Marshall Plan Scholarship
Report

12.01.2015 – 12.07.2015
Pennsylvania State University

Adaptation of Next Generation Sequencing to Detect Rare Mutations in Human Genomic DNA

Barbara Arbeithuber, MA. rer. nat.

Internal supervisor (Johannes Kepler University Linz):
a. Univ-.Prof. Dr. Irene Tiemann-Boege

External supervisor (Pennsylvania State University):
Prof. Kateryna Makova, Ph.D.

Abstract

Over the past years, next-generation sequencing (NGS) has become a powerful tool in a wide variety of applications, among them: genetics, evolutionary biology, medicine, or forensics. Sequencing costs could be significantly reduced, and read lengths increased more and more. However, when it comes to the detection of rare DNA sequence variants or mutations in an excess of non-mutated genomes, one major drawback of NGS is the high error rate. In 2012, Schmitt et al. published an adaptation for NGS library preparation, termed "duplex sequencing", which allows the separate analysis of the forward and reverse strand of a DNA duplex, and therefore significantly decreases the error rate to $<10^{-7}$. Since the detection of rare mutations is a major focus in my PhD project, the aim of my research stay at the Pennsylvania State University in the lab of Kateryna Makova was to learn duplex sequencing, and establish the method for the use with a variety of samples. I could successfully prepare and sequence duplex sequencing libraries from plasmids, mitochondrial DNA (mtDNA), human genomic DNA (gDNA), PCR amplicons and synthetic DNA.

The low error rate of duplex sequencing comes with the cost of a high number of reads needed to get the information of the sequence of a single DNA molecule. Therefore, the optimization of enrichment procedures of mtDNA from human blood was necessary prior to duplex sequencing. Also for genomic DNA, regions of interest had to be enriched before sequencing.

With this research stay, I did not only get the chance to learn the wet-lab associated parts of such a powerful method like duplex sequencing, but I could also benefit from the computational expertise of the research group. Despite not having any background in bioinformatics, I learned how to perform simple analyses of NGS, and especially duplex sequencing data.

Two publications containing duplex sequencing data obtained during my research stay are currently in preparation.

Table of Contents

Introduction

Detection of rare DNA sequence variants or mutations in an excess of non-mutated genomes is the focus in a growing number of applications, including e.g. the detection of somatic mutations in tumor biopsies in the early screening of cancer or to monitor the response after therapy. Such rare events could be detected with technologies that provide high throughput and single molecule sensitivity. Several such technologies exist, known as next-generation sequencing (NGS), which allow the sequencing of a massive amount of DNA in parallel in a microscopic format (reviewed in van Dijk et al. 2014).

However, given the high error rates of these technologies, they are unsuitable for detecting rare mutations. Schmitt et al. developed a method termed "duplex sequencing", that greatly reduces errors resulting from DNA damage and amplification in NGS experiments, by independently tagging each of the two strands of a DNA duplex, followed by the amplification and sequencing of both strands separately (Schmitt et al. 2012). While true mutations are found in both DNA stands, artefacts resulting from DNA damage, and sequencing and PCR artefacts can be identified since they are only present in one of the strands or in several reads of one strand. An overview of the principle of the duplex sequencing method is shown in **Figure 1**.

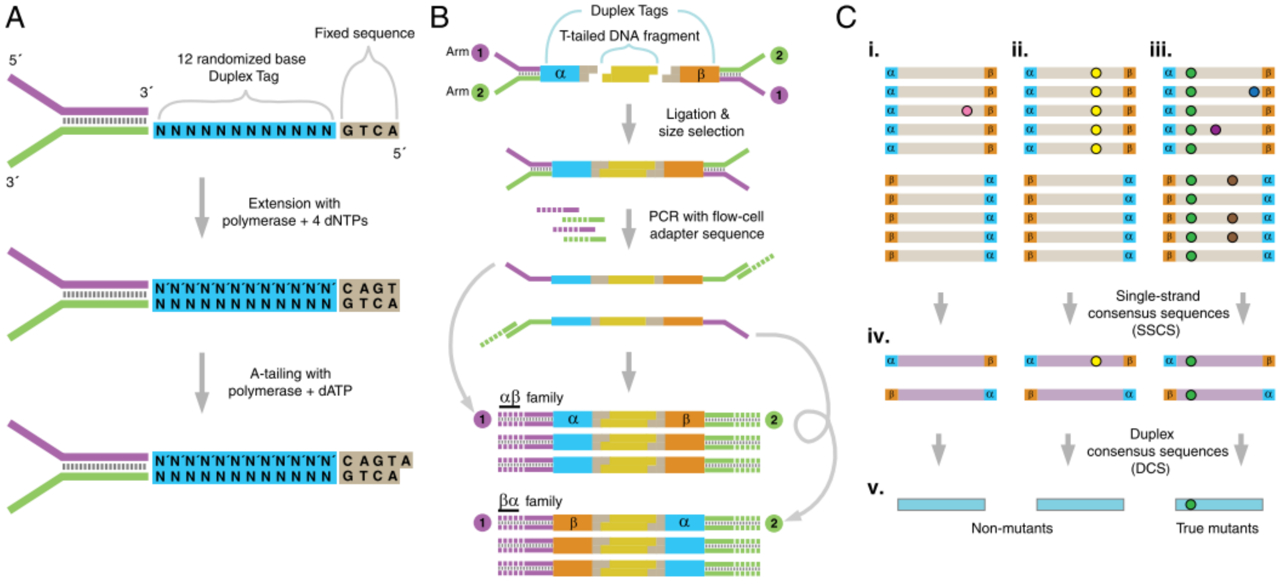

A
5´
12 randomized base Duplex Tag
Fixed sequence
3´
NNNNNNNNNNNN GTCA
5´
3´
Extension with polymerase + 4 dNTPs
N´N´N´N´N´N´N´N´N´N´N´N´ CAGT
NNNNNNNNNNNN GTCA
A-tailing with polymerase + dATP
N´N´N´N´N´N´N´N´N´N´N´N´ CAGTA
NNNNNNNNNNNN GTCA
B
Duplex Tags
T-tailed DNA fragment
Arm 1
Arm 2
α
β
2
1
Ligation & size selection
PCR with flow-cell adapter sequence
αβ family
βα family
C
i.
ii.
iii.
iv.
v.
Single-strand consensus sequences (SSCS)
Duplex consensus sequences (DCS)
Non-mutants
True mutants

Figure 1. Duplex sequencing scheme (Figure taken from Schmitt et al. 2012). (A) Duplex sequencing adapters, which contain a double-stranded 12-nucleotide randomized sequence are formed by extending DNA that is hybridized to a single-stranded oligo which contains a random sequence (shown as Ns). **(B)** Double-stranded duplex sequencing adapters are then ligated to both ends of the DNA fragments that need to be analyzed. After amplification, two "families" are formed for each DNA fragment: one of the forward strand that contains the randomized adapter sequences in the order of αβ, and the reverse stand with the random sequence order βα. **(C)** After sequencing, all reads from a family are combined to a single-strand consensus sequence (SSCS), which results in two SSCS per original DNA fragment. SSCSs contain only mutations that have been present in more than 70% of the reads, and therefore

exclude random sequencing errors, and amplification errors that occurred in later PCR cycles. These SSCSs, representing the forward and reverse strand of a double-stranded DNA fragment, are then further combined to one duplex consensus sequence (DCS), which only contains mutations that have been present in both strands.

Since several reads are necessary to form single-strand consensus sequences (SSCS), that represent the sequence of one strand of a DNA duplex (an average of 6-12 family members is optimal (Kennedy et al. 2014)), and two SSCS are needed to form a duplex consensus sequence (DCS), the number of required reads drastically increases compared to conventional NGS applications. Therefore, duplex sequencing is only applicable for small DNA samples such as plasmids or mitochondrial DNA (mtDNA). For genomic DNA, regions of interest have to be enriched before sequencing, this is necessary to avoid enormous costs for sequencing of bulk DNA.

The aim of my research stay at the Pennsylvania State University was to learn the duplex sequencing method, specifically, to prepare duplex sequencing libraries, sequence them with a MiSeq Illumina Sequencer, and learn to analyze the data according to the published protocols (Schmitt et al. 2012; Kennedy et al. 2014).

Duplex sequencing could successfully be performed on different DNA samples:

1) Duplex sequencing was performed on a well characterized plasmid, which provided an easy sample material to learn the basics of the duplex sequencing method due to its small size (~8 kb) and availability in large amounts.

2) In the next step, duplex sequencing was performed on enriched human mtDNA, which provided a human sample that is still smaller than human genomic DNA (about 180.000-times less bases). Therefore, sequencing of the whole mtDNA was possible without prior enrichment of selected regions.
However, prior to duplex sequencing, an efficient method to isolate and enrich mtDNA from human blood had to be established.

3) Duplex sequencing was also performed on targeted regions of human genomic DNA. Therefore, three disease associated microsatellites were selected, which provide an interesting target for the detection of rare mutations due to the relatively high mutation rate compared to other genomic regions (Ellegren 2004).
A targeted enrichment protocol was developed for these regions based on the published targeted enrichment method used for duplex sequencing of the *ABL1* gene (Schmitt et al. 2015).

4) Additionally, duplex sequencing was performed on (a) synthetic DNA and (b) PCR products, as parts of different research projects.

(a) A common source of sequencing artefacts is DNA damage resulting from e.g. DNA extraction or storage (Kunkel 1984; Ravanat et al. 2002), which can lead to a bias of the true number of mutations. Duplex sequencing therefore provides an optimal method to further explore the effect of such lesions. One common type of artefacts are guanine to thymine and cytosine to adenine transversions, resulting from the oxidation of guanine (→ 8-oxoguanine), where the polymerase preferentially puts an adenine opposite of 8-oxoguanine during amplification (Beard et al. 2010). We have already explored several approaches to reduce these preparation artefacts using enzymatic treatments of the DNA with the DNA glycosylase Fpg (formamidopyrimidine [fapy]-DNA glycosylase). Another common form of DNA damage is the deamination of cytosine or 5-methylcytosine. While the deamination product of cytosine is uracil, which can be easily removed by treating the DNA with a uracil DNA glycosylase (UDG) (Lindahl et al. 1977), deamination of 5-methylcytosine forms thymine, a base naturally occurring in DNA, that cannot easily be removed and therefore is a common source of artefacts. Different synthetic DNA molecules containing such lesions were analyzed by duplex sequencing, and also the effect of enzymatic treatments before sequencing library preparation was tested.

(b) One of the research interests of the Makova lab is the analysis of heteroplasmy in human mtDNA. Duplex sequencing was performed to obtain additional data for indel heteroplasmy analysis. The previously observed frequency of a human mtDNA indel, consisting of two or three 9-nucleotide repeats, was confirmed by duplex sequencing of a PCR product containing this indel region.

Libraries for all of the mentioned aims could be successfully prepared and sequenced, and library preparation protocols could be adapted to the different tasks. The analysis of the sequencing results could not be finished during my research stay at the Pennsylvania State University, and is still ongoing. For several applications, only preliminary results are reported.

Materials and Methods

Duplex adapter preparation

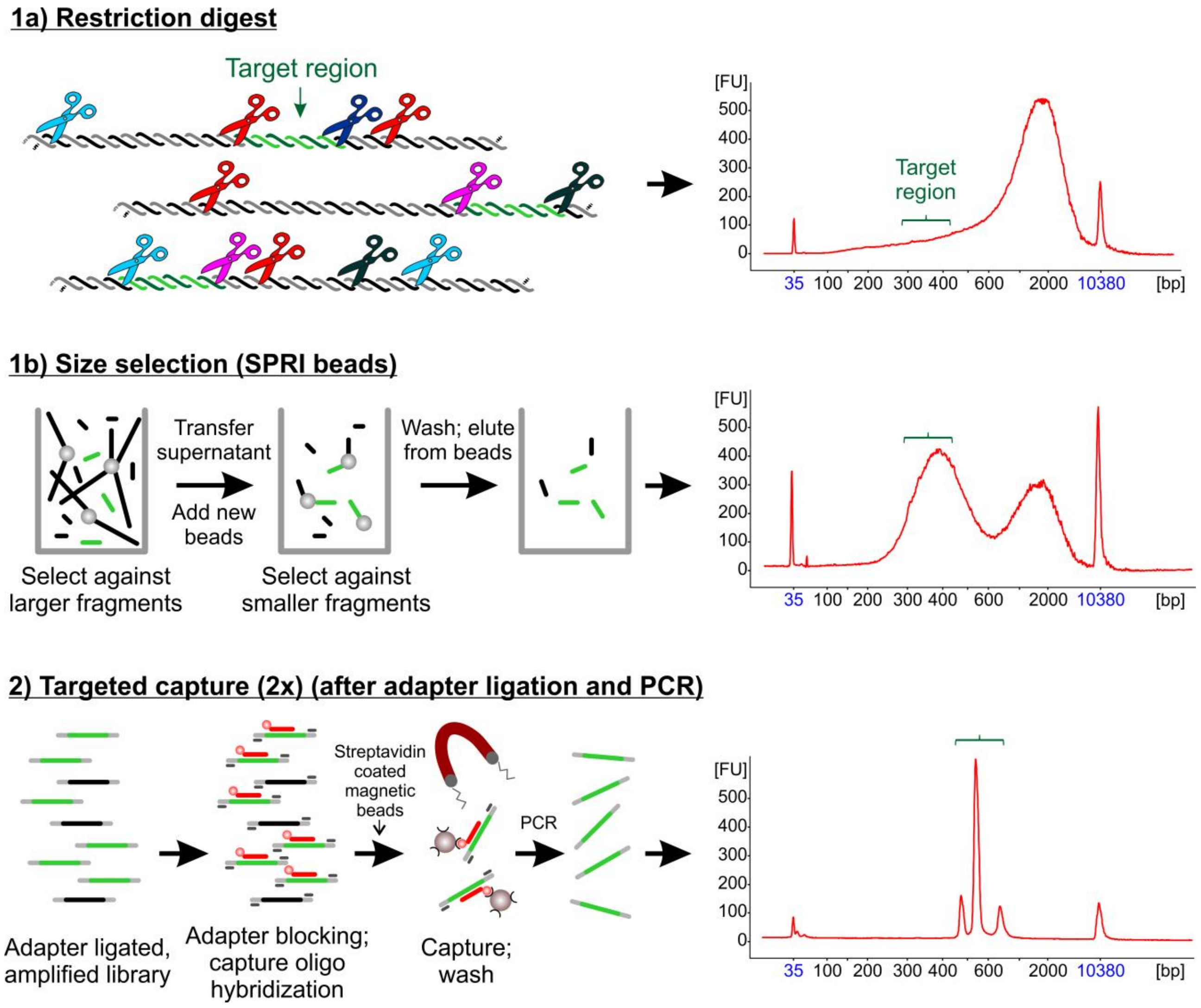
For duplex sequencing of plasmid DNA, duplex adapters were prepared as described in (Schmitt et al. 2012). For all other experiments the improved method for adapter preparation, which is based on T-tailed adapters and A-tailed DNA fragments was used, as described in (Kennedy et al. 2014), with some minor modifications in the protocol. In brief: T-tailed adapters were prepared by hybridization of the oligos MWS51 and MWS55 (sequences reported in (Kennedy et al. 2014)), followed by extension with the Klenow Fragment (3'→5' exo-) (NEB) and a restriction digest with TaaI (HypCH4III) at 60°C for 16h. Adapters were purified by ethanol precipitation with 2 volumes absolute ethanol and 0.5 volumes 5 M NH_4OAc. The different steps of adapter preparation were monitored on a 3% agarose gel (1.5% normal agarose and 1.5% low-melt agarose). Double-stranded adapters were aliquoted and stored at -20 °C until use.

Duplex library preparation and sequencing

Duplex libraries from plasmids (pML113 with and without two different short tandem repeats (STRs)) were prepared as described in (Schmitt et al. 2012). All other libraries were prepared as described in (Kennedy et al. 2014), with minor modifications depending on the application. Fragmentation of mtDNA and long-range PCR amplicons to a fragment size ~550 bp was performed with a Covaris S2 (duty cycle 5%, intensity 3, time 75 sec, sample volume 130 µl). Fragmented DNA, or amplicons were end-repaired with the End-Repair Enzyme Mix provided in the Illumina TruSeq Kit or the NEBNext End Repair Module (NEB) according to manufacturer’s instructions, A-tailed, size selected with 0.55 and 0.7 volumes Agencourt SPRI beads (Beckmann Coulter) (not necessary for short amplicons), and the adapter was ligated with 1800 U T4 ligase (NEB) with 20x molar excess at 16 °C for 30 min. The amount of adapter-ligated DNA used for the generation of amplified tag families varied between the different experiments and samples, the optimal cycle number for amplification was evaluated by real-time PCR. Purifications of DNA between the different steps of library preparation were performed with Agencourt AMPure beads (Beckmann Coulter).

For duplex sequencing of three selected human genomic regions (that contain microsatellites), targeted enrichment had to be performed. The method was adapted from what was published by Schmitt et al. (Schmitt et al. 2015) with the addition of another enrichment step (restriction digest) prior to library preparation. An overview of the different enrichment steps is shown in **Figure 2**.

The libraries were quantified with the KAPA Library Quantification Kit (Kapa Biosystems). Sequencing was performed on an Illumina MiSeq platform producing 151, 251, or 301 bp paired-end reads.


1a) Restriction digest

1b) Size selection (SPRI beads)

2) Targeted capture (2x) (after adapter ligation and PCR)

Target region
[FU]
500
400
300
200
100
0
Target region
35 100 200 300 400 600 2000 10380 [bp]
Transfer supernatant
Add new beads
Wash; elute from beads
Select against larger fragments
Select against smaller fragments
[FU]
500
400
300
200
100
0
35 100 200 300 400 600 2000 10380 [bp]
Streptavidin coated magnetic beads
PCR
Adapter ligated, amplified library
Adapter blocking; capture oligo hybridization
Capture; wash
[FU]
500
400
300
200
100
0
35 100 200 300 400 600 2000 10380 [bp]

Figure 2. Targeted enrichment of selected microsatellites from human genomic DNA. 1a) In the first step of enrichment, three DNA regions of interest, that contain disease associated microsatellites, were cut out of human blood DNA using five different restriction enzymes (all of them 6-cutter). It was inevitable, that also other regions within the DNA were digested, however, most of these regions were of a higher molecular weight as the target regions. **1b)** In the first step of size selection, it was selected against regions of higher molecular weight by binding these regions to Agencourt SPRI beads (Beckmann Coulter) and further proceeding with the supernatant. In the next step, the regions of interest and regions with a similar length were bound to the beads, and unbound DNA was washed away, resulting in an enriched DNA at a size between 300 and 400 bp. The binding capacity of the used beads was not high enough to select against all high molecular weight DNA resulting in a second, smaller peak with a maximum around 1800 bp. **2)** After duplex adapter ligation, the libraries were amplified, and the target regions were captured by hybridization of specific biotinylated oligos, followed by capture and washing (to remove unspecific regions) with streptavidin-coated magnetic beads. Adapter sequences were blocked with blocking-oligos to avoid interference with specific hybridization. Two rounds of this targeted capture were performed with amplification in between.

For library preparation of double-stranded synthetic DNA containing different DNA lesions, the end-repair time (NEBNext End Repair Module (NEB)) was elongated to 1 h (1.5 h for insert 5) instead of the suggested 30 min. This adaptation was necessary to blunt the 20 bp 3’-overhangs of the fragments (one 50 bp overhang in insert 5). All other steps were performed as described before for short amplicons. For insert 5, purification steps had to be performed by ethanol precipitation instead of Agencourt AMPure beads due to the small size (80 bp) of the fragment.

mtDNA enrichment

Two different mtDNA enrichment methods were tested to obtain sufficient enrichment for blood samples:

1) Mitochondria enrichment with dounce homogenization

The protocol was adapted for blood samples from the already published protocol for human brain tissue (Kennedy et al. 2013). As a control, the enrichment protocol was also tested with 115 mg mouse liver, for which similar amounts of mtDNA were expected as for the published human brain, but after enrichment, which perfectly worked for this sample according to real-time PCR analysis, no further experiments were performed with this sample. Additionally, the protocol was tested with 5 ml human saliva.

5 ml homogenization buffer (0.32 M sucrose, 1 mM EDTA, 10 mM Tris-HCl, pH 7.8) were added to 7 ml frozen blood (or 115 mg mouse liver, or 5 ml human saliva) and dounced with a glass pestle with 5 strokes, followed by centrifugation (1000g, 20 min, 4 °C). The supernatant was transferred to a new tube and centrifuged again (12000g, 35 min, 5 °C).

The supernatant was discarded, the pellet (mitochondria) was resuspended in 200 µl mito-DNase buffer (0.3 M sucrose, 10 mM $MgCl_2$, 0.15% BSA (w/v), 20 mM Tris-HCl, pH 7.5, 0.01 mg/mL DNase) and incubated at 37 °C for 1.5 h. With this step, contaminating nuclear DNA was removed, while not affecting mtDNA that is protected within intact mitochondria. Mitochondria were re-pelleted (12000g, 30 min) and the supernatant discarded. The mitochondria pellet was washed twice with 1 ml mito-DNase buffer followed by resuspension in 200 µl lysis buffer (150 mM NaCl, 20 mM EDTA, 1% SDS (w/v), 10 mM Tris-HCl, pH 7.8, 0.2 mg/mL Proteinase K, 0.01 mg/mL RNase) and incubation at 56 °C for 1h. After mitochondria lysis mtDNA was extracted by standard phenol-chloroform DNA extraction followed by ethanol precipitation.

2) mtDNA enrichment from extracted total blood DNA with exonuclease V

This method is based on the protocol published by Jayaprakash et al. (Jayaprakash et al. 2015) with digestion of linear nuclear DNA with exonuclease V while preserving the circular mtDNA.

mtDNA from 900 ng total blood DNA (extracted from 100 µl frozen blood with the DNeasy Blood & Tissue Kit (QIAGEN) according to manufacturer's instructions) or 100 ng pre-enriched mtDNA from blood (by dounce homogenization) were enriched in two rounds of exonuclease V digests:

1) Digestion with 40 U exonuclease V (NEB) at 37 °C for 64h followed by purification with 1.4 volumes AMPure XP beads (Beckmann Coulter) according to manufacturer's instructions and elution in 40 µl PCR-grade water.
2) Digestion with 40 U exonuclease V (NEB) at 37 °C for 24h followed by purification with 1.4 volumes AMPure XP beads (Beckmann Coulter) according to manufacturer's instructions and elution in 40 µl PCR-grade water.

Enrichment was tested with real-time PCR by analyzing the ΔCq of mtDNA specific amplification and genomic DNA (gDNA) specific amplification from non-enriched and enriched samples.
Two reactions were set up for each sample with the gDNA specific primers (CAGTGACCATCTGGCCAGAA and ATTTGCCCAGGCCCAGAAAG) and mtDNA specific primers (CCACAGCACCAATCCTACCT and GTCAGGGGTTGAGGTCTTGG), respectively. The 10 µl PCR reactions contained 1 µl enriched mtDNA, 0.2 µM each primer, and 1x EXPRESS SYBR® GreenER™ qPCR SuperMix with premixed ROX (LifeTechnologies). The reactions were carried out with an initial heating step of 95°C for 20 sec, followed by 45 cycles at 95°C for 2 sec, and 60°C for 20 sec, and a final extension at 60°C for 1 min.

Duplex sequencing data analysis

Duplex sequencing data was analyzed according to the pipeline published by Kennedy et al. (Kennedy et al. 2014). Alignments were inspected with the Integrative Genomics Viewer (IGV) (Robinson et al. 2011).

Results and Discussion

Because of the low error frequency that can be yielded with duplex sequencing, this method provides a powerful tool for mutation analysis in a high diversity of applications. During my research stay at the Pennsylvania State University, I could successfully learn and establish the duplex sequencing method (only one successful duplex sequencing experiment had been performed before together with a collaborating lab) for different applications. Optimization of the method was necessary to make it work with the available MiSeq Illumina Sequencer instead of a HiSeq Illumina Sequencer, that was used in the published protocol (Kennedy et al. 2014), especially considering the amount of adapter-ligated library used for the generation of amplified tag families to reach a specific sequencing coverage.

Duplex sequencing of plasmid DNA

Duplex sequencing was performed on three plasmids: pML113 without an STR, pML113 with a $(AT)_{12}$ + $(A)_{28}$ STR, and pML113 with a $(T)_{19}$ STR. 10 amole of each adapter-ligated plasmid library were used for the generation of amplified tag families. The major purpose to sequence these plasmids was to learn duplex sequencing with well established, small (~8kb) samples that were easily available.

Duplex consensus sequences (DCS) could be generated for all of the samples (9408, 4054, and 2697, respectively), however, only the pML113 control fell into the acceptable tag family size with an average of 5 family members (see **Figure 3**). Despite identical library preparation of all three samples, the plasmids containing STRs yielded lower average family member numbers and therefore lower numbers of DCS. A problem that occurred here during analysis is, that STRs resulted in a high sequence variability in the original sequencing reads. This made it problematic to map those reads to a reference sequence, which is normally done prior to SSCS formation, and therefore many reads are not included in the analysis. An overview of the high variability in the length of the $(T)_{19}$ STR itself is shown in **Figure 4**. To overcome this problem, a reference free duplex sequencing analysis pipeline is currently in development.

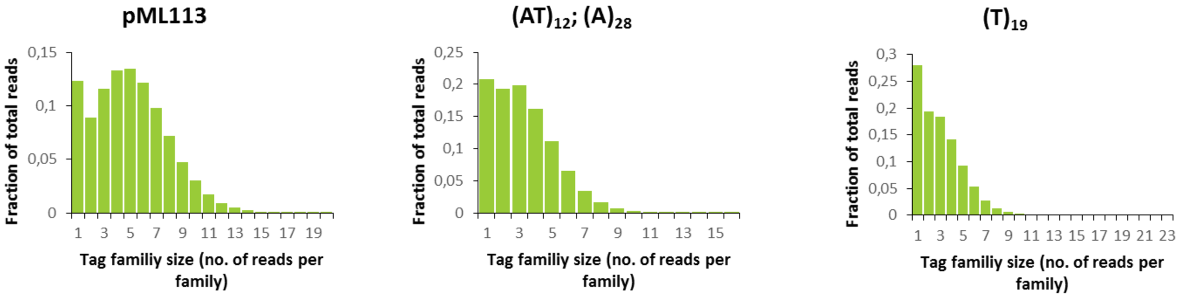


Figure 3. Tag family size distribution of plasmid samples. Three different plasmids: pML113, pML113 with an insert containing a $(AT)_{12}$+$(A)_{28}$ STR, and pML113 with an insert containing a $(T)_{19}$ STR were analyzed by duplex

sequencing. The tag family size, which represents the number of sequenced PCR products for each strand (from which SSCSs are formed) of a DNA duplex is shown for the analyzed plasmids.

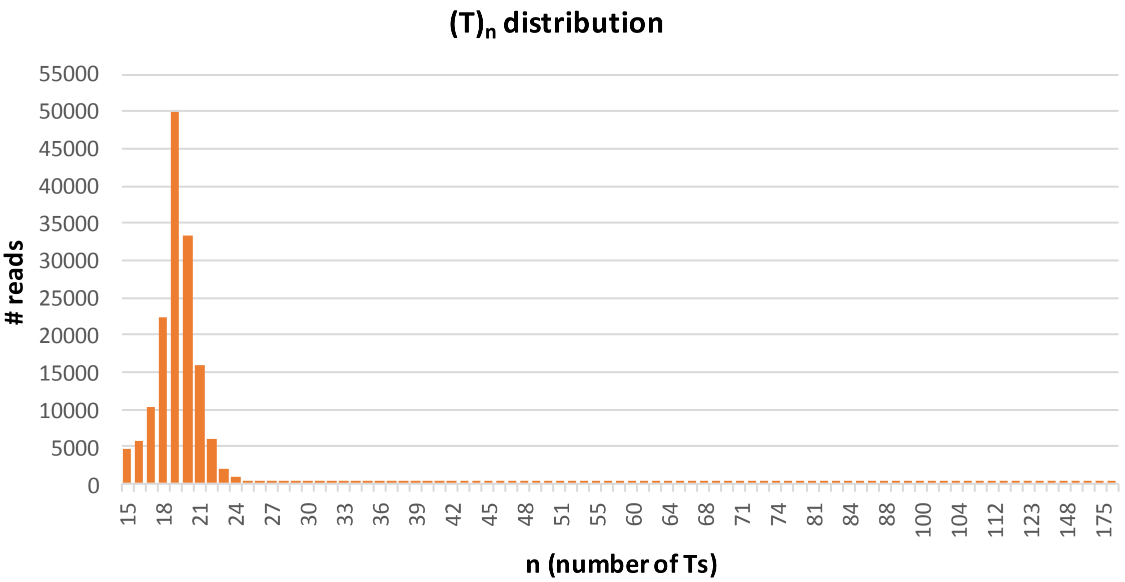


Figure 4. Distribution of the number of Ts within a $(T)_{19}$ STR. A great variability in the number of Ts was observed for paired-end reads containing a $(T)_{19}$ STR region (third analyzed plasmid). This hampers the mapping of these reads with the Burrows-Wheeler Aligner (BWA) (Li and Durbin 2010), which is used by default in the duplex analysis pipeline, leading to a loss of these reads for further analysis. The repeat number distribution was analyzed with the STR-FM pipeline (Fungtammasan et al. 2015).

Duplex sequencing of human mtDNA

1. mtDNA enrichment

Considering the size of human mtDNA (16,569 bp) in comparison to gDNA (~3 x 10^9 bp), despite the presence of several copies of mtDNA, the majority of the DNA in a cell is genomic. This makes it necessary to enrich for mtDNA prior to duplex sequencing. In comparison to tissues with a high energy requirement (e.g. brain or muscle), blood cells contain only a few copies of mtDNA. Quantification with digital droplet PCR (ddPCR) and qPCR showed, that in human blood, a cell contains on average ˜200 copies of mtDNA.

mtDNA enrichment, with the goal to efficiently enrich mtDNA from human blood, was tested with two different methods: 1) Mitochondria enrichment with dounce homogenization; 2) mtDNA enrichment from extracted total blood DNA with exonuclease V. The efficiency of enrichment was then tested with qPCR, and for the most promising samples additionally with sequencing.

With dounce homogenization, mtDNA was enriched from 7 ml fresh human saliva, 115 mg mouse liver, and 5 ml frozen human blood. According to Kennedy et al. (Kennedy et al. 2013), a ΔCq ≥ 17.5 cycles in the amplification with gDNA and mtDNA specific primers represents enrichment to having an equal mass of mtDNA and gDNA. For human saliva, not nearly enough enrichment could be yielded to use the sample for further applications, also the total amount of extracted DNA was low (see **Table 1**). A possible explanation for this observation is that the mitochondria in saliva were not intact which leads to digestion of mtDNA with DNase I. The functionality of the enrichment method could be shown with mouse liver, for which more than half of the enriched 1670.4 ng DNA were mtDNA specific. Similar as for the saliva sample, enrichment was not sufficient enough for blood. Additionally to qPCR, mtDNA enrichment was tested with sequencing, and the observed ΔCq of 10.82 represented ~8.5% of all reads to be mtDNA specific.

Table 1. mtDNA enrichment with dounce homogenization

mtDNA extracted from:	DNA yield [ng]	ΔCq
7 ml human saliva	221.8	8.16
115 mg mouse liver	1670.4	18.43
5 ml human blood	975.0	10.82

As a second method, mtDNA was enriched from already extracted total DNA by digestion of linear DNA with exonuclease V, while the circular mtDNA is preserved (Jayaprakash et al. 2015). With this method, mtDNA was enriched from freshly extracted total DNA (from 100 µl blood), or from already enriched mtDNA (from 5 ml human blood with dounce homogenization, for which only ~8.5% of the reads mapped to the mitochondrial genome). The exonuclease V digest turned out to be the more efficient method for enrichment: all of the reads of the pre-enriched sample, and 73% of the sample with total DNA extraction mapped to mtDNA, representing a ΔCq of 20.89 and 18.86, representatively (see **Table 2**).

Table 2. mtDNA enrichment with exonuclease V

mtDNA extracted from:	DNA yield [ng]	ΔCq
390 ng pre-enriched blood DNA	30.0	20.89
900 ng total blood DNA	<1.5	18.86

2. Duplex sequencing

Both exonuclease V enriched samples were analyzed by duplex sequencing. Sequences were obtained as 251 bp paired-end reads. Before analysis, the sequences were trimmed by 5 bases using Galaxy – Trim Sequences (Giardine et al. 2005; Blankenberg et al. 2010; Goecks et al. 2010), or by 100 bases to test the effect of sequence length on the duplex analysis pipeline. Trimming was necessary, since the duplex sequencing pipeline only accepts sequences of the same length, however, the sequencing data also contained sequences that were a few bases shorter, which would have been discarded from the analysis otherwise. Despite a rather high average peak family size of 28 and 40 for exonuclease V treated pre-enriched blood DNA and total blood DNA, respectively, only 96 and 89 DCS could be formed when the sequences were trimmed by 5 bases (to 246 bases) (see **Table 3**). Trimming of the sequences to 151 bases could further increase the number of formed DCS. In this preliminary analysis, we found that many reads are lost during the analysis. DCS present in the 151 base sequences are not present in the 246 base sequences and *vice versa* (see **Figure 5**). Further analysis of the data is necessary.

Table 3. DCS analysis of exonuclease V enriched mtDNA

mtDNA extracted from:	Paired-end reads	SSCS	DCS
pre-enriched blood DNA (246 bases)	80,290	3,827	96
total blood DNA (246 bases)	218,755	4,190	87
total blood DNA (151 bases)	218,755	6,618	195

A coverage up to 11 reads could be obtained for the pre-enriched blood DNA sample, up to 10 for the total blood DNA sample (see **Figure 5**). Unfortunately, not the whole mtDNA could be covered with DCS reads, however, there is potential to increase the coverage by optimization of library preparation and duplex analysis.

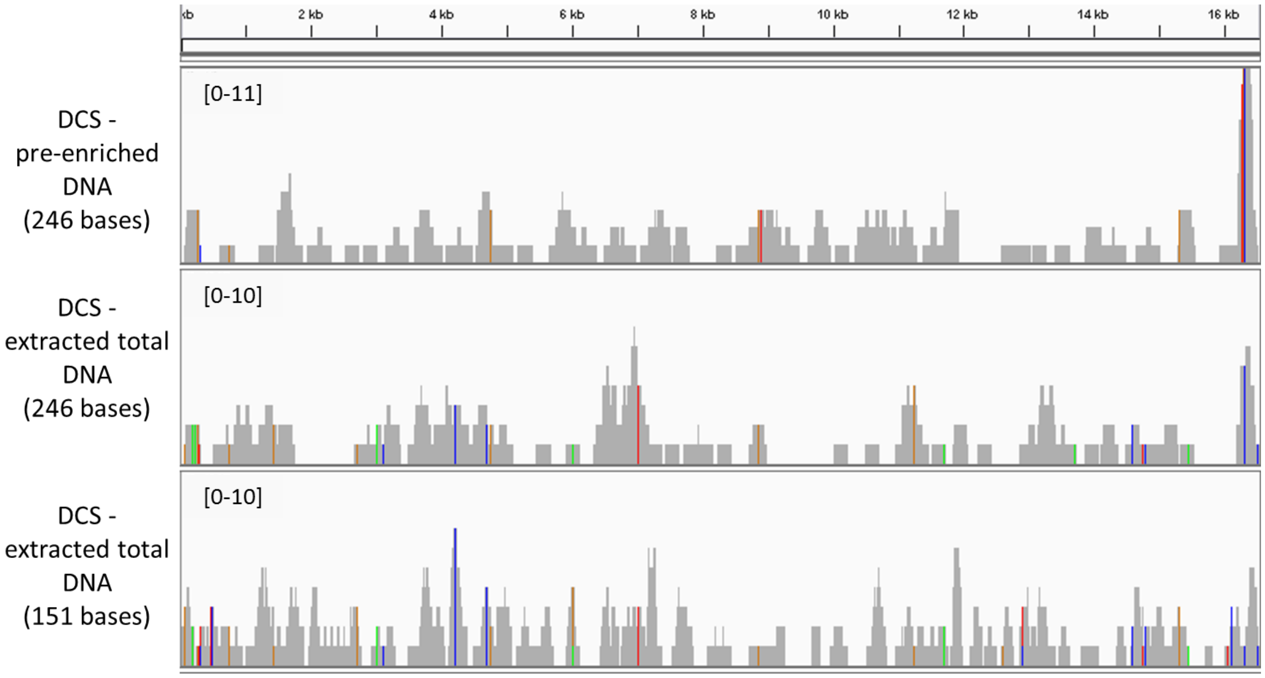


Figure 5. Coverage of the mtDNA specific duplex consensus sequences. The coverage for DCS reads mapping to the mitochondrial genome is shown for exonuclease V treated pre-enriched blood DNA and total blood DNA trimmed to 246 bases. For the second sample, coverage is also shown for sequences trimmed to 151 bases.

Duplex sequencing of human microsatellites

Duplex libraries were prepared for three disease associated human microsatellites (SCA6 (Spinocerebellar ataxia 6), AIB-I (Increased prostate cancer risk), and COMP (Multiple Skeletal dysplasias)). Enrichment of the mentioned region was necessary, which was first done by digestion of the DNA with 5 different restriction enzymes, followed by size selection, and next by targeted capture enrichment with biotinylated region specific oligos (see **Figure 2**). As already explained for mtDNA, reads were trimmed by 5 bases prior to duplex sequence analysis. All three regions could be successfully enriched, with ~99% of all reads mapping to these regions.

As already observed for the plasmids that contain STRs, problems in mapping these reads with BWA lead to a loss of a high proportion of reads during analysis. Concerning coverage, the three regions are represented with a different depth, with AIB-I yielding the highest coverage (4286 for non-overlapping DCS sequences) (see **Figure 6**). This difference in coverage could result from differences in the efficiency of targeted capture of these regions, but also from inequalities in the mapping efficiencies. When e.g. considering the SSCS obtained for SCA6, read 1 yields a coverage of 1231x while read 2 only yields 110x, for which an equal coverage would be expected. A high variability in the paired-end read sequences was observed for the analyzed microsatellite regions (as already described for STR containing plasmids), making the published reference based duplex analysis unsuitable. As soon as the reference free analysis is available, the microsatellite duplex data has to be reanalyzed.

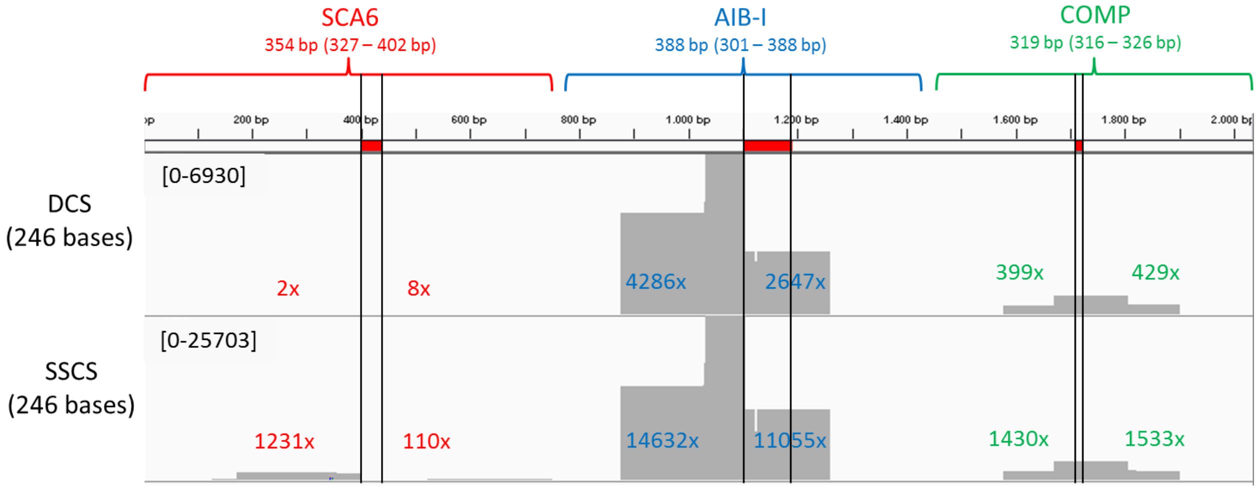


Figure 6. DCS and SSCS coverage of enriched human microsatellites. The coverage of DCS and SSCS sequences is shown for the three analyzed disease associated microsatellite regions SCA6, AIB-I, and COMP. The numbers of read 1 and read 2 (paired-end reads) present for each region are shown in the representative colors.

Duplex sequencing of synthetic DNA – analysis of DNA lesions

The effect of different DNA lesions on single molecule PCR (smPCR) was already analyzed in previous work, for which a publication is currently in preparation. To obtain additional data on the amplification behavior of these lesions (which will be included in the publication), duplex sequencing libraries were prepared for five previously designed double-stranded DNA fragments (insert 2-6) containing the different lesions and/or mismatches (see **Figure 7**). For insert 3, additionally to an untreated library, a library that was treated with USER (Uracil-Specific Excision Reagent), which excises uracil sites and therefore produces abasic sites and single-stranded breaks, making the template unamplifiable. Analogous, insert 5 was also treated with Fpg ((formamidopyrimidine [fapy]-DNA glycosylase), which excises the oxidation product of guanine -> 8-oxoguanine, and also forms unamplifiable single-stranded breaks. For insert 6, libraries were prepared for an untreated control, an aliquot heated to 65 °C for 3.5h, and an aliquot that was thawed, frozen for 14 days, followed by two additional freeze thaw cycles at -20 °C for 24h between thawing.

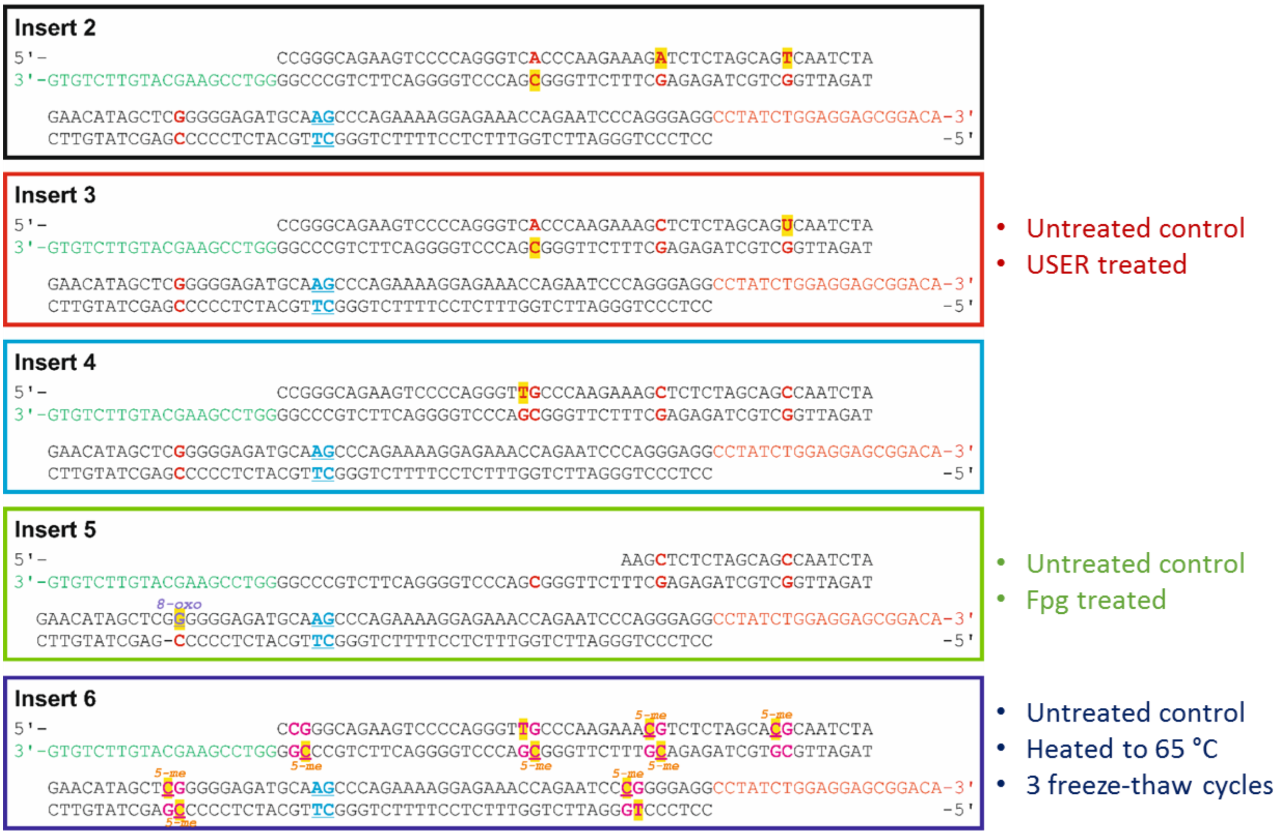

Insert 2
5'- CCGGGCAGAAGTCCCCAGGGTCACCCAAGAAAGATCTCTAGCAGTCAATCTA
3'-GTGTCTTGTACGAAGCCTGGGGCCCGTCTTCAGGGGTCCCAGCGGGTTCTTTCGAGAGATCGTCGGTTAGAT
GAACATAGCTCGGGGGAGATGCAAGCCCAGAAAAGGAGAAACCAGAATCCCAGGGAGGCCTATCTGGAGGAGCGGACA-3'
CTTGTATCGAGCCCCCTCTACGTTCGGGTCTTTTCCTCTTTGGTCTTAGGGTCCCTCC -5'
Insert 3
5'- CCGGGCAGAAGTCCCCAGGGTCACCCAAGAAAGCTCTCTAGCAGUCAATCTA
3'-GTGTCTTGTACGAAGCCTGGGGCCCGTCTTCAGGGGTCCCAGCGGGTTCTTTCGAGAGATCGTCGGTTAGAT
GAACATAGCTCGGGGGAGATGCAAGCCCAGAAAAGGAGAAACCAGAATCCCAGGGAGGCCTATCTGGAGGAGCGGACA-3'
CTTGTATCGAGCCCCCTCTACGTTCGGGTCTTTTCCTCTTTGGTCTTAGGGTCCCTCC -5'
• Untreated control
• USER treated
Insert 4
5'- CCGGGCAGAAGTCCCCAGGGTTGCCCAAGAAAGCTCTCTAGCAGCCAATCTA
3'-GTGTCTTGTACGAAGCCTGGGGCCCGTCTTCAGGGGTCCCAGCGGGTTCTTTCGAGAGATCGTCGGTTAGAT
GAACATAGCTCGGGGGAGATGCAAGCCCAGAAAAGGAGAAACCAGAATCCCAGGGAGGCCTATCTGGAGGAGCGGACA-3'
CTTGTATCGAGCCCCCTCTACGTTCGGGTCTTTTCCTCTTTGGTCTTAGGGTCCCTCC -5'
Insert 5
5'- AAGCTCTCTAGCAGCCAATCTA
3'-GTGTCTTGTACGAAGCCTGGGGCCCGTCTTCAGGGGTCCCAGCGGGTTCTTTCGAGAGATCGTCGGTTAGAT
8-oxo
GAACATAGCTCGGGGGGAGATGCAAGCCCAGAAAAGGAGAAACCAGAATCCCAGGGAGGCCTATCTGGAGGAGCGGACA-3'
CTTGTATCGAG-CCCCCTCTACGTTCGGGTCTTTTCCTCTTTGGTCTTAGGGTCCCTCC -5'
• Untreated control
• Fpg treated
Insert 6
5-me
5-me
5'- CCGGGCAGAAGTCCCCAGGGTTGCCCAAGAAACGTCTCTAGCACGCAATCTA
3'-GTGTCTTGTACGAAGCCTGGGGCCCGTCTTCAGGGGTCCCAGCGGGTTCTTTGCAGAGATCGTGCGTTAGAT
5-me
5-me
5-me
5-me
5-me
GAACATAGCTCGGGGGAGATGCAAGCCCAGAAAAGGAGAAACCAGAATCCCGGGGAGGCCTATCTGGAGGAGCGGACA-3'
CTTGTATCGAGCCCCCTCTACGTTCGGGTCTTTTCCTCTTTGGTCTTAGGGTCCCTCC -5'
5-me
• Untreated control
• Heated to 65 °C
• 3 freeze-thaw cycles

Figure 7. Synthetic DNA fragments (insert 2-6) analyzed with duplex sequencing. Five different inserts were analyzed to address different amplification behavior. Insert 2 has three different mismatches. Insert 3 has a U in one strand, and a mismatch in the second strand. Insert 4 has a T/G mismatch, mimicking the deamination event of a methylated cytosine in the context of a CpG site. Insert 5 contains an 8-oxoguanine and an additional guanine in front of the 8-oxoguanine allowing the distinction of the amplified strands. Insert 6 has 4 methylated cytosines (5-me) in a CpG context in each of the strands. Additionally it contains two T/G mismatches, mimicking deamination products of 5-methylcytosines.

A preliminary analysis was performed for some of the libraries; however, not all of the duplex sequencing data was analyzed yet. To obtain independent sequencing data for the forward and reverse strands, SSCSs were analyzed instead of DCS. 151 base paired-end reads were sequenced, which were then trimmed to 126 bases (96 bases for Insert 5) before analysis, as described for mtDNA. This trimming was necessary since the inserts had only a length of 110 bases (80 base for Insert 5), to which the duplex adapters were ligated. After trimming, the published duplex sequencing analysis was performed, however, only the SSCS bam-files were used for further analysis. An overview on the used attomoles for the PCR to generate amplified tag families, the obtained number of paired-end (PE) reads, the number of SSCS and DCS, and the average tag family size are shown in **Table 4**. Since the experiment was designed that paired-end reads were overlapping, it was necessary to merge them during sequence analysis. Therefore, the reads in the bam-file were first separated into read 1 and read 2 and further into forward- and reverse-mapping reads allowing the separate analysis of the forward and reverse strand, using samtools (Li et al. 2009; Li 2011). Only reads for which both paired-end reads were present, were further used for analysis. Therefore, the bam-files were first converted to fastq format using bamtools (Barnett et al. 2011), only reads for which both pairs were present were kept, and next, paired end reads were merged using PEAR (Zhang et al. 2014). After merging, the reads were again mapped to the sample specific reference sequence using BWA.

Table 4. Duplex sequencing analysis of the inserts

	Insert 2	Insert 3	Insert 3 USER	Insert 4	Insert 5	Insert 5 Fpg	Insert 6 crtl	Insert 6 65°C	Insert 6 frozen
Attomoles	1	1	1	1	1	1	1	1	1
Read length	126	126	126	126	96	96	126	126	126
PE reads	698,267	1,166,676	691,964	1,189,007	5,763,162	5,323,091	927,450	1,060,267	946,541
SSCS	**116,037**	**91,048**	**83,617**	**122,071**	**277,884**	**195,742**	**59,962**	**33,818**	**63,088**
DCS	**21,828**	**8,172**	**505**	**24,813**	**36,249**	**18,304**	**8,574**	**4,780**	**8,418**
Av. Tag family size	7	19	11	15	35	40	23	48	22

For insert 2, which contains 3 mismatches, a preliminary analysis was performed, testing the influence of the reference sequence on strand-specific mapping. By using a strand-specific reference, mapping can be biased towards the used strand, despite a difference in only three bases. A reference containing mixed bases can be used to avoid such a strand bias, which was done in all further analysis.

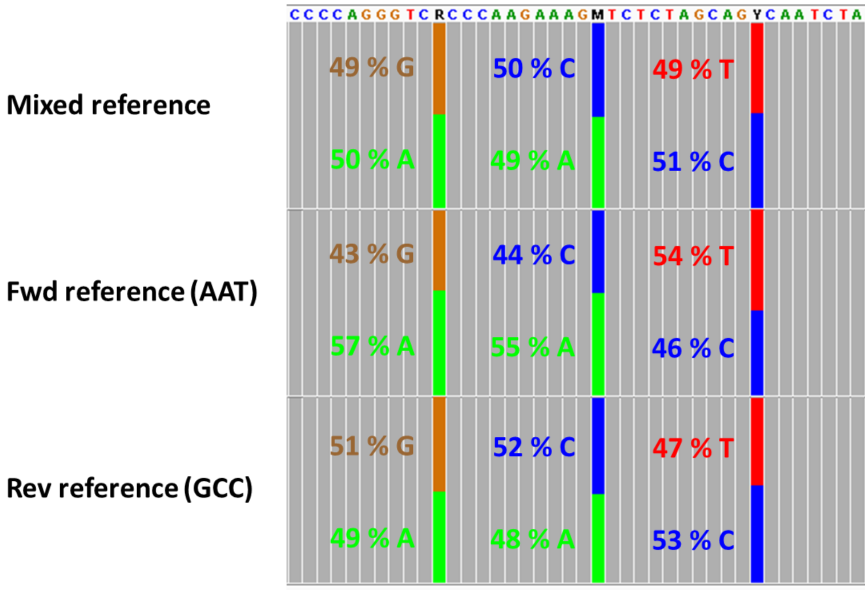


Figure 8. Dependency of the strand distribution on the identity of the reference sequence. Duplex analysis was performed with three different reference sequence, one containing mixed bases at the mismatch sites, one containing the bases specific for the forward strand, and one containing the bases specific for the reverse strand. While for the mixed reference approximately half of the SSCS reads represented the forward or reverse strand, using a strand specific reference biased the results towards the used strand.

For insert 3, which contains a uracil in the forward strand, the amplification behavior of the Kappa HiFi Hot Start DNA polymerase in the amplification of a uracil was tested. Additionally, the effect of USER, which excises uracil sites, was analyzed. For the control, 6,096 forward mapping SSCS reads, and 27,976 reverse mapping reads were obtained. Without an amplification bias, equal numbers would be expected for both strands. However, these results confirm our previous observation, that uracils hamper the amplification efficiency, but still in ~18% of the reads uracil could be amplified. For the USER treated sample, 138 forward mapping SSSCS reads, and 34,498 reverse mapping reads were obtained, reducing the amount of uracil-containing amplifiable reads to 0.4%. The remaining amplifiable reads can also contain synthesis error, for which the frequency still has to be analyzed.

For insert 5, which contains an 8-oxogunanine in the forward strand, only the analysis of the control is completed. The results confirm our observation with smPCR, that additionally to G→T mutations, which is observed the majority of reads (83.07%), 8-oxogunanine can also lead to G→C mutations (4.05% of the reads). In 9.72% of the reads, no mutation was observed, 3.13% contained an N, which results from a mixture of bases in the reads.

Duplex sequencing of amplicons – verification of indel frequencies

As part of another project, which involves the analysis of indel heteroplasmies in human mtDNA, three different libraries were prepared containing a 9-nucleotide indel, that is either present at two or three copies (see **Figure 9**). The aim of using duplex sequencing was to confirm an observed repeat-frequency distribution of this indel. Library 1 was prepared from a 685/676 bp PCR product containing the indel, amplified from a long-range PCR product of mtDNA. Library 2 was prepared from the long-range PCR product directly, which was sheared to ~550 bp before adapter ligation. For library 3, the region of the shorter PCR product was synthesized with three or two copies, mixed at the expected ratios, and denatured and re-hybridized to mimic equal conditions as with PCR.

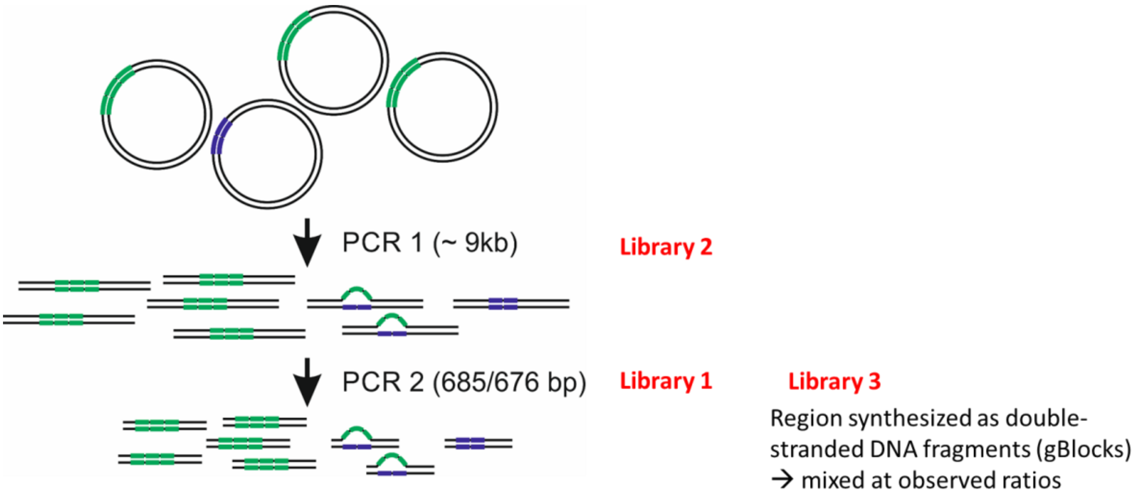

PCR 1 (~ 9kb)
Library 2
PCR 2 (685/676 bp)
Library 1
Library 3
Region synthesized as double-stranded DNA fragments (gBlocks)
→ mixed at observed ratios

Figure 9. Libraries prepared for the analysis of a mitochondrial indel. The mtDNA indel region was first amplified in a long-range PCR from total DNA, which is necessary to reduce any bias coming from homologous regions present in genomic DNA. A second PCR was made from the long-range PCR resulting in a shorter product, by which we could ensure that the whole indel was within one read in the sequencing results, and therefore to be able to obtain the copy number. As a control of the method for getting reliable copy number frequencies, a third library was prepared from synthetic DNA mixed at the expected rations of the copy numbers.

All three described duplex sequencing libraries could be successfully prepared and sequenced. The analysis is still in progress. Since the analyzed DNA are PCR products, we had to analyze SSCS instead of DCS. This was necessary since after amplifying from mtDNA, where some mitochondrial genomes have two, and others three copies, we end up with double stranded DNA that has either two or three copies in both strands, but there is also the possibility of hybrid formation (with two copies in one strand, and three in the other strand) which cannot be analyzed as duplexes.

Preliminary results of library 1 (short PCR product) again show the need for the development of a reference free duplex sequence analysis. When using a reference that contains three copies of the indel, we get a coverage of ~678,000x at the indel region containing mainly SSCS with three copies, but also some with two (see **Figure 10**). However, when using a reference with two copies, the indel region is only covered ~44,700x, mainly with SSCS with two copies, but also some with tree. Some reads were therefore present in both analyses, but the majority only in one. Libraries 2 and 3 have not been analyzed yet.

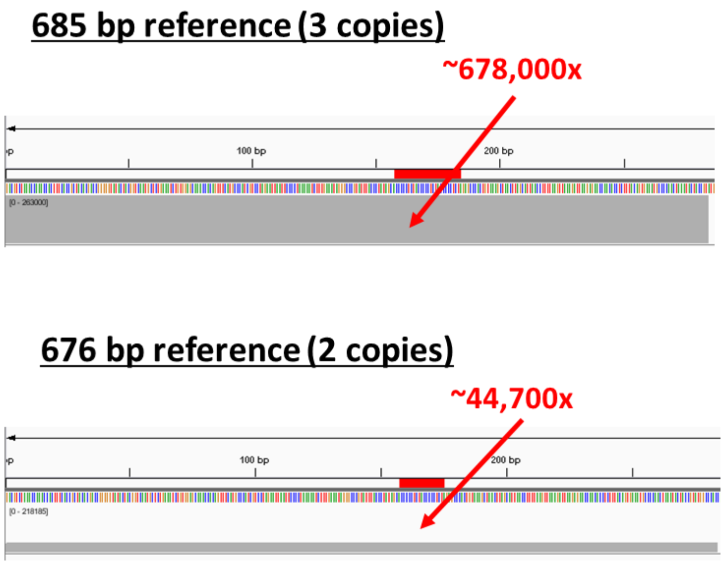


Figure 10. Dependency of the coverage on the reference sequence. When using a reference that contains three copies at the indel, the region is covered ~678.000x, when using a reference that contains two copies at the indel, the region is covered ~44.700x. The higher coverage of the three copies represents the higher frequency at mtDNA level.

To summarize, during my research stay at the Pennsylvania State University, I could successfully learn how to prepare duplex sequencing libraries, how to perform a sequencing run using an Illumina MiSeq platform, and how to perform a basic analysis of the data. Duplex sequencing could be performed for a variety of samples, including human genomic DNA. Data analysis is still in progress. Two publications containing duplex sequencing data generated during my research stay are currently in preparation:

Arbeithuber, B. and I. Tiemann-Boege. "*Analysis of sequencing artifacts resulting from DNA mismatches and lesions in single molecule PCR*." Nucleic Acids Research

Stoler N., M.S. Su, **B. Arbeithuber**, W. Guiblet, B. Rebolledo-Jaramillo, K. Makova, and A. Nekrutenko "*Heteroplasmic mosaicism at mitochondrial indel sites*."

References

Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. 2011. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**(12): 1691-1692.

Beard WA, Batra VK, Wilson SH. 2010. DNA polymerase structure-based insight on the mutagenic properties of 8-oxoguanine. *Mutat Res* **703**(1): 18-23.

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. 2010. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* **Chapter 19**: Unit 19 10 11-21.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5**(6): 435-445.

Fungtammasan A, Ananda G, Hile SE, Su MS, Sun C, Harris R, Medvedev P, Eckert K, Makova KD. 2015. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Res* **25**(5): 736-749.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J et al. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**(10): 1451-1455.

Goecks J, Nekrutenko A, Taylor J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**(8): R86.

Jayaprakash AD, Benson EK, Gone S, Liang R, Shim J, Lambertini L, Toloue MM, Wigler M, Aaronson SA, Sachidanandam R. 2015. Stable heteroplasmy at the single-cell level is facilitated by intercellular exchange of mtDNA. *Nucleic Acids Res* **43**(4): 2177-2187.

Kennedy SR, Salk JJ, Schmitt MW, Loeb LA. 2013. Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. *PLoS Genet* **9**(9): e1003794.

Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, Prindle MJ, Kuong KJ, Shen JC, Risques RA et al. 2014. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9**(11): 2586-2606.

Kunkel TA. 1984. Mutational specificity of depurination. *Proc Natl Acad Sci U S A* **81**(5): 1494-1498.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**(21): 2987-2993.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**(5): 589-595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Lindahl T, Ljungquist S, Siegert W, Nyberg B, Sperens B. 1977. DNA N-glycosidases: properties of uracil-DNA glycosidase from Escherichia coli. *J Biol Chem* **252**(10): 3286-3294.

Ravanat JL, Douki T, Duez P, Gremaud E, Herbert K, Hofer T, Lasserre L, Saint-Pierre C, Favier A, Cadet J. 2002. Cellular background level of 8-oxo-7,8-dihydro-2'-deoxyguanosine: an isotope based method to evaluate artefactual oxidation of DNA during its extraction and subsequent work-up. *Carcinogenesis* **23**(11): 1911-1918.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**(1): 24-26.

Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, Radich JP, Loeb LA. 2015. Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat Methods* **12**(5): 423-425.

Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. 2012. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* **109**(36): 14508-14513.

Tiemann-Boege I, Curtis C, Shinde DN, Goodman DB, Tavare S, Arnheim N. 2009. Product length, dye choice, and detection chemistry in the bead-emulsion amplification of millions of single DNA molecules in parallel. *Anal Chem* **81**(14): 5770-5776.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. *Trends Genet* **30**(9): 418-426.

Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**(5): 614-620.