different barcode lengths. B) Simulations for the average number of molecules with overlapping barcodes for different number of molecules for 10 nucleotides long sequence.

Appendix 3

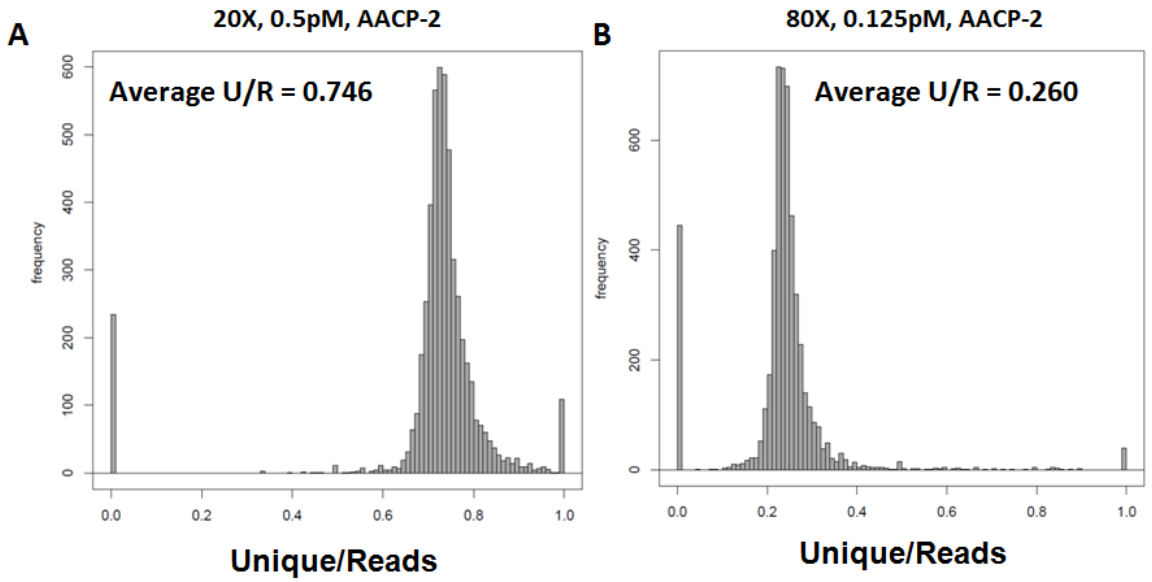

Figure 28. Frequency histograms for the distribution of the quotient "unique molecules/ total reads" (U/R), (A) For sample 2 with 20X dilution, and (B) for sample 3 with 80X dilution. The average value of U/R are shown for each sample.

Appendix 4

Approximation methods were used to relate the ratio of “observed unique molecules / total reads” to the number of unique observed molecules of total available. The table can be used to obtain the y-value corresponding to an x-value.

Table 13. Relationship between the x and y-values of figure 23.

X	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0		1	1	1	1	1	1	1	1	1
0.1	1	0.9999	0.9998	0.9996	0.9993	0.9988	0.9982	0.9974	0.9963	0.995
0.2	0.9935	0.9916	0.989	0.9864	0.9835	0.9804	0.9769	0.973	0.9688	0.2902
0.3	0.9595	0.9542	0.9487	0.9428	0.9368	0.9304	0.9234	0.9165	0.909	0.9012
0.4	0.8932	0.8851	0.8765	0.8672	0.8582	0.849	0.8388	0.8287	0.8188	0.8083
0.5	0.7972	0.7863	0.7749	0.7633	0.7517	0.74	0.7278	0.7151	0.7023	0.6896
0.6	0.6764	0.6633	0.6497	0.6356	0.6216	0.608	0.593	0.5788	0.564	0.5494
0.7	0.5336	0.5181	0.5029	0.4872	0.471	0.4552	0.439	0.4223	0.4062	0.3895
0.8	0.3724	0.3549	0.338	0.3206	0.3028	0.2857	0.267	0.2491	0.2306	0.2126
0.9	0.1941	0.1747	0.1558	0.1368	0.1177	0.0985	0.0792	0.0596	0.0399	0.0201
1	NA									

Appendix 5

Formula for estimating the observed unique molecules (eq.3) and for calculating the variance of the estimated number of observed unique molecules (eq. 4):

(Eq. 3) $$m[U|M,R] = M\left(1-\left(1-\frac{1}{M}\right)^{R}\right)$$

(Eq. 4) $$var[U|M,R] = M\left(1-\frac{1}{M}\right)^{R} + M^2\left(1-\frac{1}{M}\right)\left(1-\frac{2}{M}\right)^{R} - M^2\left(1-\frac{1}{M}\right)^{2R}$$

Where m are the observed unique molecules, M the available molecules in population and R the number of reads.

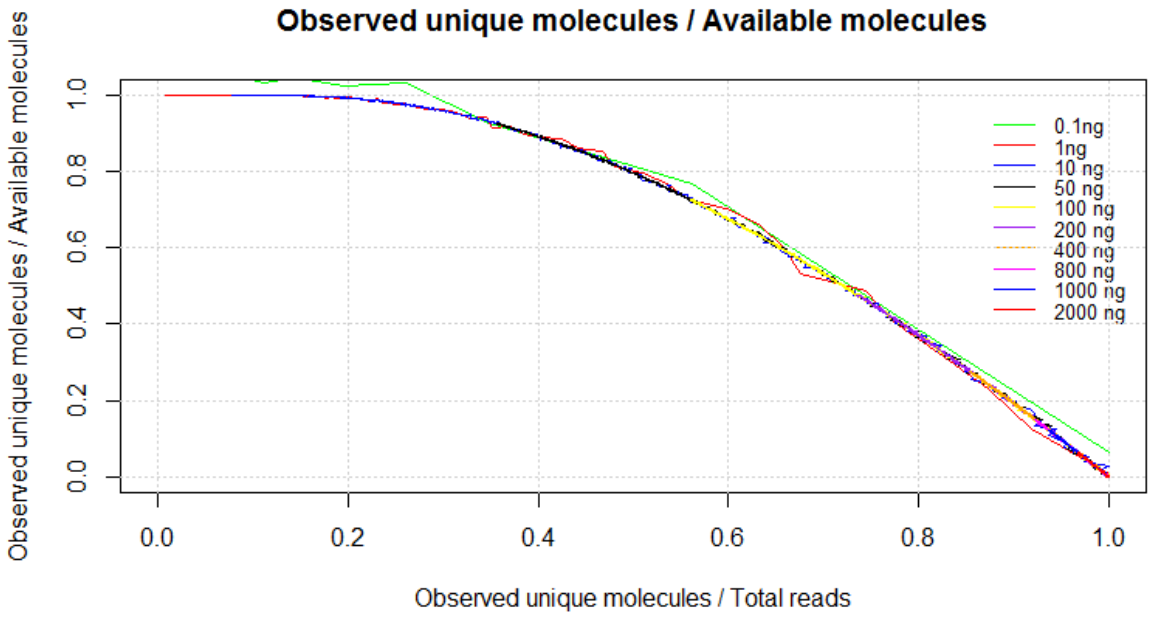

Figure 29. Simulations for observed "unique molecules/Available molecules". Simulations to relate the ratio of "observed unique molecules/total reads" to the ratio of "unique observed molecules/available fragments". Using this information, one can estimate how many different molecules there are in the population. The simulations were made for various amounts of available molecules "represented by the amount of input DNA" to show that the result does not vary for different number of available molecules.

Appendix 6

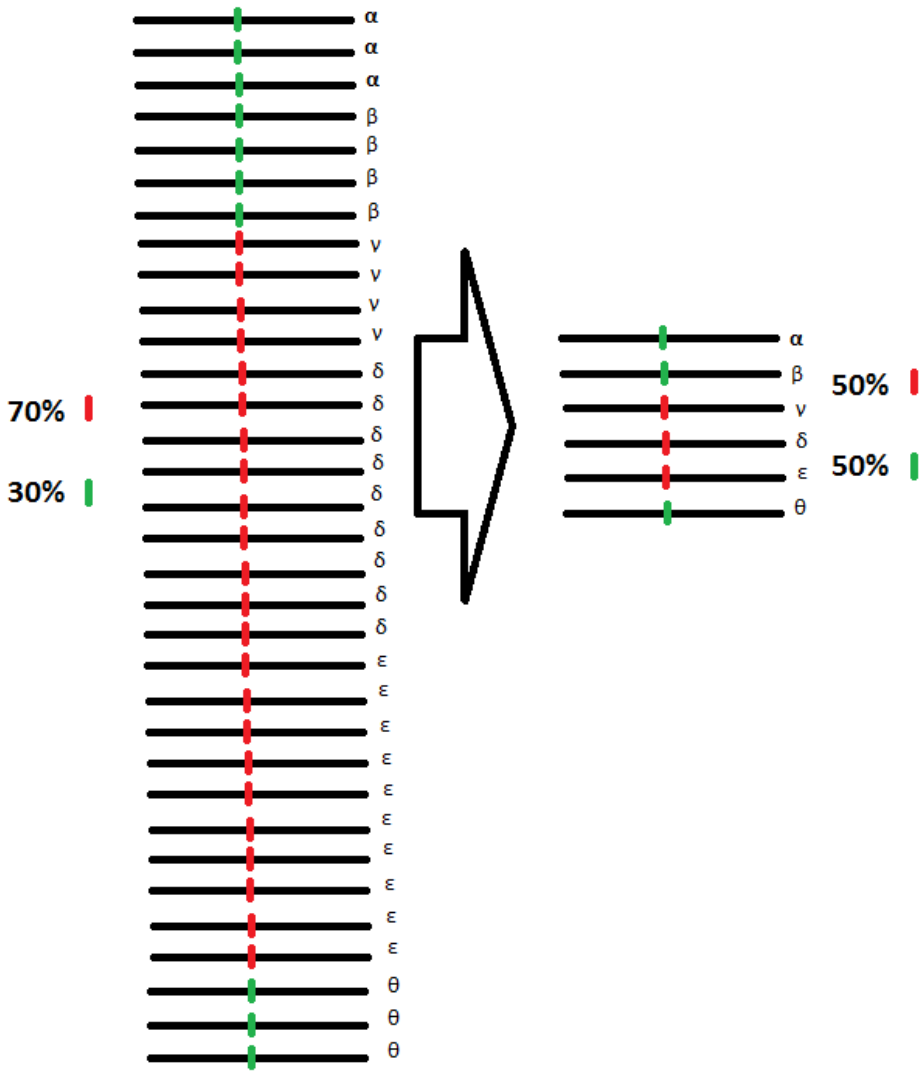

Figure 30. Illustration for the correction of skewed allele frequencies with the help of molecular barcodes. The black lines represents reads of the same sequence. The Greek letters (α-θ) represent molecular barcodes. The green and red marks represent two different types of nucleotides.

Figure 31 can give a simplified example on how molecular barcodes can be used to separate technical mutations from real mutations. In (A) the colors are base variations. By observing that the mutations does not exist in the same position on the other fragments with the same barcode, one can conclude that these mutations is technical and has probably arisen due to polymerase errors.

For figure 31 (B), assume an amplification mutation has occurred in an early PCR cycle during the library preparation for the HaloPlex enrichment. The blue mark in figure (B) represent the mutated base after it has been further amplified. Without molecular barcode, it is not possible to distinguish if it is a real mutation or if it is an technical one. By using molecular barcodes and demanding the mutation to exist at a fragment with different molecular barcode (very unlikely that a PCR mutation will occur at exactly same place on another molecule), one can reject it as an real mutation.

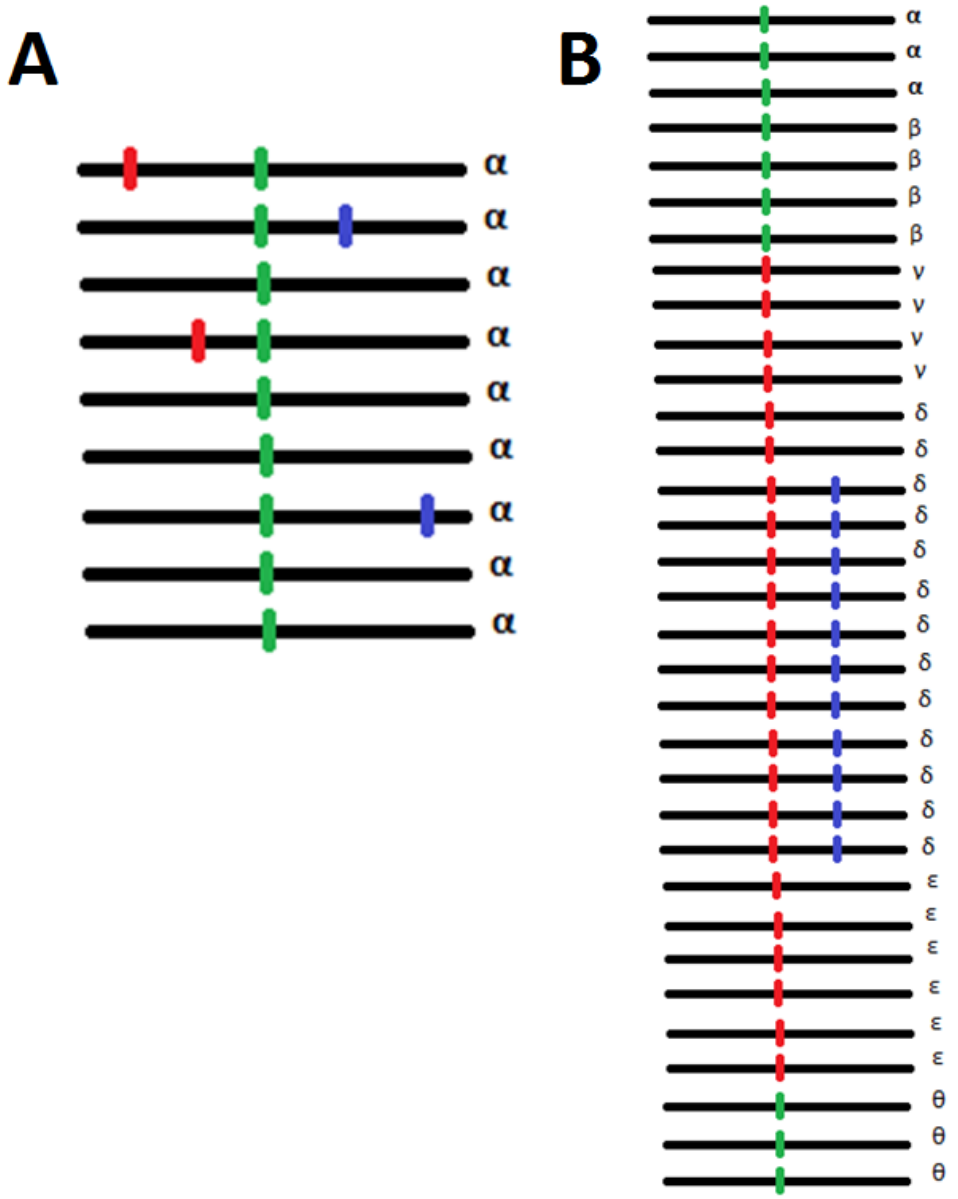

Figure 31. Figure illustrating fragments with single nucleotide variations represented by the colored marks. The Greek letters (α-θ) represent molecular barcodes. The green, blue and red marks represent different types of nucleotide bases.

Appendix 7

 # Algorithm for probability calculations, figure 26 (A).

```
m <- 7500 # Define the maximum number of molecules
P<- numeric(m) # Define a probability vector with empty elements
                # create a for-loop that counts the probability for each
                # number of molecules.
for (i in 1:m){
P[i] <- prod(1 - (0:(i-1))/4^13) }  #calculate the probability and store in vector.
NumberOfMolecules <- 1:m  # Define the X-axis
plot(NumberOfMolecules, P ,
  xlab="Number of molecules", ylab="Probability [no overlap]")
```

 # Algorithm for plotting the probability versus the length of the barcode in nucleotides, figure 27 (A).

```
p <- numeric(3000) #Define the number of calculation points
                #Create a for-loop and calculate the probabilities.
for (i in 1:3000){
j <- ((267386880*i)/3000)   #
p[i] <- prod(1 - (0:(7500-1))/j) }  # Calculate the probability and store it in a vector.
NumberOfCombinations <-seq(4^10, 4^14, 89129.96) # Define the x-axis.
plot(log(NumberOfCombinations,4) , p, xlab="Length of the molecular barcode[Nucleotides]" ,
main="Probability [7500 molecules with varying length of barcode]" , ylab="Probability [no overlap]")
# plot x in log4 versus y.
```

 # Algorithm for simulating the assignment of barcodes to molecules, figure 26 (B).

```
n <- 7500  #Define the number of molecules
m <- 1000 # Define the number of simulation runs.
x <- numeric(m) # create vector with m number of locations
for (i in 1:m)  {
                b <- sample(1:4^10, n, repl=T) # randomly assigning a barcode to each molecule
                x[i] <- n - length(unique(b))  }    # count the number of unique barcode-> subtract it from the
                                                      number of possible barcode ->
                                                    # and you get the number of molecules with overlapping
                                                      barcodes.
mean(x) # mean number of molecules with overlapping barcodes
mean(x==30) # returns an approximation of the probability that 30 molecules have overlapping barcodes.
cut <- (0:(max(x) + 1)) - 0.5 # Define the x-axis of the histogram
hist(x, breaks=cut, freq=F, col=8, xlab="Number of molecules with overlapping barcodes",
ylab="Density[probability]", main="Probability for x number of molecules with overlapping barcodes " ) #
```

 # plotting the number of molecules with overlapping barcodes versus the number of molecules, figure 27 (B)

```
k<-numeric(7500)
for (n in 1:7500) {
m <- 40
x <- numeric(m)
for (i in 1:m)  {
  b <- sample(1:4^10, n, repl=T)
  x[i] <- n - length(unique(b))  }
k[n]<-mean(x)
}
kx<- 1:7500
plot(kx, k , main="Number of molecules with overlapping barcodes versus the number of molecules ",
    xlab="Number of molecules", ylab="Number of molecules with overlapping barcodes")
```

 # Simulations and approximations for the number observed molecules of total available, figure 22.

```
S=1:30000
un=3000
uniqueObs=numeric(length(S))
result=numeric(length(S))
uresult=numeric(length(S))
for (i in 1:length(S)){
  uniqueObs[i]=un*(1-(1-(1/un))^i)
  result[i]=uniqueObs[i]/i
  uresult[i]=uniqueObs[i]/un
}
plot(result, uresult, type="l", col="red")
```