The source and fate of mitochondrial DNA mutations using high-sensitivity next-generation sequencing technologies

INAUGURAL-DISSERTATION

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zur Köln

vorlegt von

MARITA ISOKALLIO

aus Huittinen

Berichterstatter:	Dr. Dario Valenzano
	Prof. Dr. Andreas Beyer
Tag der mündlichen Prüfung:	14.12.2017

"It's a dangerous business [...] going out your door.
You step onto the road, and if you don't keep your feet,
there's no knowing where you might be swept off to."

- J.R.R. Tolkien -

"Science [...] is made up of mistakes,
but they are mistakes which it is useful to make,
because they lead little by little to the truth."

- Jules Verne -

TABLE OF CONTENTS

I ABBREVIATIONS

AF	Allele frequency
CP	Codon position
CRT	Cyclic reversible termination
CSB I–III	Conserved sequence blocks 1, 2 and 3
CsCl	Cesium chloride
D-loop region	Triple-stranded structure at mtDNA control
dsDNA	Double-stranded DNA
E7.5, E14	Embryonic day 7.5 or 14
EDTA	Ethylenediaminetetraacetic acid
ETAS I–II	Extended termination associated sequences 1 and 2
EtBr	Ethidium bromide
ExoV	DNA exonuclease V
F1	Founder mouse of a female mouse lineage, MKO genotype
F1 score	Harmonic mean of precision and sensitivity
FP	False positive variant
gDNA	Genomic DNA (including both nuclear and mitochondrial DNA)
HSP	Heavy-strand transcription promoter
H-strand	Heavy-strand of mtDNA (based on GC-content)
LSP	Light-strand transcription promoter
L-strand	Light-strand of mtDNA (based on GC-content)
MKO	Hemizygote mtDNA mutator mouse, genotype *PolgA*$^{D275A/KO}$
mtDNA	Mitochondrial DNA
mtDNA-seq	The optimized mtDNA enrichment and sequencing approach
MTERF1	Mitochondrial transcription termination factor 1

mtRNA	Mitochondrial RNA
NCR	Non-coding region, mtDNA control region
N1, N2, N3	F1 female offspring lines, carry mtDNA variants, WT genotype
nDNA	Nuclear DNA
NuMTs	Nuclear sequences of mitochondrial origin
NZB	Wild-type mouse carrying mtDNA from NZB mouse
OriL, OriH	Origin of replication of light- (L) and heavy- (H) strand of mtDNA
pAM1	The entire mtDNA cloned into pACYC177-vector backbone
PCS	Post-PCR cloning and sequencing
PGC	Primordial germ cell
POLG	Mitochondrial DNA polymerase γ
PolgA	Mitochondrial DNA polymerase γ catalytic subunit A
POLRMT	Mitochondrial RNA polymerase
pp	Percentage points
PPV	Positive predictive value, precision
R1, R2	Read 1 and read 2 obtained by paired-end sequencing
RCA	Rolling circle amplification (multiple displacement amplification)
RMC	Random mutation capture
SD	Standard deviation
smPCR	Single-molecule PCR
SNA	Single nucleotide addition
ssDNA	Single-stranded DNA
TFAM	Mitochondrial transcription factor A
TFB2	Mitochondrial transcription factor B2
TP	True positive variant
TPR	True positive rate, recall, sensitivity
TWINKLE	Replicative mitochondrial helicase
UMI	Unique molecular identifier
WT	Wild-type mouse, genotype *PolgA*$^{WT/WT}$

II ZUSAMMENFASSUNG

Pathogene Mutationen in der mitochondrialen DNA (mtDNA) sind dafür bekannt, mehrere Erbkrankheiten zu verursachen. Aufgrund fehlender Methoden zur transgenen Manipulation der mtDNA ist es kaum möglich, die mtDNA Sequenzen und Funktionen zu untersuchen. Die mtDNA-Mutator-Maus wird als Sättigungsmutagenese-Modell verwendet, um eine hohe Variantenbelastung innerhalb der mtDNA zu erzeugen. Bisher wurde bei diesem Modell gezeigt, dass der OriL essentiell für die mtDNA-Replikation ist und dass eine starke negative Selektion potentiell schädlicher mtDNA-Mutationen in der Keimbahn stattfindet. Traditionell wurden mtDNA-Mutationen anhand von Sanger-Sequenzierung oder Post-PCR-Klonierung und Sequenzierung nachgewiesen. Diese Methoden können allerdings nicht das gesamte mtDNA-Genom darstellen. Zudem sind sie aufwendig, teuer und nicht sensibel genug. Seit einigen Jahren werden Hochdurchsatz-Sequenzierungsverfahren als billigere Ansätze verwendet, um mtDNA-Varianten über das gesamte Genom zu detektieren. Allerdings werden diese, wegen ihrer hohen Fehlerrate, als ungeeignet zum Detektieren von Varianten angesehen. Im Gegensatz dazu sind empfindlichere Hochdurchsatz-Sequenzierungsmethoden wie Duplex-Sequenzierung mit einem hohen Arbeitsaufwand verbunden, da sie eine umfangreiche Optimierung und eine hohe Sequenzierungstiefe erfordern. Dadurch erhöhen sich die Kosten auf ein unerschwingliches Niveau.

In dieser Arbeit werden verschiedene Mitochondrienanreicherungs- und Amplifikations-methoden untersucht, um mtDNA frei von nuklearer DNA-Kontamination anzureichern. Es wird eine Standard Illumina HiSeq Sequenzierung genutzt. Die Datenanalyse wird sorgfältig für das mtDNA-Genom optimiert, da dessen Eigenschaften sich von denen des Kerngenoms unterscheiden. Schließlich wird das optimierte mtDNA-Anreicherungs- und Sequenzierungs-protokoll, mtDNA-seq, unter Verwendung einer Titration von Spike-In Proben, welche bekannte mtDNA-Varianten besitzen, validiert. Mit mtDNA-seq ist es möglich,

mtDNA-Varianten zuverlässig zu detektieren. MtDNA detektiert sogar Mutationen unterhalb einer Allelfrequenz von 0,05%. Dies ist etwa zehnmal niedrigerer als die allgemein angewandte Varianten-Nachweisgrenze.

Die optimierte mtDNA-seq wird angewendet, um noch offene mitochondriale Probleme zu adressieren. Es wird das Variantenprofil des gesamten mtDNA-Genoms erzeugt, wodurch mehrere komplette Mutations-Coldspots innerhalb von Kontrollbereichen der mtDNA entdeckt werden. Diese bisher unbeschriebenen Coldspots könnten potenzielle Regulationsorte für die mtDNA-Replikation und die replikations-assoziierte Transkription sein. Die molekulare Mechanismen hierfür sind bisher ebenfalls unbekannt. Zur Untersuchung der Entwicklungs-stufen und des Mechanismus der negativen Selektion wird die hemizygote mtDNA-Mutator-Maus verwendet, um mtDNA-Varianten in weibliche Linien zu isolieren. Da mtDNA-seq die Detektion extrem seltener mtDNA-Varianten ermöglicht, können neue Ergebnisse gewonnen werden, welche den bisherigen Wissensstand zur starken negativen Selektion in der N2-Generation von Mäusen erweitern. Die Ergebnisse deuten darauf hin, dass jede mtDNA-Variante zufällig auf die Nachkommen übertragen werden kann. Jedoch scheinen sich die schädlichsten Mutationen nicht klonal auszubreiten, nicht einmal in N1-Generation. Um die mtRNA-Verarbeitung zu verstehen, wird in einer Pilotstudie der amplicon-Sequenzierungsansatz verwendet. Ziel dieser Studie ist es, Allel-Mismatches zwischen mtDNA und mtRNA-Varianten zu detektieren und dadurch auf mögliche mtRNA-Verarbeitungsdefekte hinzuweisen.

III ABSTRACT

Pathogenic mutations in mitochondrial DNA (mtDNA) are known to cause numerous inherited diseases. However, the lack of methods to transgenically manipulate the mtDNA limits the possibilities to learn about mtDNA sequence and function. The mtDNA mutator mouse is used as a saturation mutagenesis model to generate high variant load into mtDNA. With this model, it has been previously shown that, for instance OriL is essential for mtDNA replication or that strong purifying selection of potentially deleterious mtDNA mutations takes place in the germ line. Traditionally, mtDNA mutations have been detected by Sanger sequencing or post-PCR cloning and sequencing, which are unable to represent the entire mtDNA genome, and are laborious, expensive, or of low sensitivity. More recently, high-throughput sequencing methods have been utilized as cheaper approaches to detect mtDNA variants over the entire genome. However, the high error-rate of these technologies is considered as a limiting factor regarding variant detection sensitivity. On the other hand, high-sensitivity high-throughput sequencing methods, such as Duplex Sequencing, are often laborious requiring extensive optimization and high sequencing depth, ultimately raising the costs to a prohibitive level.

In this thesis, various mitochondria enrichment and amplification methods are explored in order to enrich mtDNA free from nuclear DNA contamination. Standard Illumina HiSeq sequencing is utilized and data analysis steps are carefully optimized to be suitable for mtDNA genome, which has characteristics very different from the nuclear genome. Finally, the optimized mtDNA enrichment and sequencing protocol, mtDNA-seq, is validated utilizing a titration of spike-in samples harboring known mtDNA variants. With mtDNA-seq it is possible to detect mtDNA variants reliably even below allele frequency of 0.05 %, which is approximately ten times lower variant detection threshold than what has been generally applied in other studies.

The optimized mtDNA-seq is applied to address open mitochondrial

biology research questions. The variant profile of the entire mtDNA genome is generated, and several complete mutational coldspots are discovered at the control region of the mtDNA. These novel coldspots are hypothesized to be potential regulation sites for mtDNA replication and replication-associated transcription by as-yet-unknown molecular mechanisms. To clarify the developmental stage and mechanism of purifying selection, hemizygote mtDNA mutator mouse is utilized to isolate mtDNA variants into female lineages. As it is possible to detect extremely rare mtDNA variants by mtDNA-seq, these new results expand the previous study showing strong purifying selection by N2 generation of mice. The results suggest that by chance any mtDNA variant may be transmitted to the offspring, however, the most deleterious mutations do not seem to clonally expand even in N1 generation mice. To understand the mtRNA processing, amplicon sequencing approach is utilized in a preliminary study. The aim in this study is to detect allelic mismatches between mtDNA and mtRNA variants, which potentially indicate mtRNA processing defects.

1 REVIEW OF THE LITERATURE

1.1 Mitochondrial genetics and disorders

1.1.1 Mitochondria

According to the widely accepted endosymbiotic theory of the origin of mitochondria, mitochondria evolved from an α-proteobacterium engulfed by an archaeon (reviewed by Martin et al. 2015). The symbiotic relationship was thought to be based on the ability of the proteobacterium to efficiently produce ATP in exchange of carbohydrates produced by the host. Over the course of evolution, such symbiosis lead to formation of mitochondria – cell organelles, which are known as the powerhouses of the cell, but mitochondria also contribute to many other cell functions (reviewed by Nunnari & Suomalainen 2012). Mitochondria are dynamic in nature, forming a fusion-fission network, the steady state of which varies between different cell types: in cardiomyocytes, the network consists of connected tubular structures, which are distributed throughout the cell, whereas in oocytes mitochondria network is more localized as fragmented aggregates. Such a system requires delicate control of fusion, fission, positioning and motility, and it is enabling the cells to respond variable energy demands (reviewed by Labbé et al. 2014).

The energy production (i.e. oxidative phosphorylation system) is located inside a double-membrane structure, in the inner membrane and mitochondrial matrix. The respiratory chain is formed by four respiratory complexes, coenzyme Q and cytochrome *c*, which receive electrons from the citric acid cycle. The electrons go through a series of reduction and oxidation reactions resulting in proton transfers across the mitochondrial inner membrane. This proton gradient drives the ATP production and is an essential part of mitochondrial function. The key proteins required for oxidative phosphorylation are encoded in mitochondrial genome (mitochondrial DNA, mtDNA, reviewed by Larsson 2010). During their evolution, most of the original bacterial genetic material has been transferred to the nucleus, and novel genes

required for mitochondrial function have evolved. Mitochondria still do harbor their own tightly packed circular mtDNA molecules (**Fig. 2.1**) – size of which varies from ~6 kb in *Plasmodium falciparium* to ~16 kb in mammals or even >200 kb in some plants (reviewed by Gray 2012).

The DNA strands of the mammalian mitochondrial genome can be categorized into heavy- and light-strands (H- and L-strand) according to their density separation, as H-strand is more G-rich than L-strand. H-strand is the sense strand for 12 proteins, the core components of the mitochondrial respiratory complexes, and the two rRNAs as well as 14 tRNAs required for mitochondrial translation machinery, whereas L-strand is the sense strand for only one protein and 8 tRNAs. The rest of the >1000 proteins required for oxidative phosphorylation, mtDNA replication and expression, mitochondrial protein synthesis, iron-sulfur cluster synthesis or other metabolic functions are encoded in the nuclear genome, translated in the cytosol and transported to mitochondria (**Fig. 2.1**, reviewed by Larsson 2010). This bi-genomic system requires well-coordinated expression of nuclear and mitochondrial genes.

The mitochondria network harbors thousands of mtDNA molecule copies. At a given time, a proportion of mtDNA molecules are naked, but most of them are packed into nucleoid structures (Farge et al. 2014). The exact nucleoid composition is still a debated topic. Studies with mouse embryonic fibroblasts (MEFs) suggest the nucleoid typically consists of mitochondrial transcription factor A (TFAM) protein and a single copy of mtDNA (Kukat et al. 2011). Approximately 1000 TFAM proteins coat a single mtDNA molecule (i.e. one TFAM every ~16 bp [(Kukat et al. 2011]), and TFAM can bind mtDNA in a single, co-operative and cross-strand fashion *in vitro* (Kukat et al. 2015). Binding of human TFAM to mtDNA is sequence specific at the promoter regions but non-specific elsewhere (Fisher & Clayton 1988; Fisher et al. 1992), and furthermore, human TFAM is shown to bend the mtDNA 180° (Ngo et al. 2011; Rubio-Cosials et al. 2011). All these qualities explain the capability of TFAM to efficiently compact the mtDNA molecule into a nucleoid structure of ~100 nm in size (Kukat et al. 2011).

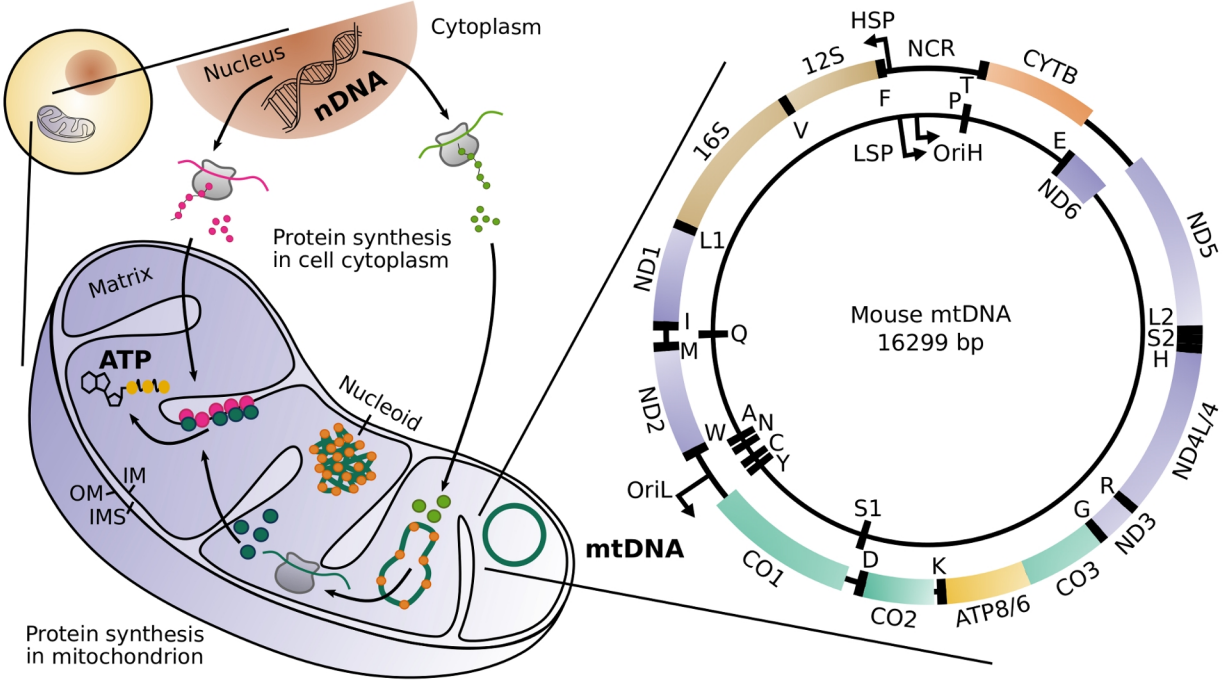

Nucleus
nDNA
Cytoplasm
Protein synthesis
in cell cytoplasm
Matrix
ATP
Nucleoid
IM
OM
IMS
Protein synthesis
in mitochondrion
mtDNA
HSP
12S
NCR
CYTB
16S
V
F
P
T
LSP
OriH
E
ND6
ND5
ND1
L1
I
M
Q
Mouse mtDNA
16299 bp
L2
S2
H
ND2
W
A
N
C
Y
ND4L/4
OriL
S1
R
G
ND3
CO1
D
K
CO3
CO2
ATP8/6

Figure 2.1. Mitochondrial function and genome. *Mitochondria are cell organelles, which consist of outer and inner membranes (OM and IM) enclosing inter-membrane space (IMS) and forming the inner matrix of mitochondria. Mitochondria produce ATP by oxidative phosphorylation, which is dependent on proteins encoded in nuclear genome (nDNA) as well as in mitochondrial genome (mtDNA). Approximately >1000 proteins are synthesized in the cell cytoplasm (pink and light green) and imported to mitochondria not only for oxidative phosphorylation but also for other mitochondrial functions e.g. iron-sulfur cluster synthesis, cell signalling and, of course, for mtDNA maintenance, such as compacting mtDNA into nucleoids (orange proteins) or mtDNA replication and transcription (light green proteins). Only 13 proteins are encoded in mtDNA, and they are essential parts of the respiratory complexes (components of each complexes are denoted with different colors: complex I, ND1–6, purple; complex III, CYTB, orange; complex IV, CO1–3, green; complex V, ATP6 and 8, yellow). Furthermore, mtDNA encodes the two rRNAs and 22 tRNAs required for mitochondrial translation. The base composition of mtDNA is biased, and the different strands are called as heavy- and light-strands (H- and L-strands) according to their densities. Both strands harbor their own origin of replication (OriH and OriL) as well as transcription promoters (LSP and HSP). The densly packed mtDNA contain only one major non-coding region (NCR) also known as control region. The illustration is based on Gustafsson et al. (2016).*

Others have additionally suggested the nucleoid to contain also mtDNA replication proteins (e.g. mitochondrial DNA polymerase γ (POLG), mitochondrial single-stranded DNA binding protein (mtSSB), mitochondrial helicase TWINKLE as well as mitochondrial inner membrane proteins) (reviewed by Gilkerson 2009). This is suggested to hold the nucleoid tethered to the inner membrane rather than freely floating nucleoids, providing a possible model how mtDNA and nucleoids are segregated between the dynamic network of mitochondria and to the daughter organelles (Gilkerson 2009). In line with this, mitochondria and endoplasmic reticulum interact at certain contact sites which direct mitochondrial division and nucleoid segregation (reviewed by Labbé et al. 2014). Moreover, the level of mtDNA compaction into nucleoids may function as a regulator for mtDNA replication and transcription (Farge et al. 2014).

Different from nucleus of the same cell, mtDNA are replicated through a relaxed replication process independent from the cell cycle. Despite a relatively low error rate of POLG, $5.6x10^{-7}$ mut/bp/doubling (Zheng et al. 2006), each mtDNA molecule will go through many more rounds of replication, which increases the probability of variant-harboring mtDNA copies and higher per gene substitution rate in comparison to nuclear DNA (nDNA). Different mtDNA molecules, harboring distinct variants, can simultaneously exist within a cell – a condition called as heteroplasmy. Through vegetative segregation, variable proportions of mitochondria harboring different mtDNA molecules may end up to the daughter cells (**Fig. 2.2a**). It has been also shown *in silico* that random drift may drastically shift the levels of different mtDNA molecules during the life time of a human being when some mtDNA molecules are clonally expanded and others are not (Chinnery & Samuels 1999, **Fig. 2.2b**). According to their model, it is an effective factor in avoiding a pathogenic allele from fixing (i.e. becoming homoplasmic), however, rare variants may also expand to relatively high levels. If the relative level of a pathogenic mtDNA mutation frequency reaches a critical biochemical threshold, defects in mitochondrial function can be detected (**Fig. 2.2**, Durham et al. 2007). Moreover, if the functional defect is

compensated by more mtDNA replication, relative levels of wild-type mtDNA molecules may drop, by chance, even more (Chinnery & Samuels 1999).

1.1.2 Mitochondrial disorders

It can be easily understood that disruptions at any level of such complex systems as mitochondria, can have severe consequences. Indeed, mitochondrial dysfunction is associated with variety of heterogenous, inherited human disorders as well as common diseases such as neurodegenerative disorders or metabolic syndromes (Nunnari & Suomalainen 2012). In United Kingdom, it has been estimated that one in 4300 adults are affected by mitochondrial disorders, making them one of the most common group of inherited neurological disorders (Gorman et al. 2015). Furthermore, even one in 200 healthy individuals are estimated to be carrier of certain pathogenic mtDNA mutations (Elliott et al. 2008).

Pathogenic mutations or deletions in nuclear genes encoding mitochondrial proteins are known to cause of several mitochondrial disorders, such as mutations or deletions in *Polg* causing progressive external ophthalmoplegias (PEO, van Goethem et al. 2001). These nuclear-gene derived disorders follow Mendelian rules and are relatively well known, however a typical characteristic for a mitochondrial disorder is that the exact same mutation may cause variable symptoms or onset ages between individuals. Disorders caused by mtDNA mutations or deletions are even more complex than the ones of chromosomal origin. The transmission of mtDNA in mammals is solely maternal and the multicopy nature and variable levels of mutation present in the mtDNA molecules further complicate the interpretation of the relationship between the mutation and the disorder symptoms of an individual. For example, myoclonic epilepsy and ragged-red fiber disease (MERRF) is caused by a mutation in mt-tRNA Lys (K) and although the relative mutation frequency mostly correlates with the phenotype, the disease onset age varies and as little as 15 % presence of

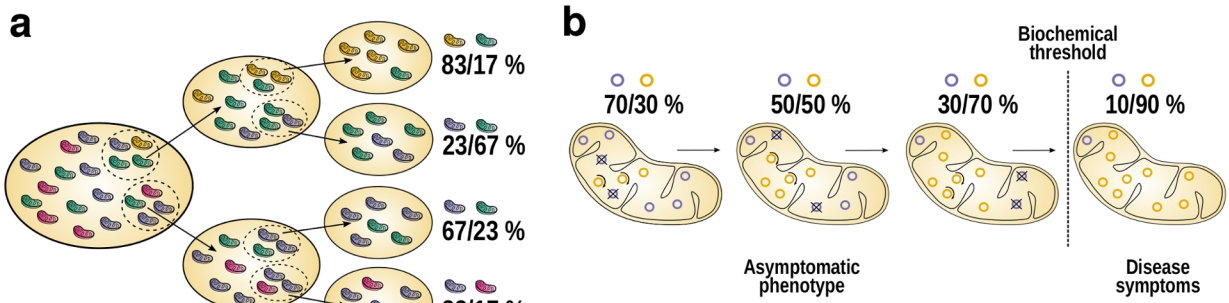


Figure 2.2. Vegetative segregation and relaxed replication. *As a mitochondrion harbors thousands of mtDNA molecules, it is likely that some of them harbor variants. Thus, a cell may carry normal and mutated mtDNA molecules, a condition called as heteroplasmy. During cell division, mitochondria are divided to daughter cells through vegetative segregation (**a**). Each daughter cell may then contain very variable proportions of different mtDNA molecules. Furthermore, mtDNA molecules go through constant turnover, i.e. relaxed replication (**b**), and no mechanism exist to ensure each molecule is replicated. Thus, proportion of different mtDNA molecules may rapidly shift over time. If a pathogenic mtDNA mutation reaches high levels (past a biochemical threshold), disease symptoms occur due to defects in respiratory chain system. The illustration is based on Stewart & Chinnery (2015).*

wild-type mtDNA molecules was enough for one individual to escape disease symptoms (Shoffner et al. 1990). On the other hand, a single mutation on mt-tRNA Leu (L1) can appear as different disorders: mitochondrial encephalopathy lactic acidosis and stroke-like episodes (MELAS), maternally inherited deafness and diabetes (MIDD), and progressive external ophthalmoplegia (PEO, Nesbitt et al. 2013). This variability makes clinical genotype-phenotype assessments very difficult for novel pathogenic mtDNA mutations. Moreover, it challenges the understanding of the mechanistic details of the disease progression.

Severity and the onset of a mitochondrial disorder may directly be affected by the segregation of heteroplasmic mtDNA molecules to daughter cells, or via germ line to the offspring. Although, random drift is the most dominant factor, selection of certain mtDNA molecules, even if harboring a phenotypically neutral variant, has been shown to take place in a tissue- or mutation-specific manner in somatic cells (e.g. Jenuth et al. 1997; Pyle et al. 2007). How exactly replication of certain mtDNA molecules, mtDNA compaction into nucleoids, nucleoid clustering or dynamic fusion-fission network affect the segregation or turnover of different mtDNA molecules is still not fully understood (reviewed by Jokinen & Battersby 2013).

1.1.3 Models for mitochondrial DNA mutation research

There are no curative treatments available for mtDNA disorders. The current approaches only aim to maintain the health of the patient, and recently also to avoid inheritance of pathogenic mtDNA mutations by mitochondrial replacement therapies or preimplantation diagnostics (Poulton & Bredenoord 2010, Chinnery et al. 2014). In contrast to nuclear genome, studies on mtDNA variants are extremely difficult since there is no method for mitochondrial reverse genetics *in vivo* to study and confirm genotype-to-phenotype causation (reviewed by Patananan et al. 2016). To study different mtDNA disorders, the solution has been to utilize variety of mouse models, which are generated by direct introduction of existing mtDNA mutations by cytoplasmic fusion strategy or indirectly by modifying the nuclear genes which can affect

the mtDNA composition, such as *Polg* or *Twinkle* (reviewed by Tyynismaa & Suomalainen 2009).

The most relevant mouse model to this thesis is the mtDNA mutator mouse (Trifunovic et al. 2004; Kujoth et al. 2005). Trifunovic et al. (2004) created a homozygote knock-in mouse expressing proof-reading deficient POLG, in which the critical aspartate residue of exonuclease domain in *PolgA* was replaced with alanine (*PolgA*^{D257A}). These mice showed normal replication efficiency, but the exonuclease activity was significantly reduced (Trifunovic et al. 2004) showing a mtDNA-specific mutation load of $6.6x10^{-4}$ mut/bp (~40x the background mutation rate, Ross et al. 2013) as well as notable amount of truncated, linear mtDNA molecules (Trifunovic et al. 2004, Macao et al. 2015). After ~25 weeks of age, these mice begin to show ageing symptoms like kyphosis, alopecia, decreased body fat, osteoporosis, anemia, and reduced fertility. Moreover, their median lifespan is only ~48 weeks (Trifunovic et al. 2004).

As a saturation mutagenesis model, the mtDNA mutator mouse has proven to be valuable tool in addressing research question related to mtDNA biology. Recently, the heterozygote mtDNA mutator mouse was utilized to establish a new mouse model which harbored a mt-tRNA Ala (A) variant and presented with mitochondrial disorder phenotype (Kauppila et al. 2016). Other studies have utilized the mtDNA mutator mouse model e.g. to address the effect of mtDNA mutations on ageing (Vermulst et al. 2007; Edgar & Trifunovic 2009; Williams et al. 2010; Ameur et al. 2011; Ross et al. 2013; Baines et al. 2014), to study mtDNA transmission (Stewart et al. 2008a; Ross et al. 2013; Ross et al. 2014), or to understand various processes involved in mtDNA maintenance (Hance et al. 2005; Wanrooij S. et al. 2012; Baines et al. 2014; Macao et al. 2015).

1.1.4 Mitochondrial DNA variant detection by traditional methods

Many mtDNA mutation studies simply focus on diagnostic detection of near-homoplasmic or high-frequency, clonally expanded mtDNA

mutations or rearrangements. This is because the diseases symptoms often occur only after the levels of pathogenic mtDNA reach a relatively high threshold, often ~70–90 % (Durham et al. 2007). Detection of such high-level mutations is relatively straightforward with PCR- or blotting-based methods (reviewed by Moraes et al. 2003) or by Sanger sequencing, which has a detection threshold of ~15–30 % (Hancock et al. 2005; Rohlin et al. 2009). However, especially with ageing studies, there is interest in detecting the total variant load of a tissue, including *de novo* mutational events, which are not yet highly clonally expanded, and are difficult to detect with the mentioned methods. Traditionally three methods have been used to sensitively measure the total mtDNA variant load of a tissue: post-PCR cloning and sequencing (PCS), single-molecule PCR (smPCR) and random mutation capture assay (RMC, compared by Greaves et al. 2009).

In PCS, the target DNA is first amplified by high-fidelity PCR, then cloned into a vector and single clones are expanded and sequenced by Sanger sequencing. As reviewed by Kraytsberg and Khrapko (2005), the advantages of PCS are fast amplification of the target mtDNA without mitochondria isolation and utilization of easy-to-use commercial kits for cloning and even robotics for plasmid purification (Kraytsberg & Khrapko 2005). Furthermore, as mtDNA is very small genome, ~16 kb, it is even possible to analyze the entire mtDNA genome by lambda-phage based PCS (Ross et al. 2013, Hagström et al. 2014). One potential drawback, however, is propagation of PCR-errors which are indistinguishable from genuine variants; often used high-fidelity DNA polymerase *Pfu* introduces 1.6×10^{-6} errors per nucleotide per cycle (Lundberg et al. 1991) or even less, down to error rate of 4.4×10^{-7} with the engineered enzymes like Phusion® (New England Biolabs, Inc.). Furthermore, DNA polymerases have a tendency to jump between templates, which is an issue in highly mutated samples, where template switching would create new combinations of mutated molecules (Hagström et al. 2014). PCR-step may also introduce bias by preferential amplification of one but not another allele (Kraytsberg & Khrapko 2005).

To overcome the issues of PCS, smPCR was suggested as more accurate method to study mtDNA variant loads (Kraytsberg et al. 2008). In smPCR, the source DNA is serially diluted until only a fraction of PCR reactions amplify an mtDNA product. The key assumption is that this way only a single molecule is analyzed, and thus, PCR-errors are easily distinguished as a heteroplasmic peak in the sequencing. Despite overcoming the disadvantages of PCS, smPCR introduces some new issues. First of all, the method requires optimization of the PCR to succeed from such low amount of template DNA. This also causes another major drawback – highly increased risk of sample contamination (Kraytsberg & Khrapko 2005). Also, large numbers of samples are required for the serial template dilutions (Greaves et al. 2009).

Another method to sensitively measure mtDNA variant loads without the risk of PCR-induced errors, is RMC. The method is based on restriction digestion of wild-type DNA prior to quantitative PCR, thus, only molecules harboring a mutated restriction site will be amplified. An important step in this method is the quantification of the starting material in order to be able to determine the final variant load (Greaves et al. 2009). RMC revealed lower mtDNA variant loads than the other methods (Vermulst et al. 2007; Greaves et al. 2009), which was suggested to indicate that RMC is more sensitive and effectively diminishing the PCR-errors as artefacts. Furthermore, RMC is insensitive to DNA damage, such as oxidized deoxyguanosine (8-oxo-dG), which would be mistakenly paired with adenosine by DNA polymerases (Shibutani et al. 1991), and thus, artificially detected as a fixed GC>TA variant. Again, sensitivity to DNA damage is a potential factor increasing the variant loads observed by PCR-based methods in comparison to RMC (Vermulst et al. 2007). A major acknowledged drawback of RMC is that the result may not be an accurate reflection of the total variant load of the entire mtDNA genome because the target restriction site represents only very short part of the genome (Vermulst et al. 2007; Greaves et al. 2009), thus RMC might not detect clonally expanded variants which are rare across the mtDNA genome (Greaves et

al. 2014).

In comparison to the published literature, Vermulst et al. (2007) detected very low mtDNA variant load in young wild-type mice, on average $6.0x10^{-7} \pm 0.9x10^{-7}$ mut/bp, whereas older mice showed $1.1x10^{-5} \pm 0.3x10^{-5}$ mut/bp (Vermulst et al. 2007). Meanwhile, Ross et al. (2013) detected $2.0x10^{-5}$ mut/bp in WT mice by PCS and estimated the method error rate to be $<3.5x10^{-6}$ mut/bp (Wanrooij S. et al. 2012, Ross et al. 2013). These results suggest that, despite the discussed disadvantages, these methods are very sensitive approaches to measure the mtDNA variant load. However, they require extensive optimization or hands-on time and fail to represent the entire mtDNA genome or the costs become prohibitive. Thus, over the recent years the focus has turned to deep sequencing technologies.

1.2 Deep sequencing technology overview

For over a decade already, advancements in various deep sequencing technologies have decreased the costs significantly and replaced earlier high-throughput methods, like microarrays, as a routine method used in DNA variant detection. These technologies are often referred to with variable umbrella terms and abbreviations with sometimes confusing, mixed usage: deep sequencing, high-throughput sequencing (HTS), massively-parallel sequencing (MPS), second-, third-, fourth- or next-generation sequencing (NGS), DNA-seq or simply by the company name, who first invented the sequencing platform in question (e.g. SOLiD, IonTorrent, PacBio, MinION or Illumina sequencing). Since this PhD thesis focuses on deep sequencing of mtDNA, hereafter simply the term 'sequencing' is used to refer to deep sequencing unless otherwise mentioned.

This chapter first introduces different sequencing technologies. The reviewed technologies include two long-read sequencing technologies: single-molecule real-time sequencing (SMRT) by Pacific Biosciences of California, Inc. (PacBio) and MinION sequencing by Oxford Nanopore Technologies, as well as different short-read sequencing technologies:

IonTorrent sequencing and SOLiD sequencing currently owned by ThermoFisher Scientific and HiSeq sequencing by Illumina Inc. Comparison of key values such as required input DNA or run time of the different sequencing platforms is summarized in **Table 1.1**. Finally, the main data analysis steps and potential artefacts introduced during sequencing are discussed.

Table 1.1. Comparison of different sequencing platforms. *Input DNA, run time, yields, read lengths and accuracies of different sequencing platforms from the provider's web pages at the time of writing (08/2017).*

Technology	Instrument	Input (ng)	Run time	Yield (per run)	Read length	Accuracy (%)
Sanger	3703xl	1–300	2 h	~96 kb	<1 kbp	99.99
SMRT	PacBio Sequel	10–100	0.5–10 h	5–8 Gb	>20 kb	>99.999
Nanopore	MinION	0.01	Real time	Up to 17 Gb	Up to 200 kb	~92 (with 1D reads)
SOLiD	SOLiD 4	10	12 d	100–300 Gb	35 bp	>99.94
IonTorrent	Ion 318 v2	10	4.4–7.3 h	0.6–2 Gb	200–400 bp	99.99
Illumina	HiSeq 2500	1–100	5 d	~400 Gb	2 x 100 bp	>80 % of reads with accuracy 99.9
Illumina	HiSeq 3000/ 4000	1–50	<1–3.5 d	750–1300 Gb	2 x 150 bp	>75 % of reads with accuracy 99.9

1.2.1 Sequencing library and cluster generation

In short-read technologies, the DNA is first sheared to short fragments, either by mechanical or enzymatic methods. Mechanical methods include commonly used acoustic shearing by Covaris Adaptive Focused Acoustics™ (AFA) technology (COVARIS, Inc.), which, similar to

sonication by Bioruptor® ultrasonicator (Diagenode Inc.), utilizes the acoustic cavitation to fragment the DNA. Another, older, mechanical method is nebulization, in which the DNA is forced through a small hole (Sambrook & Russell 2006). In addition to Covaris shearing, enzymatic fragmentation is often applied, of which the potentially most common method is tagmentation. In tagmentation, the DNA is fragmented and sequencing adapters are simultaneously ligated to the fragment, thus reducing the processing steps (Adey et al. 2010). Another enzymatic approach is developed by New England Biolabs, and it is based on two enzymes; one generates nicks on the double-stranded DNA (dsDNA) and the other enzyme breaks the DNA at the nicked site (NEBNext dsDNA Fragmentase, New England Biolabs). Although claimed to be random, all fragmentation methods may introduce bias (Poptsova et al. 2014) or even artefactual variants (Costello et al. 2013, as discussed later in this chapter).

In contrast, with long-read sequencing technologies, direct usage of long PCR amplicons or restriction digested DNA is possible. However, if fragmentation is required, the long fragments can be achieved by different methods, such as Covaris g-Tube, which can fragment the DNA up to 20 kb fragments based on centrifugal forces (COVARIS, Inc.). Furthermore, long-read sequencing technologies do not need to amplify the template DNA for sequencing, but instead require larger amount of input DNA fragments to which hairpin adapters are ligated either to both ends (SMRT) or only to one end (MinION). This differs from short-read sequencing technologies which generally require a template amplification step after the adapter ligation. Moreover, short-read sequencing technologies reach massive parallelization by amplification of the DNA to form clonal DNA template clusters on a solid surface (reviewed by Goodwin et al. 2016).

Two main techniques are used to form the clonal sequencing clusters – emulsion PCR and solid-phase bridge amplification. The latter is applied by Illumina, whereas the other technologies discussed here utilize the emulsion PCR. In emulsion PCR, as the name already indicates, DNA

fragments are diluted into water-oil droplets to such degree that a droplet contains only a single molecule (Dressman et al. 2003). Each droplet forms a microscale PCR reaction as it also contains primer, dNTPs and DNA polymerase. The droplets are combined with beads covered with primers complementary to the adapter sequence, and when subjected to PCR cycling conditions, the template DNA hybridize with the bead-bound primer leading to extensive amplification of a single template. The complementary strand is dissociated and the bead is left with thousands of single-stranded DNA (ssDNA) templates (Dressman et al. 2003). The beads can then be immobilized into wells or on a glass slide for sequencing.

In the solid-phase bridge amplification technique the DNA template cluster synthesis takes place directly on a slide which has covalently bound primers. One of the primers is complementary to the adapter sequence and ssDNA fragments applied to the slide can hybridize with the primer. After amplification, the original DNA strands are dissociated whereas the newly formed complementary DNA strand, covalently bound to the slide, is hybridized to the other adapter bound on the slide and again amplified – hence the technique name 'bridge amplification'. This way millions of distinct clusters of clonal DNA templates are formed (Illumina, Inc.).

1.2.2 Long-read sequencing technologies

Of the long-read sequencing technologies, SMRT was established almost a decade ago (Travers et al. 2010), whereas MinION sequencing has been in developmental use since 2014 (Jain et al. 2016). In SMRT, or specifically SMRTbell, the sequencing takes place in a miniscule scale multi-well plate with transparent bottoms called zero-mode waveguides (ZMW, Levene et al. 2003). The DNA polymerase is fixed to the bottom of the well, and a single DNA molecule is sequenced in each well. The SMRTbell hairpin adapters allow primer binding to the single-stranded hairpin loop, which is further bound by the polymerase and incorporation of fluorescently labelled dNTPs is monitored by laser and camera from each individual wells (reviewed by Goodwin et al.

2016). The two hairpin adapters circularize the template allowing the strand-displacing DNA polymerase to proceed through the DNA molecule multiple times. Formation of a consensus sequence is efficiently excluding sequencing errors and increasing the accuracy (Travers et al. 2010). At the time of writing, PacBio reports the average read-length to be >20 kb and maximum >60 kb.

Portable MinION (or larger-scale GridION and in the future PromethION) sequencing is commercial sequencing platform based on nanopore technology (Oxford Nanopore Technologies). One end of the DNA fragment is attached to a leader sequence which can attach the DNA to a nanopore, which is located across a membrane. The membrane separates electrolyte solution, and an ionic current is formed when the solution is moved through the nanopore. When the template DNA is moving through the nanopore, changes in the ion current reflect k-mers of the DNA sequence blocking the pore. Similar to SMRT, the hairpin adapter at the other end of the DNA fragment enables sequencing of the both strands and formation of consensus sequence increases the accuracy (Jain et al. 2016). At the time of writing, Oxford Nanopore Technologies promises to reach even >200 kb read-lengths.

1.2.3 Short-read sequencing technologies

Sequencing-by-ligation technology

Sequencing by oligonucleotide ligation and detection (SOLiD) technology is short-read sequencing technology based on short, fluorescently labelled probes in which one or two nucleotides are known and the rest are degenerate bases (reviewed by Voelkerding et al. 2009). The sequencing consists of multiple cycles of series of reactions including competitive annealing of the probe to the template DNA, ligation, removal of unbound probes and reading the fluorescent signal. Once the first round is finished, the formed dsDNA is denatured and sequencing cycling is continued with an offset of one base and this is continued until 35 nucleotides are sequenced multiple times. The resulting sequencing read-out is termed as color space -coding and has to be deconvoluted (Voelkerding et al. 2009).

Sequencing-by-synthesis technologies

Sequencing-by-synthesis technologies include multiple providers and approaches, here, two commonly used platforms, IonTorrent and Illumina, are reviewed. The former applies single-nucleotide addition (SNA) approach, which was also utilized by 454 pyrosequencing (not discussed here). Illumina, as well as Qiagen's GeneReader, utilizes cyclic reversible termination (CRT, reviewed by Goodwin et al. 2016).

IonTorrent sequencing is also termed as semiconductor sequencing and it is the first technology enabling sequencing without optical signal detection (Rothberg et al. 2011). Instead, IonTorrent, as the name indicates, detects the released proton when a dNTP is incorporated. Such detection is enabled by sequential addition and washing of each nucleotide and the H^+ ion is released only when polymerase incorporates the dNTP to the newly synthesized strand. The change in H^+ concentration is then measured as a change in pH value by a sensor including ion-sensitive field-effect transistor (ISFET) and complementary metal-oxide semiconductor (CMOS). The raw voltage signal is then processed to base calls (Rothberg et al. 2011), however, accurate homopolymer sequencing is challenging due to the fact that the same nucleotides are incorporated during single measurement and the change in pH value cannot be exactly measured (Goodwin et al. 2016).

The basic idea of cyclic reversible termination approach is similar to Sanger sequencing in which the dNTP bound to a terminator blocks the elongation. In contrast to SNA, in CRT all four nucleotides are added and only one is incorporated by the polymerase. The labelled dNTP is imaged and the elongation termination is reversed by removal of the fluorophore and blocking group. Earlier Illumina utilized four-channel detection to image the sequence, however, the newer machines use much faster two-channel detection. In two-channel detection, instead of using four dyes, one for each dNTP, mixture is used and for example G is detected as non-labelled cluster (Illumina 2016). Furthermore, since HiSeq3000, the flow cell type is now a patterned flow cell, on which the cluster generation takes place in nanowells. This way the clusters are

evenly spaced decreasing the time required for sequence analysis when the software does not need to predict the cluster location. Moreover, it increases the cluster density on the flow cell, thus decreasing the costs (Illumina 2015).

1.2.4 Key sequencing data analysis steps and their potential artefacts

The accuracy of variant detection lies not only in good sample quality but also on proper data analysis steps, starting from the fluorescence image analysis until the variant filtering thresholds. As reviewed by Nielsen et al. (2011), base-calling algorithms determine the nucleotide content of a read from the read image (SOLiD, Illumina) including a measure of uncertainty called base-call quality score (Nielsen et al. 2011), which is given as Phred score (Ewing et al. 1998) defined as $Q_{Phred} = -10 \log_{10} P$, where P is error probability. Thus, Phred score 30 means that the chance of an incorrect base-call is 0.1 %. Often the sequencing platform provider's base-calling algorithm is used, however, according to Nielsen et al. (2011), development of better base-calling algorithms have improved the error rates up to ~30 %, and for example, earlier extremely high error rates of MinION have been significantly improved by development of MinION compatible tools, including base-calling algorithms and read aligners (Jain et al. 2016).

Very often in Illumina sequencing, the samples are multiplexed in order to reduce the costs as several samples may be sequenced on a single lane. The index sequence is included into the sequencing adapter and it is sequenced after the actual read sequencing in a separate process with a new primer. After base-calling, the multiplexed reads need to be de-multiplexed i.e. separated based on their index sequence (Kircher et al. 2012). Sequencing errors, errors in the synthesis of the oligos or failure in library preparation pose a risk that a read is mistakenly assigned to another sample – a phenomenon called as index hopping or index switching (Illumina 2017; Sinha et al. 2017). The multiplex design is aimed to be complex enough such that several bases need to be erroneous before the indices are misassigned, thus, effectively lowering

the probability of a sequencing error to cause the misassignment. Furthermore, during de-multiplexing only zero or one mismatches are usually allowed in the index sequence. This, nevertheless, does not exclude the possibility that oligos are contaminated during synthesis or library preparation (Kircher et al. 2012), which inevitably leads to indistinguishable indices and misassigned reads.

Index switching has been a known artefact for years (Kircher et al. 2012), and it is unavoidable, causing a low base level of errors. This is not an issue for most sequencing applications (Illumina 2017), but for example, in accurate low-frequency variant detection, genotyping or single-cell RNA-seq it is a significant problem. The switch to patterned flow-cell usage significantly increased the occurrence of index hopping (Illumina 2017), which has recently raised serious concerns (Hadfield 2016, accessed 08/2017; Linck 2017, accessed 08/2017; Sinha et al. 2017), although, no peer-reviewed publication is available on the topic at the time of writing (08/2017). As already suggested by Kircher et al. (2012) as well as noted by Illumina (2017) and also discussed by Hadfield (2016, accessed 08/2017), dual-indexing is an effective solution to avoid index switching. It, however, is currently only possible in 6- or 8-plex combinations, thus, reducing the throughput (Illumina 2017) and require paired-end read mode, eventually increasing the sequencing costs.

Data pre-processing generally includes trimming off of low-quality bases and adapter leftovers. In Illumina reads for example, especially with the earlier chemistries, the base-calling quality heavily decreased towards the end of the read. This was and is due to the enzyme activity and de-phasing (Fuller et al. 2009). De-phasing means that while incorporating new nucleotides to the cluster of probes, due to a missed incorporation or a failed termination reaction the synchronization between different strands within the cluster is lost. This will lead to an ambiguous fluorescence signal from that cluster and decrease the reliability of the base-call (as reviewed by Reinert et al. 2015).

Read alignment is the next fundamental analysis step. The aligner has to

tolerate errors and variants in the read, yet being able to assign it to the most correct location i.e. approximate string matching problem (as reviewed by Reinert et al. 2015). The basic idea in the read alignment is to create string indices either from the reference genome, reads or both, thus, querying of these substrings is much faster than querying the entire data set (Nielsen et al. 2011; Reinert et al. 2015). A commonly utilized data compression algorithm is Burrows-Wheeler transformation (Burrows & Wheeler 1994), for example implemented in Bowtie/Bowtie2 (Langmead et al. 2009; Langmead & Salzberg 2012) and BWA (Li & Durbin 2009), which, based on number of citations, are the most commonly used DNA read aligners (Fonseca, accessed 08/2017). Over the years, a plethora of read aligners have been developed (list of aligners updated in 2015, Fonseca, accessed 08/2017) and the suitability of each aligner should be determined for each application as aligners vary e.g. in their capability to, for instance, introduce long gaps (here termed as splice-aware aligners), to conduct global or local alignment, or to tolerate different read lengths or mismatches. Often alignment sensitivity is a trade-off with the run time (Otto et al. 2014).

As in base-calling, aligners also report a quality score (mapping quality), which is for example utilized to increase the accuracy of variant calling (Nielsen et al. 2011). Each mismatch or gap in the read decrease the mapping quality, as well as failure to find a ”unique” match, which is especially difficult in genomes containing repeats or low-complexity regions (Reinert et al. 2015). One key issue in utilizing the mapping qualities is the lack of standardization and documentation of the tools: different aligners have variable maximum mapping quality values and variable thresholds for describing ”uniquely mapped” reads (based on own experiences as well as Urban 2014, accessed 08/2017; Bradnam 2015, accessed 08/2017). Thus, an important step in sequencing data analysis is to find the most suitable aligner, define the alignment strategy (global/local, mismatch or gap penalties etc.) and to consider only highly accurately (i.e. ”uniquely”) aligned reads for the downstream processing.

Variant calling is the final step and, again, a plethora of algorithms and reviews on their performances exists (e.g. Altmann et al. 2012; Pabinger et al. 2013 and Sandmann et al. 2017 just to mention a few). In human genetics, the most widely used variant callers seem to be GATK (DePristo et al. 2011), SAMtools (Li et al. 2009) and VarScan (Koboldt et al. 2009). Recently VarDict has been suggested to be a good alternative (Lai et al. 2016; Sandmann et al. 2017). Moreover, GATK Best Practices are widely accepted for variant detection (van der Auwera et al. 2013). These tools and practices, however, are mostly developed for genotyping a diploid human genome, thus, their suitability for other type of applications may be limited. Nonetheless, many steps of the Best Practices (van der Auwera et al. 2013) generally apply in accurate variant detection despite the exact experiment in question: only high quality bases should be considered when calling a variant, the variant should be supported by many reads (total coverage), the read distribution over forward and reverse strands should be balanced and also the variant supporting reads should follow the distribution of the reference reads (Fisher's exact test of strand bias, reported as Phred score). Additionally, variants are often filtered with hard-coded thresholds for minimum allele frequency (AF) or minimum quality of the variant call, which can reduce the number of false positive variants but also lead to false negative results.

Many tools are based on Bayesian (e.g. GATK, SAMtools) or heuristic (e.g. Varscan) approaches, whereas LoFreq* (Wilm et al. 2012) is based on Poisson-binomial distribution (Sandmann et al. 2017). Some tools specifically model sequencing (LoFreq*) or PCR errors (VarDict) to increase sensitivity. And indeed, one key difference between variant callers that is relevant for the work presented in this thesis, is their capability to detect low-frequency variants. For example, GATK is recommended for detection of AF >20 % variants (DePristo et al. 2011), which is understandable given that the tool is originally aimed for diploid genome variant analysis. Moreover, not all tools are capable of handling high-coverage data like VarDict is (Lai et al. 2016). And, for example LoFreq* is especially designed to detect low-frequency

variants from high-coverage data without assumptions on ploidy as it has been developed for viral data (Wilm et al. 2012), which would more resemble the nature of mtDNA.

Thus, sensitivity and accuracy of the variant detection is dependent on all of the key data analysis steps presented above. Additional fine tuning (yet sometimes controversial) analysis steps include for example de-duplication, local re-alignment and quality score re-calibration (van der Auwera et al. 2013). In summary, the data analysis should be carefully designed keeping in mind the specific application and suitability of each tool for the data set and the research question.

1.2.5 Other artefacts

Also other than the above-mentioned data analysis artefacts might exist in the variant results. These originate from the sample processing and since those are of biological or chemical origin, they are very difficult to distinguish from true variants. For example, DNA damage may lead to misincorporation of a nucleotide or during PCR enrichment early polymerase errors may expand exponentially, or the enrichment itself might be biased and not all alleles are amplified equally.

All sample preparation steps should be conducted without unnecessary damage to the DNA. It is well-known that heat and acid exposure cause apurinic/apyrimidinic sites (Cabral Neto et al. 1992), which may either block the DNA polymerase or lead to substitutions (as reviewed by Eckert & Kunkel 1991). For example, cytosine deamination leading to uracil, and thus CG>TA variants, commonly occurs in formalin fixed, paraffin embedded samples, as well as during PCR thermocycling (Chen et al. 2014). Another common error potentially introduced by DNA polymerases is mispairing T and G, thus leading to artefactual AT>GC variants. Moreover, polymerase jumping (i.e. strand-switching) is also said to increase when the DNA is damaged (Eckert & Kunkel 1991). This could, for example, lead to increased index hopping during library preparation PCR. Li & Stoneking (2012) observed 15 % chimeric reads when 40–90 samples were multiplexed together (Li & Stoneking 2012).

One commonly observed sequencing artefact are GC>TA variants (Chen et al. 2017). This variant is commonly used as a signature of oxidative damage to DNA as an oxidative lesion, 8-oxo-dG, is sometimes paired with A by DNA polymerase (as reviewed by Kauppila & Stewart 2015). Oxidative damage may be introduced during DNA extraction if oxidized phenol is used (Claycamp 1992), thus it is advisable to use other means to extract the DNA. Furthermore, acoustic shearing of DNA during sequencing library preparation has been shown to induce oxidative damage, especially at certain GC-rich motifs (Costello et al. 2013). Costello et al. (2013) suggested that a contaminant present in the sample could induce the oxidative damage. Also, the temperature easily rises during sonication if too harsh shearing conditions are used for low-input DNA samples. It was reported that the addition of a chelator into the sample might reduce the damage (Costello et al. 2013). Similar reports have been published and the use of repair enzymes in the library preparation is suggested as a solution since the exact cause of the damage has not been identified and thus cannot be excluded (Arbeithuber et al. 2016; Chen et al. 2017).

1.3 Recent high-sensitivity variant detection methods

As discussed by Fox et al. (2014), the development of sequencing platforms continues, yet reliable variant detection <1 % still seem to be problematic due to various biases introduced during sample processing or sequencing signal detection. During past years, an increasing number of more complex sequencing approaches have been proposed in order to control for various biases and to increase the variant detection accuracy.

1.3.1 PELE-Seq

Paired-End Low-Error Sequencing (PELE-Seq) aims to remove PCR- and sequencing errors by sequencing overlapping reads which are tagged by two indices (Preston et al. 2016). Paired-end sequencing is conducted with short insert size (100 bp), thus the insert is sequenced twice and a formation of a consensus sequence is possible. The idea is similar to long-read sequencing approaches, and can be applied to

remove sequencing errors from the reads as a sequencing error is likely present only in one of the reads. Incorporation of dual-indexing to the DNA fragments increases the variant calling accuracy as the variant is required to be present in the read with both indices (Preston et al. 2016). PELE-Seq analysis utilizes LoFreq* variant caller (Wilm et al. 2012) and reliable variant detection threshold is set to AF of 0.2 % even with extremely high coverage (~60000x). They suggest inclusion of control samples with each sequencing run in order to empirically determine the suitable variant calling thresholds for each experiment. In comparison to standard sequencing library, PELE-seq allowed equal sensitivity with highly improved precision, and Preston et al. (2016) especially recommended the method for accurate and cost-efficient amplicon or small genome sequencing (Preston et al. 2016).

1.3.2 Circle sequencing

To improve the sequencing error rates, Lou et al. (2013) developed a library preparation method that utilizes circular DNA templates and rolling circle amplification (Lou et al. 2013). In this circle sequencing, the short DNA template (amplicon, cDNA or DNA fragments) is denatured and ssDNA is circularized. The circularized template DNA is amplified with random primers by φ29, which possesses strand-displacement activity and can replicate continuously around the circular template, in a process referred to as rolling circle amplification (RCA). The priming takes place also on newly synthesized DNA strands and the final DNA product is branched, tree-like structure containing physically linked, multiple copies of the single template DNA. The original fragment length should be approximately one third of the desired sequencing read length, thus the original DNA fragment is present multiple times within a single read. Such linkage allows efficient error removal when a consensus sequence is formed. The error rate reported for the circle sequencing was 2.8×10^{-4}. The main artefactual variants were CG>TA, likely arising from spontaneous deamination of cytosine to uracil. By addition of uracil-DNA glycosylase (UDG) and formamidopyrimidine-DNA glycosylase (Fpg) to excise deaminated

cytosine and 8-oxo-dG, respectively, the error rate was improved to $7.6x10^{-6}$ (Lou et al. 2013). This approach could be beneficial to any amplicon sequencing experiment or even included to the standard sequencing library preparation PCR step. Due to consensus sequence formation, the maximum theoretical efficiency of the method is 33 %, however, Lou et al. (2014) observed only ~20 % efficiency (Lou et al. 2013). Thus, the increase in accuracy also increases the sequencing costs significantly.

1.3.3 Unique molecular identifiers

Safe-SeqS

Safe-Sequencing System, Safe-SeqS, introduced by Kinde et al. (2011), utilizes unique molecular identifiers (UMI) to increase the variant detection accuracy from captured genes (Kinde et al. 2011). They simply ligated standard Illumina adapters to the DNA fragments and sequenced the formed libraries observing an error rate of $2.4x10^{-4}$ mut/bp with a stringent variant calling. However, Safe-SeqS analysis is based on the clever idea of using randomly sheared DNA fragment ends as endogenous unique molecular identifiers to form consensus sequences of the original DNA templates. This way the error rate was decreased to $3.5x10^{-6}$ mut/bp. The approach was improved by also adding exogenous 12–14-nt single-stranded identifiers, which significantly increased the number of UMIs and thus the number of targets that could be analyzed. Safe-SeqS claimed realible detection of AF 0.001 % ($1x10^{-5}$) (Kinde et al. 2011).

Duplex Sequencing

Schmitt et al. (2012), and later updated by Kennedy et al. (2014), extended on the single-stranded UMI applied in Safe-SeqS, to the use of double-stranded, 12-nt random UMIs at both ends of dsDNA fragment – a method called Duplex Sequencing (Schmitt et al. 2012; Kennedy et al. 2014). This way, both strands of the DNA template become uniquely labelled. After PCR-amplified library preparation and high-depth sequencing, single-strand consensus sequences as well as duplex

consensus sequences can be formed. Single-strand consensus sequence is effective in removing sequencing errors, whereas duplex consensus sequence is used to eliminate even first-cycle PCR-errors which are indistinguishable from real variants by all the other sequencing methods relying on single-strand sequencing. The key in Duplex Sequencing library preparation is that the strands of dsDNA are not separated from each other before UMI ligation – limiting the usage of certain DNA extraction or amplification methods, e.g. plasmid prep which is based on physical separation of the DNA strands from each other (Kennedy et al. 2014).

Moreover, the input DNA and adapter amounts, PCR amplification and sequencing depth have to be in a delicate balance to obtain optimal distribution of copies of each UMI pairs (Kennedy et al. 2014). With single-strand consensus sequence, Schmitt et al. (2012) observed error rate of $3.4x10^{-5}$ mut/bp (Schmitt et al. 2012), whereas application of duplex consensus sequence analysis decreased it to an estimation of $3.8x10^{-10}$ mut/bp. Duplex Sequencing was capable to detect AF 0.00001 % ($1x10^{-7}$), and they claim to reach even $5x10^{-8}$ AF detection threshold (Kennedy et al. 2014). However, as with circle sequencing, a significant amount of data is wasted – the efficiency of Duplex Sequencing was estimated to be only 0.8 % (Lou et al. 2013).

CypherSeq

The latest improvement to UMI or circle sequencing is represented by CypherSeq (Gregory et al. 2015). In CypherSeq, the pUC19 plasmid backbone with sequencing adapters and 7-nt double-stranded UMIs surrounding the cloning site is utilized. Template DNA is ligated into the vector and amplified either by PCR, RCA, or by cloning before sequencing. As in the other methods, the barcodes allow formation of consensus sequences and thus efficient error removal. The clear advantage of CypherSeq over earlier methods is the simple amplification of the sample and compatibility with all sequencing platforms. CypherSeq was shown to detect AF $2.4x10^{-7}$ – without detecting any erroneous variants (Gregory et al. 2015).

1.3.4 Summary

Although all of the presented methods are efficiently decreasing the error rate and enhancing the detection precision (as summarized in **Table 1.2**), this comes with a cost. Mostly, the hands-on time is increased in comparison to standard sequencing, and more importantly, the required sequencing depth might raise the costs to prohibitive levels. Indeed, as mentioned by Lou et al. (2013), one should consider whether the project truly requires detection of ”the rarest of rare variants” (Lou et al. 2013). Furthermore, the mentioned studies did not discuss the method sensitivity comprehensively but only reported the lower error-rates and allele frequency detection thresholds.

Table 1.2. Comparison of detection thresholds and error rates of high-sensitivity sequencing methods.

Method	Sample preparation	Detection threshold	Error rate	Reference
PCS	~1 kbp PCR	n.a.	$3.5x10^{-6}$	Ross et al. (2013)
PELE-seq	Short fragments/ amplicons	$2x10^{-3}$	n.a.	Preston et al. (2016)
Circle sequencing	Short fragments/ amplicons	n.a.	$7.6x10^{-6}$	Lou et al. (2013)
Safe-SeqS	No additional sample preparation steps	$1x10^{-5}$	$3.5x10^{-6}$	Kinde et al. (2011)
Duplex Sequencing	Input optimization	$1x10^{-7}$	$3.8x10^{-10}$	Schmitt et al. (2012), Kennedy et al. (2014)
CypherSeq	PCR/RCA/cloning	$2.4x10^{-7}$	n.a.	Gregory et al. (2016)

n.a. = not annotated

1.4 Mitochondrial DNA variant detection by deep sequencing

1.4.1 Nuclear sequences of mitochondrial origin – NuMTs

A cell harbors several to thousands of mitochondria which can harbor altogether thousands of mtDNA molecules, yet from a total genomic DNA (gDNA) extraction <1 % is mtDNA. Although this sounds like a very small amount, due to the small size of the genome, low-depth sequencing of gDNA (one Gbases) can yield >600x coverage of mtDNA. Therefore, mtDNA reads can be obtained as a by-product of whole-genome sequencing, often with >1000x coverage (Li et al. 2012). In theory, such coverage would allow reliable detection of mtDNA variants present at as low as ~1 % allele frequency (AF). However, a major challenge is caused by nuclear sequences of mitochondrial origin (NuMTs), as the nuclear genome harbors not only genes required for mitochondrial function but also mitochondrial pseudogenes. These NuMTs are chunks of mtDNA which are naturally transferred to the nucleus and incorporated to the nDNA via non-homologous end-joining at double-strand breaks (Hazkani-Covo et al. 2010). NuMTs may be polymorphic and present or absent in different individuals, and moreover, the mtDNA is constantly transferred to the nucleus creating new 100 % homologous NuMTs and between-individual variation in their NuMT content (Hazkani-Covo et al. 2010, Calabrese et al. 2012).

NuMTs have been identified in most species, including *Mus musculus* genome. The mouse genome sequence was estimated to harbor ~37 kbp of NuMTs as 137 BLAST hits (Hazkani-Covo et al. 2010), or 172 chunks ranging from 33 bp to 4.7 kbp in length with 66–100 % identity (Calabrese et al. 2012). Malik et al. (2016) estimated that >95 % of the mouse mtDNA genome is present in nDNA (Malik et al. 2016). The number of detected NuMTs depends on search strategy as well as nuclear genome version and completion level (Hazkani-Covo et al. 2010).

The presence of NuMTs and their above-mentioned characteristics complicate the design of mtDNA primers and, indeed, some reported pathogenic mtDNA mutations have been actually NuMT polymorphisms (Yao et al. 2008). Furthermore, it has been acknowledged that NuMTs may affect amplification- or blotting-based mtDNA content measurements (Malik et al. 2016). Similarly it may hamper accurate mtDNA variant detection by high-throughput sequencing, as NuMTs may be 100 % identical to mtDNA and increase the wild-type reads and might cause even false negative variant detection results. Or, NuMTs reads harbor a variant, which will be impossible to distinguish from a true positive variant result. Various sequencing approaches have been applied to reliably detect mtDNA variants and are discussed below in separate sections.

1.4.2 Indirect and capture-enriched mitochondrial DNA sequencing methods

Whole-genome sequencing or, even more, exome sequencing is commonly used for nDNA-focused studies. In exome sequencing, the coding regions of nuclear genome are capture-enriched before sequencing. However, even half up to of the sequencing reads originate from non-target sources, and for example, the mtDNA genome typically reaches even 100x coverage (reviewed by Samuels et al. 2013). Thus, mtDNA can be sequenced indirectly by utilizing these by-products of these data sets (Picardi & Pesole 2012). A tool, MitoSeek, exists for extracting mtDNA reads from exome or whole-genome sequencing data and to perform variant detection as well as copy number determination (Guo et al. 2013). Samuels et al. (2013) stressed that NuMTs are present in these by-product data sets, and MitoSeek takes the conservative approach to avoid NuMTs in the data set; reads are first aligned to the nuclear reference genome and only unmapped reads are used for mtDNA analysis (Guo et al. 2013; Samuels et al. 2013). Yet, they have later determined that such alignment strategy leads to poorer results in variant detection than less conservative alignment, in which either only the mtDNA reference genome or the full reference genome is used

(Zhang et al. 2016).

Another method, sequencing of capture-enriched mtDNA (Maricic et al. 2010), is also prone to artefactual variants caused by NuMTs. Li et al. (2012) showed that NuMTs are present in capture-enriched mtDNA sequencing data and they interfere with accurate low-frequency variant detection (Li et al. 2012). With a statistical method, they were able to reliably detect variants at a level of AF 5 % without any false positive results. The sensitivity could have been improved by higher sequencing depth. Furthermore, they compared the results to whole-genome sequencing, which showed significant proportion of variants originating from NuMTs (Li et al. 2012). Thus, indirect or capture-enriched mtDNA sequencing seems to be suitable only for high-frequency mtDNA variant detection (Griffin et al. 2014) or, for example, for *de novo* assembly of an unknown mitochondrial genome.

1.4.3 Amplification-based mitochondrial DNA enrichment and sequencing

Generally, long-range PCR amplification of mtDNA is considered as a method of choice to cost-efficiently avoid the presence of NuMTs yet to obtain high-coverage over the mtDNA genome (Payne et al. 2015). Different approaches have been suggested to amplify the entire mtDNA: in multiple short or two or more longer amplicons (Dames et al. 2013; Payne et al. 2015), or an even better way to avoid NuMTs amplification would be to amplify the mtDNA in a single long amplicon (Cui et al. 2013). Li et al (2012) utilized long-range PCR of mtDNA sample as a reference when comparing the variant results from whole-genome or capture-enriched mtDNA sequencing. However, even with long-range PCR, 23 % of the detected variants were likely NuMTs (Li et al. 2012). These results highlight the need of careful primer design such that NuMTs amplification is minimized, and the primer design should be tested by BLAST, and failure of NuMTs amplification could be empirically verified by utilizing ρ_0 DNA, which has been depleted from mtDNA (Payne et al. 2015).

Long-range PCR amplification and sequencing is utilized in human studies to detect mtDNA variants or deletions (e.g. (He et al. 2010; Li et al. 2010; Zaragoza et al. 2010; Payne et al. 2011; Sosa et al. 2012; Payne et al. 2013; McElhoe et al. 2014; Gardner et al. 2015; Pyle et al. 2015). Often the variant detection threshold has been high (AF >10 %) mainly due to limited coverage and higher error-rates of the earlier technologies, whereas the later publications used lower thresholds (AF ~0.2–1 %). Payne et al. (2015) suggested that sequencing platform error-rate will be the main limiting factor in variant detection. With a high-accuracy platform, as low as AF 0.1 % can be reached at 8000x coverage, and higher coverage would not further improve the sensitivity. Furthermore, Payne et al. (2015) remind us of the importance of using high-fidelity DNA polymerase (**Table 1.3**) for the amplification in order to avoid unnecessary artefacts (Payne et al. 2015).

Table 1.3. Fidelities of different DNA polymerases.

Polymerase	Error rate	Fidelity relative to *Taq*	Longest product length	Provider(s)	Reference(s)
Phusion®	$4.4x10^{-7}$	>50x	<20 kbp[a]	a, b	a, b
KOD	$5.7x10^{-6*}$	~4x	~15 kbp	c	(Takagi et al. 1997)
PrimeSTAR GXL	$6.2x10^{-5}$	60x[d]	>30 kbp	d	d
φ29	$<9.5x10^{-6}$	~2x[*]	>70 kbp[a]	a, b	(Esteban et al. 1993; Paez et al. 2004)

*a = ThermoFisher Scientific, b = New England Biolabs, Inc., c = Merck Chemicals GmbH, d = TaKaRa Bio Inc., * estimated from the reported values and Taq fidelity of $2.28x10^{-5}$ (reported by ThermoFisher Scientific)*

In another similar long-range PCR approach the mtDNA is first enriched by a plasmid preparation kit – based on the fact that mtDNA genome is circular (Quispe-tintaya et al. 2015). This mtDNA enrichment strategy

was recently utilized to study mtDNA heteroplasmy in Chinese hamster ovary cells. They were able to detect variants at AF >=1 %, and did not detect presence of NuMTs (Kelly et al. 2017). As plasmid prep is already enriching the mtDNA in relation to nDNA, it is logical to think that this approach could be less susceptible for the presence of NuMTs, however, an additional processing step could cause other artefacts.

Instead of long-range PCR amplification, recent mtDNA variant detection approaches have focused on RCA (MitoRCA-seq by Ni et al. 2015, MitoRS by Marquis et al. 2017), which has been earlier used for viral studies (Johne et al. 2009). In addition to NuMTs amplification, biased amplification or the presence of a variant at the primer site, or sensitive reagent setup is of concern in long-range PCR amplification – issues which can be overcome by RCA (Marquis et al. 2017). The method is also called multiple displacement amplification (MDA) as it is based on the ability of the DNA polymerase φ29 to displace the non-template strand and generate multiple copies of a circular template in hours at 30 °C. Furthermore, amplification can be primed simultaneously at multiple sites with exonuclease-resistant primers to increase the reaction efficiency (Dean et al. 2001). Small circular mtDNA is preferably enriched from linear nDNA with mtDNA-specific primers even from picograms of DNA (Marquis et al. 2017). The mtDNA RCA-enrichment has been also commercialized by QIAGEN, which provides REPLI-g Mitochondrial DNA kit (QIAGEN GmbH). Marquis et al. (2017) combined the amplification to tagmentation-based library preparation, whereas Ni et al. (2015) utilized restriction digestion enzymes and size selection on agarose gel before further fragmentation of the DNA by Covaris.

With ρ_0 DNA, NuMTs contamination in RCA was shown to be <0.06 %, and even that amount of amplification was suggested to originate from the incomplete depletion of mtDNA (Marquis et al. 2017). However, Ni et al. (2015) determined the NuMTs contamination with an alignment-based approach reaching ten times higher estimates for NuMTs originating reads. Yet, even this was much lower than what was

estimated earlier for long-range PCR approach. Unlike long-range PCR, RCA does not introduce coverage bias (Marquis et al. 2017). Furthermore, RCA was determined as an accurate method for variant detection: φ29 error rate is at a level of 10^{-6} (**Table 1.3**) and variant detection was reliable at AF 1 % on ~3000–30000x coverage data (Marquis et al. 2017) or 0.3 % on ~55000x coverage data (Ni et al. 2015).

Taken together, amplification-based methods are fast, inexpensive, easily controlled and scalable to high-throughput for mtDNA variant detection studies. However, the accuracy is greatly dependent on optimal PCR without significant biases and with careful primer design, yet still the risk of enriching NuMTs cannot be fully excluded. Moreover, as all DNA polymerases make errors, it is impossible to eliminate these completely from any analysis. Although high-fidelity enzymes exist, early-cycle errors will always be possible and will be indistinguishable from true variants. If the application does not require extremely sensitive variant detection and if the amount of available sample is limited, amplification-based methods, especially RCA, seem very promising approaches for mtDNA variant studies.

1.4.4 Traditional mitochondrial DNA enrichment and sequencing

Traditionally mitochondria are enriched from other cellular material by differential or density gradient centrifugation methods (Frezza et al. 2007; Wieckowski et al. 2009). Such methods enrich mitochondria several folds and mitochondria are generally used for functional assays or blotting experiments. Only a few studies have utilized gradient centrifugation to extract mtDNA for sequencing. Williams et al. (2010) presented a Mito-seq approach utilizing homozygote mtDNA mutator mouse, Nycodenz gradient and paired-end sequencing. They observed <10 % and >78 % of sequencing reads aligning to mtDNA reference genome in two groups of samples, in which the latter samples were enriched with optimized protocol, thus, containing significantly less nDNA contamination. Furthermore, they concluded by ρ_0 assay that NuMTs did not confound the variant detection analysis as only 0.002 %

of the reads aligned to mtDNA. However, the variant detection threshold was set to AF 1 % and one of the main aim in their study was a breakpoint detection (Williams et al. 2010).

Another study with homozygote mtDNA mutator mouse by Ameur et al. (2011) utilized sucrose gradient for mitochondria enrichment and SOLiD sequencing (Ameur et al. 2011). They reported 35–69 % of the reads to be aligned to mtDNA reference genome and as well concluded that nDNA contamination varies sample-by-sample and cannot be fully excluded with gradient enrichment (Ameur et al. 2011). Similar to Williams et al. (2010), Ameur et al. (2011) estimated by numerical calculations that the effect of NuMTs should still be negligible – approximately less than one NuMT variant containing read out of $5x10^4$ reads (Ameur et al. 2011). Since SOLiD is very different technology from Illumina sequencing, the variant frequency per position was also determined directly from the reads and no minimum detection threshold was reported, however, as they describe AF >0.5 % as high-frequency variants (Ameur et al. 2011), the detection threshold was likely much below that.

1.4.5 Other mitochondrial DNA enrichment strategies for sequencing

Other mtDNA enrichment strategies recently published are based on enzymatic degradation of nDNA. Jayaprakash et al. (2015) developed an approach called Mseek (Jayaprakash et al. 2015), in which total DNA (gDNA) is extracted and treated with exonuclease V (ExoV). ExoV specifically digests linear ssDNA or dsDNA, thus, circular mtDNA stays intact and is enriched. The digestion is, however, very long, 48 hours at 37 °C, after which the nDNA contamination level is controlled by PCR, and in case of successful PCR, the digestion is continued for additional 16 hours (Jayaprakash et al. 2015). While being a very simple and inexpensive approach with minimal hands-on time, the approach is very slow. A major drawback of their method was very inefficient sequencing library preparation as only 40 % of the reads were mtDNA origin and the rest were mostly adapter dimers. In the end, they also recommend to

combine the sample preparation with long-range PCR (Jayaprakash et al. 2015).

Another similar approach, published by Gould et al. (2015), utilizes Plasmid Safe ATP-dependent DNase to digest nDNA (Gould et al. 2015). They utilized the enzymatic treatment after extraction of DNA from an enriched mtDNA preparation (kit-based mtDNA enrichments). Yet, they were able to obtain only ~36–62 % mtDNA sequencing reads. This was an improvement in comparison to only enriched mtDNA prepration, however, the samples were still highly nDNA contaminated. An intriguing suggestion from Gould et al. (2015) was to create mitoplasts before mtDNA extraction. The idea behind it is that nDNA attached to the mitochondria would be efficiently removed while removing the outer membrane, and thus, such approach could yield in highly pure mtDNA preparation – ideally close to 100 % (Gould et al. 2015).

1.5 Data analysis approaches for mitochondrial DNA variant detection

Early on Li et al. (2010) utilized long-range PCR and low-coverage data (~70x) as well as simulated data sets to suggest an analysis workflow for mtDNA variant detection. They concluded that a highly important step in reliable variant detection is double-strand validation i.e. the variant has to be present on both strands at least in two independent reads. Further requirements were that duplicate reads should be excluded as they may have huge impact on detected AFs in low-coverage data and minimum base-call quality should be at least 20. However, due to low-coverage, their variant detection threshold remained very high, AF >=10 % (Li et al. 2010).

Later they developed an approach to detect low-level mutations. Their approach was based on the fact that errors occur strand- and position-dependently and that there are error hotspots and these were included into statistical models to separate them from true mutations. With their new approach, variant detection was possible down to AF >2 % from

500x data. They were able to efficiently exclude sequencing errors, however, they noted that contamination or chimeric reads cannot be distinguished and experimental protocols should be refined in order to reduce such artefacts (Li & Stoneking 2012).

Guo et al. (2012) utilized GATK variant caller to detect AF 1 % mtDNA variants from long-range PCR amplified data at ~4000x coverage (Guo et al. 2012). They applied relatively stringent criteria that each strand had to be covered >200x and strand-bias Phred score should not deviate from zero. Later on they developed their approach as the MitoSeek pipeline to analyze by-product reads from exome or whole-genome sequencing data sets. They introduced their own variant calling approach, as GATK is developed for a diploid genome and is inaccurate for mtDNA variant calling. In their approach, they compare empirical allele counts from tumor and normal samples and the algorithm automatically adjusts the detection threshold suitable to the depth of the data set (Guo et al. 2013). Similarly, Calabrese et al. (2014) have developed MToolBox. They utilize SAMtools for variant calling and implement variant calling quality score as well as minimum number of supporting reads. Different from other tools, MToolBox also assign a haplogroup and annotates the variants (Calabrese et al. 2014). Further advancement to analysis pipelines was mit-o-matic, which in contrast to earlier command-line tools, is cloud-based (Vellarikkal et al. 2015). The most sensitive analysis pipeline is mtDNA-server, which models several quality aspects in their variant calling approach (Weissensteiner et al. 2016).

A major drawback with the pipelines is their relatively poor flexibility as certain minimum thresholds may be hard-coded, or the user has poor control over many parameters, and some analysis options are not even included or cannot be excluded. mit-o-matic provides the possibility to choose between three different aligners and accounts for circularity of the mtDNA genome by adding some bases from the beginning to the end of the reference genome (Vellarikkal et al. 2015). Similarly, Ding et al. (2015) have developed their own likelihood-method to detect

mtDNA variants and they consider the mtDNA genome circularity by dual approach: First aligning the reads and calling variants on normal reference genome and then to a reference genome in which the genome junction is shifted to the middle. The resulting variants are combined such that only the middle part of the genome is considered from the normal reference genome analysis, and only the junction region from the shifted reference genome analysis (Ding et al. 2015). Although mtDNA is different from nDNA and requires specific considerations in analysis steps, one additional drawback of these various pipelines is their poor comparability to other methods.

The above-mentioned pipelines are available for human mtDNA only and no pipeline for mouse mtDNA exists. Thus, there is a need to establish an accurate mouse mtDNA variant detection analysis. This could be achievable by following the above discussed guidelines and strategies for quality measures in sample preparation and sequence analysis and by utilizing the already existing tools.

2 PROJECT AIMS

This PhD thesis focuses on deep sequencing of the whole mitochondrial genome in order to reliably detect extremely rare mtDNA variants. The mitochondrial DNA mutator mouse is used as a model for mtDNA saturation mutagenesis throughout the thesis projects.

The aims of the thesis are divided into three subprojects as follows:

1. Optimization of mitochondrial DNA extraction method.
2. Selection of the sequencing method for low-frequency mitochondrial DNA variant detection.
3. Application of the optimized approaches to address mitochondrial biology research questions.

First, the projects aim to develop an improved mtDNA sequencing approach for reliable detection of extremely rare mtDNA variants. With the improved method it is possible to build a detailed picture and expand earlier studies on how mtDNA may be mutated and transmitted. Such information is of key importance in understanding mitochondrial biology, especially mtDNA maintenance. Moreover, it may help to develop preventive measures for transmission of mitochondrial disorders.

3 Materials and methods

3.1 Experimental animals

All animal experiments were performed in strict accordance with guidelines of the Federation of the European Laboratory Animal Science Association (FELASA). Protocols were approved by the Landesamt für Natur, Umwelt und Verbraucherschutz, Nordrhein-Westfalen, Germany.

All mouse lines were originally generated by using inbred C57BL/6NCrl (Charles River Laboratories, Germany, strain code 027) background. Mice were fed *ad libitum* on a standard mouse food (ssniff M-H Low Phytoestrogen, Ssniff Spezialdiaeten GmbH) or enhanced diet (ssniff M-Z Low-Phytoestrogen) during breeding or with newly weaned mice and maintained at 21 °C in a 12-hour light/dark cycle by Comparative Biology of Max Planck Institute for Biology of Ageing.

3.1.1 Animals for optimization of the methods

To optimize the mtDNA extraction, multiple tissues (mostly liver and brain, but also heart and kidneys) from fifty mice were utilized: half of the mice were from the mtDNA mutator mouse lineage (as described below in **Chapter 3.1.2**), but significant amount of the tissues were also obtained as a surplus from experiments conducted by others (e.g. liver, brain or kidneys from a heart-specific knock-out mouse lineage). Use of surplus tissues reduced the number of mice required for the project, and additionally diminished the waste produced in other experiments.

For spike-in control samples, a single wild-type mouse carrying NZB mtDNA (NZB) was used. The mtDNA of these mice deviates from the reference strain at 89 positions as listed in **Appendix 1**.

3.1.2 Animals for creating the variant profile of the entire mitochondrial genome

In total, six mice carrying mutated mtDNA were generated by crossing males heterozygous for the exonuclease-deficient mtDNA polymerase gamma (*PolgA*$^{WT/D257A}$) to females heterozygous for knock-out alleles of PolgA (*PolgA*$^{WT/KO}$) and selecting genotypes *PolgA*$^{D257A/KO}$ (MKO) as experimental animals lacking maternally transmitted mtDNA mutations. Three true wild-type (*PolgA*$^{WT/WT}$, WT) mice were bred as separate lines, and one WT mouse was a littermate from an aforementioned cross (genotype *PolgA*$^{WT/WT}$, not carrying maternally inherited mtDNA mutations). Immediately after dissection of target tissues, mitochondria were isolated according to the protocol below.

3.1.3 Animals for studying purifying selection and mitochondrial RNA processing

The same mouse breeding scheme was used for purifying selection and mitochondrial RNA processing projects. Both projects utilized the MKO mices as founders (F1) to generate mtDNA mutations. These founders were mated with *PolgA*$^{WT/WT}$ males in order to create wild-type female lineages carrying maternally inherited mtDNA mutations. First, N1 generation females were selected for *PolgA*$^{WT/KO}$ genotype and were mated to *PolgA*$^{WT/WT}$ males. The following generations (N2 and N3) were either of *PolgA*$^{WT/KO}$ or *PolgA*$^{WT/WT}$genotype. Heterozygote knock-out mice (*PolgA*$^{WT/KO}$) have half of the PolgA transcript present and up to 15 % less mtDNA copies in comparison to wild-type mice, but they have no difference in phenotype (Hance et al. 2005).

For the purifying selection project, four F1 mice were used as founders. Two mice per generation, per lineage were dissected. For mitochondrial RNA processing project, a single F1 founder was used to produce N1 generation mice. Of these, three littermates were used. One of them was the mother of three N2 generation littermate mice. For RNA extraction, ~50 mg piece of tissue was snap-frozen in liquid N_2 and stored at -80 °C until RNA extraction.

3.2 Mitochondria isolation and DNA extraction protocols

3.2.1 Gradient centrifugation methods

Unless otherwise mentioned, in all gradient centrifugation methods, mitochondria were isolated by first homogenizing the isolated tissue in 15 ml of isolation buffer (320 mM sucrose, 20 mM Tris, 1 mM EGTA, 3 mM $CaCl_2$, pH 7.2 at room temperature, 0.2 % w/v BSA) at 1000 rpm for 12 strokes with a glass-teflon homogenizer (Potter S, Sartorius). Then, 20 ml of isolation buffer was added and the cell debris was removed by centrifugation (1200 g, 10 min, 4 °C). The supernatant was collected into a fresh tube and the pellet was resuspended into 35 ml isolation buffer by vigorous shaking and re-pelleted (800 g, 10 min, 4 °C). Mitochondria were pelleted from the supernatants by centrifugation (8500 g, 10 min, 4 °C) and the pellet was washed once by re-suspension into 35 ml buffer and re-pelleted.

Sucrose gradient

The final pellet was resuspended into 1 ml of 0.6 M sucrose. Sucrose gradient was formed by placing 5-ml layer of 1.5 M (or 1.75 M) sucrose and 5-ml layer of 1 M sucrose carefully on top of each other into an ultracentrifuge tube (Beckman Coulter, Inc.). Mitochondria suspension was placed on top of the layers and the tube was filled with 0.6 M sucrose. Carefully balanced tubes were centrifuged either with SW-28 rotor (22000 g, 30 min, 4 °C) or with SW 41 Ti rotor (15000 rpm, 24 min, 4 °C, Beckman Coulter, Inc.). Mitochondria were collected between the layers with a syringe and pelleted by centrifugation (16000 g, 10 min, 4 °C).

The pellets were subjected to DNA extraction with Gentra® Puregene® Tissue kit (QIAGEN GmbH) with an adjusted protocol. The pellets were resuspended by vortexing in 800 µl of Cell lysis solution preheated to 65 °C, and 8 µl of Proteinase K (10 mg/ml) was added. Mitochondria were lyzed at 55 °C, 600 rpm for 3 hours or until the solution was clear.

RNA was digested by adding 8 µl of RNase A Solution and incubated at 37 °C for 15–60 min. Samples were incubated on ice for 1–2 min before addition of 270 µl Protein Precipitation Solution and vigorous vortexing. The incubation on ice was continued for 5–30 min. Proteins were pelleted by centrifugation (16000 g, 3 min, room temperature) and supernatant was collected into a fresh tube. DNA was precipitated by adding 850 µl isopropanol and incubating at room temperature for 16–20 hours. DNA was collected by centrifugation (16000 g, 30 min, room temperature), washed with 1 ml of 70 % EtOH and repelleted. DNA was dissolved into nuclease-free H_2O.

CsCl-gradient

The CsCl-gradient protocol followed the sucrose gradient protocol until DNA extraction step. Instead of DNA extraction by the kit, the mitochondria pellet was resuspended into 1.6 ml TE-buffer by vortexing. Mitochondria were lyzed by addition of 0.4 ml SDS 10 % and 133 µl Proteinase K (10 mg/ml) and incubated for 10 min at room temperature. Then the sample was incubated with 330 µl of 7 M CsCl solution at 4 °C for 16–20 hours. Mitochondrial membranes were pelleted by centrifugation (17000 rpm, 10 min, 4 °C), the supernatant was collected into a fresh tube and the volume was measured. Solid CsCl was added to the solution at a concentration of 0.93 g/ml. The sample density was adjusted to 1.57 ± 0.01 g by adding either solid CsCl or TE-buffer. Then, 10 µl/ml of SYBR® Safe DNA Gel stain (ThermoFisher Scientific) was added. The sample was transferred into an ultracentrifuge tube and balanced with a balancing solution (0.57 g/ml CsCl in TE-buffer). DNA was separated by centrifugation in a MLS-50 rotor (27000 rpm, 69 hours, 20 °C, Beckman Coulter, Inc.) and afterwards the DNA band was visualized with UV-light and collected with a syringe. SYBR® Safe DNA Gel stain was removed by adding equal volume of 1:1 mix of 1-butanol and 7 M CsCl solution into the sample. The phases were separated by centrifugation (1500 rpm, 3 min, room temperature) and the aqueous phase was collected. The separation was repeated 1–4 times, until the pink color disappeared from both

phases. The collected sample was diluted with 5 ml ddH_2O and DNA was precipitated 16–20 hours at -20 °C by adding 10 ml absolute EtOH. DNA was pelleted by centrifugation (12000 g, 15 min, 4 °C), washed with 5 ml 70 % EtOH and repelleted by centrifugation (12000 g, 5 min, 4 °C). Traces of EtOH were evaporated and DNA pellet was dissolved into nuclease-free H_2O.

Percoll gradient

In Percoll gradient protocol, tissues were homogenized at 100 rpm for 10 strokes, cell debris was pelleted by centrifugation (1000 g, 10 min, 4 °C) and mitochondria collected from the supernatant by centrifugation (12000 g, 10 min, 4 °C). The mitochondria pellet was resuspended into 200 µl of isolation buffer. Percoll gradients were prepared by adding 18 w-% of Percoll (Sigma-Aldrich Co.) into 5 ml of cold isolation buffer in an ultracentrifugation tube and loading the resuspended mitochondria on top of it. The mitochondria were separated from the other cell membranes or organelles by centrifugation with MLS-50 rotor (40000g, 20 min, 4 °C). The supernatant was carefully removed and mitochondria were resuspended into the remaining Percoll solution by swirling the tube. Finally, Percoll was removed by adding 10 volumes of isolation buffer and pelleting the mitochondria by centrifugation (6300 g, 10 min, 4 °C). DNA was extracted as described in the sucrose gradient protocol.

3.2.2 Mitochondria isolation kit

Mitochondria extraction kit – Tissue and Mitochondria isolation kit mouse tissue (Miltenyi Biotec GmbH) were used together according to the manufacturer's instructions in order to isolate highly pure mitochondria utilizing the Anti-TOM22 labelled magnetic beads. However, these experiments resulted in highly nDNA contaminated samples from liver, brain and heart. Thus, the method was improved by adding DNase I treatment step (as discussed below in **Chapter 3.2.3**) after the bead purification, but the DNA was completely lost in these experiments. Finally, only Mitochondria extraction kit – Tissue was used to homogenize the tissue (N1 generation mice, livers), and the

mitochondria were collected by centrifugation (8500 g, 10 min, 4 °C). Next, the protocol was similar as described below for mtDNA-seq (**Chapter 3.2.3**) from DNase I treatment step on, except the final DNA was dissolved into nuclease-free H_2O.

3.2.3 mtDNA-seq

Mitochondria were isolated using protocol described by Kennedy et al. (2013) with minor modifications. Briefly, 500–800 mg of mouse liver tissue and a full brain (~430 mg) were collected and homogenized in mitochondria isolation buffer (MIB, 320 mM sucrose, 20 mM Tris, 1 mM EGTA, 1 % BSA, pH 7.2 at room temperature, 1 % w/v BSA) with a glass-teflon homogenizer. Cell debris was removed by centrifugation (800 g, 10 min, 4 °C), supernatant was transferred to a fresh tube and the centrifugation was repeated one more time. Finally, mitochondria were pelleted (8500 g, 10 min, 4 °C). Pellets were resuspended into Mito-DNase buffer (Kennedy et al. 2013) containing 0.03 mg/ml DNase I and 0.02 mg/ml RNase A. The homogeneous solution was divided into subfractions representing 100–150 mg of the original tissue sample, and incubated at 37 °C for 1–1.5 hours. Mito-DNase buffer was removed by pelleting the mitochondria (13000 g, 30 min, 4 °C). Pellets were washed twice by resuspending into MIB (containing 0.2 % w/v BSA) and centrifugation (13000 g, 15 min, 4 °C). Clean pellets were snap-frozen in liquid N_2 and stored at -80 °C until the mtDNA extraction on the same day.

For mtDNA extraction, mitochondria pellets were lyzed overnight at 56 °C in lysis buffer (20 mM Tris, 150 mM NaCl, 20 mM EDTA, 1 % SDS, pH 8.75 at room temperature, 0.02 mg/ml Proteinase K, 0.02 mg/ml RNase A), and DNA was purified by chloroform: isoamylalcohol extraction in presence of 1.2 M potassium acetate. Before ethanol precipitation for 3 hours at -80 °C, the DNA preparations were treated with 100–200 µg RNase A at 37 °C for 45–60 min. DNA was pelleted at 16000 g for 15 min, the precipitate was washed twice with 500 µl of 70 % ethanol and pelleted at 16000 g for 15 min at room temperature. The DNA pellet was dissolved into 18–35 µl of 5 mM Tris,

pH 8.5 (Macherey-Nagel). Level of nDNA contamination was confirmed by PCR with GoTaq® DNA Polymerase (Promega GmbH) and *PolgA*-specific primers (5'-CTTCGGAAGAGCAGTCGGGTG-3' and 5'-GGGCTGCAAAGACTCCGAAGG-3'), at conditions:

$$\frac{94\,°C}{1\,min}\left(\frac{94\,°C}{30\,s}\,\frac{58\,°C}{30\,s}\,\frac{72\,°C}{45\,s}\right)^{30\,x}\,\frac{72\,°C}{3\,min}\,\frac{8\,°C}{\infty}.$$

Subfractions highly pure from nDNA (i.e. barely visible PCR product on the gel) were combined and concentration was quantified with the fluorometric method (Qubit™ dsDNA HS Assay kit, ThermoFisher Scientific). The total yield of highly pure liver or brain mtDNA was >2 µg and ~100–200 ng, respectively.

3.3 Genomic DNA extraction

Genomic DNA (gDNA) was extracted from a ~50-mg piece of liver tissue (snap-frozen in liquid N_2 immediately after dissection and stored at -80 °C) that was minced with a mortel. The DNA extraction was as described above for mtDNA-seq (**Chapter 3.2.3**), but ethanol precipitation was shortened to 1 min and the DNA was dissolved into 200 µl of 5 mM Tris, pH 8.5 (Macherey-Nagel).

3.4 Total RNA extraction

The total RNA was extracted from a 50-mg piece of a snap-frozen tissue with TRIzol™ Reagent (ThermoFisher Scientific) according to manufacturer's instructions with minor modifications. Briefly, the frozen tissue was first minced with a mortel. The homogenization was finalized by adding 1 ml TRIzol™ Reagent and using Lysing Matrix D tubes (MP Biomedicals, LLC) with settings 24x2, 6 m/s, 4x20 s. Tubes were incubated on ice for 5 min before addition of 200 µl chloroform followed by 2–3 min incubation on ice. Samples were centrifuged at 12000 g for 15 min at 4 °C. Aqueous phase was collected to a fresh tube and incubated with 500 µl isopropanol for 20 min on ice. RNA was pelleted by centrifugation at 16000 g for 30 min at 4 °C and washed

with 1 ml of 75 % EtOH. RNA was re-pelleted at 16000 g for 5 min at 4 °C, and after removal of the supernatant, the leftover EtOH was spinned down at 16000 g for 1 min at 4 °C. Traces of EtOH were evaporated by heating the sample at 55 °C for 2 min. The final RNA sample was dissolved into 50 µl nuclease-free H_2O by heating the sample at 55 °C for 2 min.

Any contaminating DNA was removed using 6 µg of the RNA for TURBO DNA-free™ treatment (ThermoFisher Scientific) according to manufacturer's instructions. The integrity and concentration of the samples were measured from 1 µl of the sample by a miniaturized gel electrophoresis system, Experion™ RNA StdSens Analysis kit (Bio-Rad Laboratories, Inc.).

3.5 Mitochondrial DNA cloned into a plasmid backbone, pAM1

A pAM1-plasmid containing a full mouse mtDNA – deviating from the mouse mtDNA reference genome (C57BL/6J, NC_005089.1) at positions 4794.C>T, 9348.G>A, 9461.T>C, 10918.A>G and 12048.T>C – restriction digested with HaeII (restriction site RGCGCY, at position 2603), cloned into a 2.5-kb pACYC177 plasmid backbone also restricted with HaeII was kindly obtained from Prof. Dr. Nils-Göran Larsson. A single clone was expanded and DNA extracted using Plasmid Midi kit (QIAGEN GmbH) according to manufacturer's instructions.

3.6 Amplicon PCR

3.6.1 Amplicon PCR without tagged primers

Amplicons were designed to represent the entire mtDNA genome in 2.8–7 kb amplicons as indicated in **Table 3.1**. The amplification PCR was done with Phusion® High-Fidelity DNA Polymerase (New England Biolabs, Inc.) according to the manufacturer's instructions, in 20–60 µl reaction volume using 10 µM final concentration of primers, 1.25 mM of dNTPs and 1 µl template DNA with following conditions:

$$\frac{98\,°C}{30\,s}\left(\frac{98\,°C}{5\,s}\ \frac{55^{2}/58^{1,4}/59^{3}\,°C}{2\,min}\ \frac{72\,°C}{2\,min}\right)^{15x}\ \frac{72\,°C}{10\,min}\ \frac{8\,°C}{\infty},$$

where the annealing temperature was adjusted for each amplicon 1–4 as indicated in the superscript.

Table 3.1. List of amplicon primers without tags.

Primer	Sequence (5' → 3')	Start position	Amplicon length (bp)
Amplicon 1 forward	TTGATGAGGATCTTACTCCC	9376	7057
Amplicon 1 reverse	TCTATGGAGGTTTGCATGTG	113	
Amplicon 2 forward	GAAACTTTATCAGACATCTGG	15773	2841
Amplicon 2 reverse	ACTTTGACTTGTAAGTCTAGG	2315	
Amplicon 3 forward	TTGACCTTTCAGTGAAGAGG	2115	4444
Amplicon 3 reverse	TTGCTCATGTGTCATCTAGG	6559	
Amplicon 4 forward	CCATTCCACTTCTGATTACC	4240	5923
Amplicon 4 reverse	GTAGGTTGAGATTTTGGACG	10163	

The DNA samples used were low-yield mtDNA-seq samples from three N1 generation brains and two N2 generation brains. Additionally, one N2 generation liver was used as a template. All DNA templates were diluted to approximately 1 ng/µl concentration. The liver sample carried heavy RNA contamination causing inefficient PCR and, different from the brain samples, the amplification was repeated for 20 cycles. In order to obtain high yield of amplicons, reactions were replicated 6–15 times.

For sequencing, each technical replicate sample of an amplicon was pooled and purified with NucleoSpin® Gel and PCR Clean-up kit (Macherey-Nagel) according to manufacturer's instructions and the final sample was concentrated to ~10 µl volume with Eppendorf™ Vacufuge™ Concentrator (Eppendorf Vertrieb Deutschland GmbH).

Representative mtDNA samples were formed by pooling all four amplicons in equimolar ratios into minimum of 20 ng/µl concentration in 10 µl final volume.

3.6.2 Amplicon PCR with tagged primers

The improved amplicon design included non-mtDNA 18-bp tags to the 5' end of the primers. Moreover, the overlaps between amplicons were harmonized and number of amplicons was reduced to three (**Table 3.2**). PCR reactions were as with non-tagged primers, except 25 ng/µl input DNA was preferred instead of 1 ng/µl (this doubled or tripled the final amplicon yield). Triplicate of 20-µl reactions were used with the following conditions:

$$\frac{98\,^{\circ}C}{30\,s}\left(\frac{98\,^{\circ}C}{5\,s}\,\frac{51^{2}/57^{3}/58^{1}\,^{\circ}C}{10\,s}\,\frac{72\,^{\circ}C}{2^{2}/3^{1,3}\,min}\right)^{25x}\frac{72\,^{\circ}C}{5\,min}\,\frac{8\,^{\circ}C}{\infty},$$

where the annealing temperatures and extension times were adjusted for each amplicon 1–4 as indicated in the superscript. Also the number of cycles was increase to 25 in comparison to the PCR without tagged primers.

Table 3.2. List of amplicon primers with tags.

Primer	Sequence (5' → 3')	Start position
Amplicon 1 forward	**TGTAAAACGACGGCCAGT**TTGATGAGGATCTTACTCCC	9376
Amplicon 1 reverse	**CAGGAAACAGCTATGACC**TCTATGGAGGTTTGCATGTG	113
Amplicon 2 forward	**TGTAAAACGACGGCCAGT**GAAACTTTATCAGACATCTGG	15773
Amplicon 2 reverse	**CAGGAAACAGCTATGACC**GATAGTAGAGTTGAGTAGCG	4373
Amplicon 3 forward	**TGTAAAACGACGGCCAGT**CAAGCCCTCTTATTTCTAGG	3756
Amplicon 3 reverse	**CAGGAAACAGCTATGACC**GTAGGTTGAGATTTTGGACG	10163

Reactions were purified with NucleoSpin® Gel and PCR Clean-up kit according to manufacturer's instructions, by adjusting the sample volume to 100 µl with H_2O and including also the optional wash step. The DNA was eluted twice into 15 µl by centrifugation first 30 g for 1 min and then 11000 g for 1 min, as suggested in the kit manual for long PCR fragments.

3.7 Rolling circle amplification

Approximately 40 ng (1 µl) of mtDNA enriched from a single N1 embryo by the mtDNA-seq method (**Chapter 3.2.3**) and gDNA extracted from WT liver were used as templates for REPLI-g rolling circle amplification with 2 µl of human mt-primer mix (QIAGEN GmbH) according to manufacturer's instructions. The amplified samples were purified by ethanol precipitation as described for genomic DNA above, except that the final DNA pellet was dissolved into 100 µl.

3.8 Illumina HiSeq library preparations and sequencing

3.8.1 Illumina HiSeq DNA-seq

In total, >100 ng of each highly enriched mtDNA sample was sent for sequencing at Max-Planck Genome-center Cologne (MP-GC, Germany). The sequencing libraries were prepared at MP-GC and library preparation steps were always adjusted according to the centre's current best practices or the sample concentration. Here, for clarity, the exact details are given only for those samples presented in **Chapters 4.3.1** and **Chapter 4.3.2**. Other libraries were prepared and sequenced very similarly, with minor modifications to the protocols. This was not considered to be a factor affecting the final results and thus omitted here.

The DNA samples were sheared by Covaris Adaptive Focused Acoustics technology (COVARIS, Inc.) and library preparation was done with NEBNext Ultra DNA Library preparation kit (old) or with NEBNext Ultra II DNA Library preparation kit (new, New England Biolabs, Inc.).

Samples of the first two mouse lineages in **Chapter 4.3.2** were sheared as 70 ng DNA in 120 µl to average fragment size of 250 bp with settings: intensity 5, duty cycle 10 %, 200 cycles per burst and 180 s treatment time with old kit and eight PCR cycles. N1 and N2 generation samples were sequenced as 1x100bp, whereas N3 generation samples were 1x150bp (single-end mode). The samples from additional two mouse lineages, however, were sheared as 100 ng DNA in 120 µl containing 1 mM EDTA to average fragment size of 400 bp with settings: intensity 5, duty cycle 5 %, 200 cycles per burst and 55 s treatment time with new kit and seven PCR cycles. All of these samples were sequenced as 2x150 bp (paired-end mode).

The conditions for samples presented in **Chapter 4.3.1** were as follows: WT1–4 and MKO1–4 samples were sheared as 100 ng in 55 µl to average fragment size of 400 bp with settings: intensity 5, duty cycle 5 %, 200 cycles per burst and 55 s treatment time with the old kit and eight PCR cycles. MKO5 was prepared otherwise similarly but with the new kit and seven PCR cycles, and MKO6 was prepared as the first set of N3 generation samples mentioned above. Furthermore, replicates of WT3–4 and MKO1 and MKO5 were also sequenced with the paired-end approach, libraries prepared as mentioned above for the additional two mouse lineages.

All samples were sequenced with Illumina HiSeq3000, with HiSeq3000/4000 SR Cluster Kit and the corresponding SBS Kit (Illumina) until one Gbase (single-end) or two Gbases (paired-end) of sequences were achieved. Only amplicon samples without tagged primers were sequenced with HiSeq2500.

3.8.2 Illumina HiSeq RNA-seq

Approximately 3 µg of extracted total RNA was sent for sequencing at MP-GC. First, 1 µg of the RNA was used for rRNA depletion with RiboZero rRNA Removal Kit (Human/Mouse/Rat) (Epicentre) following manufacturer's instructions. The sequencing libraries were prepared with NEBNExt Ultra Directional RNA Kit (New England

Biolabs Inc.) following manufacturer's instructions. Libraries were sequenced with HiSeq2500 until 5 Gbases of sequence was achieved.

3.9 Sequencing data analysis

3.9.1 Analysis of mtDNA-seq data

Sequencing reads were trimmed with Flexbar version 2.5 (Dodt et al. 2012) for quality and adapter leftovers (parameters `-q 28 -m 50 -ao 10 -at 1 -ae ANY`). Standard mouse mtDNA reference genome (C57BL/6J, NC_005089.1) was used, however, it deviates from our mouse strain at positions 4891 and 9461 (T>C), and for some of our mice, also at position 9027 (G>A). Reads were aligned to the reference genome with BWA version 0.7.12-r1039 (Li & Durbin 2009), invoking mem (Li 2013) (parameters `-T 19 -B 3 -L 5,4`). To fix the problem of alignment to a circular reference genome, a dual alignment approach was used such that reads were aligned to both, the normal reference genome and to a split genome in which the first half of the normal reference genome was transferred to the end of the normal reference genome. Only uniquely aligned reads were filtered for further analysis with samtools (parameter `-q 1`). Per position coverage was calculated using bedtools version 2.22.1 (Quinlan & Hall 2010) genomecov (parameter `-d`).

Variants were called with LoFreq* version 2.1.2 (Wilm et al. 2012): first indel qualities were set with indelqual (parameter *--dindel*), variants were called with command lofreq call-parallel (parameters `--parallel-threads 20 -N -B -q 30 -Q 30 --call-indels --no-default-filter`) and filtered with lofreq filter (parameters `--snvqual-thres 70 --sb-incl-indels -B 60 --no-defaults`). Variants were further filtered and annotated using SnpEff version 4.2 (Cingolani et al. 2012a; Cingolani et al. 2012b) with mitochondrial codon usage table. Filtering conditions required minimum fifteen alternative base supporting reads (expression `DP*AF >= 15`) and minimum of three alternative base supporting reads on each strand estimated based on the DP4 values (expression

`DP4[2] >= 3 & DP4[3] >= 3`). However, in comparison to the strand bias filtering, the effect of this minimum number of read filtering step was miniscule, and mainly applied to ensure that every detected variant fulfilled at least this threshold.

Dual alignment approach results for coverage and variant calling were combined as follows: the results comprising genome positions 200–16099 were taken from the alignment to the normal reference genome, whereas the results for the genome junction region (positions corresponding to the original genome positions 1–199 and 16100–16299) were obtained from the alignment to the split genome.

3.9.2 Analysis of pAM1 data

Data analysis steps for pAM1-plasmid samples were similar to mtDNA-seq samples, except pAM1 sequence was used as a reference genome (**Appendix 2**). The analysis followed that for mtDNA-seq (described in **Chapter 3.9.1**), except without separate junction approach. Variant calling results were kept only for variants on mtDNA (excluding the restriction site positions 2306–2308) of the plasmid and positions were corrected to correspond the original mtDNA positions. Further filtering, annotation and analysis followed the analysis for mtDNA-seq, except the removal of strain variants. One variant immediately downstream of the restriction site, at position 2609, was excluded since it appeared in all of the pAM1-containing samples and most likely originated from the restriction-ligation reaction or read alignment step. Thus, it was considered as an arterfactual variant not present in mtDNA-seq samples.

3.9.3 Analysis of amplicon sequencing data

Analysis of amplicon sequencing data produced without tagged primers followed the analysis of mtDNA-seq. Analysis of amplicon sequencing data produced with tagged primers had an additional trimming step to remove the tag-containing reads with Flexbar after the above-mentioned quality and adapter trimming (**Chapter 3.9.1**). Flexbar barcode trimming (parameter `-b`) required a fasta format file indicating the primer sequences in forward and reverse order. Primers were required to

overlap at least 15 nt, with maximum two mismatches at any end of the read and output including unassigned reads i.e. other than first PCR-cycle reads were kept for the downstream analysis (parameters `–bo 15 –bt 2 –bu –be ANY`).

3.9.4 Analysis of RNA-seq data

RNA-seq reads analysis was similar as explained for mtDNA-seq (**Chapter 3.9.1**), but reads were aligned only to mtDNA reference genome with an aligner designed for RNA data, STAR version 2.4.1d (Dobin et al. 2013) (default parameters, except for genome indexing `--genomeSAindexNbases 6`). Moreover, reads were not aligned considering the junction region as this was not relevant for RNA data.

3.10 Post-processing of the variant calling results

3.10.1 Variant loads

In this thesis, the terms ”unique” and ”total” variant loads are used to separate two different variant loads from each other and to better characterize the sample differences. Although slightly misleading term, the ”unique variant load” does not represent exactly unique variants observed within a sample, but it is rather used to describe the number of different variant types observed within a sample (e.g. 6958.C>T and 9829.T>A are counted as two ”unique” variants). Due to the nature of the high-throughput technology, in order of a variant to be detected, it has to be present in multiple reads, thus, an observed variant cannot be unique (in the exact meaning of the term) but the variant is always amplified or clonally expanded. The total variant load, on the other hand, represents the number of all variant reads observed within a sample (e.g. 6958.C>T and 9829.T>A observed on 40 and 160 reads, respectively, are counted as 200 variant reads in total).

Unique variant load was calculated as:

$$unique\ variant\ load = \frac{number\ of\ observed\ variant\ types}{number\ of\ aligned\ bases}.$$

Total variant load was calculated as:

$$total\ variant\ load = \frac{number\ of\ variant\ reads}{number\ of\ aligned\ bases},$$

where the number of variant reads was obtained by multiplying depth (DP value) by allele frequency (AF value) reported by LoFreq*. Both loads were calculated either over the whole mtDNA genome or per defined regions, separately.

3.10.2 Spike-in sample comparisons

The variant results from the NZB spike-in samples were compared by several values representing accuracy of the variant detection.

Precision

Precision or positive predictive value (*PPV*) is the proportion of true positive results of all observed positive results:

$$PPV = \frac{true\ positives}{true\ positives + false\ positives}.$$

False discovery rate (*FDR*) can be simply obtained as 1 – PPV.

Sensitivity

Sensitivity, recall or true positive rate (*TPR*) is the proportion of true positive results of all expected true results:

$$TPR = \frac{true\ positives}{true\ positives + false\ negatives}.$$

F1 score

F1 score or F1 measure (*F1 score*) determines the accuracy of the test as a weighted average of precision:

$$F1\ score = \frac{2 * PPV * TPR}{PPV + TPR}.$$

False positive and negative rates

False positive rate (*FPR*) is the proportion of true positive results of all expected true negative results:

$$FPR = \frac{false\ positives}{false\ positives + true\ negatives}.$$

False negative rate (*FNR*) is the proportion of false negative results of all expected true results:

$$FNR = \frac{false\ negatives}{true\ positives + false\ negatives} = 1 - TPR.$$

3.10.3 DNA and RNA variant comparisons

Resulting variant lists obtained by amplicon sequencing and RNA-seq were filtered for minimum AF 0.5 % as this was considered as reliable limit in amplicon sequencing. The final variant lists were merged by common variants (position and the alternative base) in order to reduce the data and focus only on variants of interest. To detect variants showing highly different variant AF values observed in DNA and RNA sample for the same, shared variant, log2-fold-change was calculated to represent the allelic imbalance of each common variant:

$$allelic\ imbalance = \log_2\left(\frac{RNA\ variant\ AF}{DNA\ variant\ AF}\right).$$

3.11 Rodent sequence alignment

Sequence alignment was conducted by James B. Stewart. Briefly, in total 112 rodent strain mtDNA the control region sequences were aligned with MAFFT (Katoh & Standley 2013). A neighbor-joining phylogenetic tree was generated with MAFFT to divide the aligned sequences into three distinct phylogenetic clades at varying distances from the reference genome sequence (NC_005089). Variable sites were assembled for the ”mouse” clade only (70 sequences), ”mice/rats” clade

(106 sequences) or all 112 aligned sequences (”rodents”).

3.12 Statistics, plots and code availability

All plots were produced with R (https://www.r-project.org), version 3.1.2. Packages *ggplot2 (Wickham 2009)* and *circlize* (Gu et al. 2014) were utilized for plotting and color schemes were from package *RColorBrewer (Neuwirth 2014)*. Error bars representing 95 % confidence interval were calculated with *stat_summary*, function *mean_cl_normal*, and exact n values are reported in each figure legends. The curves were fitted by *geom_smooth* with method *loess* (local polynomial regression fitting).

The exact commands to conduct the optimized data analysis steps per sample (as described in **Chapter 3.2.3**) are given in **Appendix 3**.

4 RESULTS AND DISCUSSION

4.1 Optimization of the mitochondrial DNA extraction method

A key for a successful mtDNA variant detection study is a highly enriched mtDNA sample free from nDNA contamination. Differential centrifugation (Frezza et al. 2007) or density gradient methods (CsCl, sucrose and Percoll gradients) have been routinely used to enrich mitochondria (Boore et al. 2005, Wieckowski et al. 2009). Such methods enrich the mitochondria from cellular debris and the extracted mtDNA is pure enough for most mitochondrial studies, but the mtDNA still does contain significant amount of nDNA. Many approaches attempt to simplify the laborious or time-consuming tissue homogenization and centrifugation steps. Various companies sell mitochondria enrichment kits or even mtDNA extraction kits, and also regular plasmid preparation kits have been applied to enrich mtDNA from nDNA (Quispe-Tintaya et al. 2013).

Of the commercial kits, Abcam's Mitochondrial DNA Isolation kit (ab65321), for example, is based on successful differential centrifugation enrichment of mitochondria after enzymatic treatment of the cells. However, the enzyme mix composition is proprietary information. BioVision also sells a kit including a step of enzymatic treatment to the DNA. It can be speculated that these enzymatic steps are based on enzymes like ExoV (Jayaprakash et al. 2015) or PlasmidSafe (Gould et al. 2015), which are supposed to degrade linear nDNA but not the circular mtDNA. Bioo Scientific even provides a complete kit starting from mtDNA extraction with ExoV treatment and finishing with Illumina library preparation (Bioo Scientific Corporation). Such methods, however, are not suitable for MKO mice which also carry linear, truncated mtDNA molecules that would be degraded along with the linear nDNA. Furthermore, with ExoV protocol the DNA is treated at least for 48 hours at 37 °C (Jayaprakash et al. 2015), a long treatment protocol may expose the DNA to damage and

other unintended artefacts. Miltenyi Biotec GmbH provides another approach with a gentleMACS Dissociator device and mitochondria enrichment with anti-TOM22 coated magnetic beads, which has been shown to enrich functional mitochondria (Franko et al. 2013). Mitochondria enrichment with kits is more expensive than traditional methods, yet the kits may significantly reduce the hands-on time.

Despite the utilized enrichment method, nDNA contamination is not addressed or is a neglected issue. It can hamper accurate mtDNA variant detection due to the presence of NuMTs. Moreover, a huge proportion of sequencing capacity is wasted for non-target nDNA reads. Thus, to increase the reliability and sensitivity of mtDNA variant detection, the method to extract mtDNA was optimized. Traditional enrichment methods were tested for mitochondria isolation from several tissues (brain, heart, liver, spleen, kidney) in combination with various commercial or traditional DNA extraction methods (e.g. QIAGEN Gentra Puregene Tissue kit or phenol:chloroform extraction). However, mtDNA was poorly enriched from nDNA contamination, which was detected as clearly positive PCR amplification of nDNA-encoded *PolgA* (for example **Fig. 4.1a**). Alternatively, the total mtDNA yield of these samples was very low (below 50 ng) and did not meet the minimum requirement for standard Illumina sequencing library preparation. One reason for high nDNA contamination may be that the protocols for tissue homogenization are rigorous (e.g. in Frezza et al. [2007] and in Wieckowski et al. [2009] 4–8 strokes with 1500–1600 rpm), which easily leads to increased release of nDNA from the broken nuclei.

The problem of nDNA contamination and/or low mtDNA yield, especially in human mtDNA studies, is often overcome by long-range PCR (Payne et al. 2015), or recently by rolling circle amplification (Ni et al. 2015, Marquis et al. 2017). Amplification-based methods, however, are thought to be prone to polymerase errors leading to artefactual variants. Thus, suitability of these methods for low-frequency mtDNA variant detection was investigated using low-yield or highly nDNA-contaminated mtDNA samples as templates.

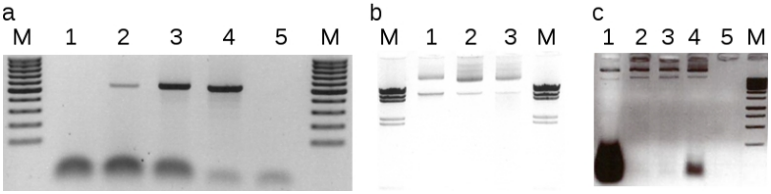


Figure 4.1. Quality control examples of liver and brain mtDNA samples. a) *Qualitative PCR to test for presence of nuclear DNA in sample. An example of 20-µl PolgA-PCR from pure liver and brain mtDNA samples (lanes 1 and 2, respectively) as well as clearly contaminated brain mtDNA sample (lane 3). Wild-type genomic DNA (lane 4) and H_2O (lane 5) were used as controls for the PCR and only 5 µl was loaded on the gel to avoid overloading. M = GeneRuler 100 bp DNA Ladder.* ***b)*** *An example showing the integrity of 100-ng liver (lane 1) and brain (lanes 2 and 3) mtDNA samples (same samples as used for the PCR in a). The lowest band is linear and the strongest is nicked mtDNA molecules. M = Lambda DNA/HindIII Marker.* ***c)*** *Examples of RNA contaminated liver mtDNA samples: mtDNA extraction without RNase A treatment and insufficient RNase A treatment of the final mtDNA preparation on lane 1 and 4, respectively. Lane 2 and 3 represent high-quality liver mtDNA samples, lane 5 is an empty well. M = GeneRuler 1 kb DNA Ladder. All gels contained 1 % agarose and 5 µg/ml ethidium bromide.*

Kennedy et al. (2013) emulated the earlier protocols in which enriched mitochondria were treated with DNase I (Kasamatsu et al. 1971) to efficiently remove nDNA contamination. Yet Kennedy et al. (2013) used simple differential centrifugation for the mitochondria enrichment instead of sucrose gradient (Kennedy et al. 2013). The idea of DNase I treatment is based on the fact that mtDNA is protected inside the isolated, intact mitochondria, whereas the contaminating nDNA remain outside of the outer membrane of mitochondria, and thus may be digested by the enzyme. A slightly modified protocol of that presented by Kennedy et al. (2013) was tested and also applied in combination with a mitochondria isolation kit alone or together with a magnetic bead isolation kit (Miltenyi Biotec GmbH). One key step in this protocol was to reduce the tissue homogenization to 3–5 strokes with only 200 rpm, potentially decreasing the total mitochondria yield, but keeping the

nuclei intact and reducing the overall nDNA contamination. According to the *PolgA*-PCR, mtDNA samples extracted by these methods were extremely pure and free from nDNA contamination (e.g. **Fig. 4.1a**). The high total mtDNA yield was suitable for sequencing. The integrity of the extracted mtDNA was always verified by agarose gel electrophoresis (**Fig. 4.1b**), and a rigorous RNase A treatment of the mtDNA was found to be necessary (**Fig. 4.1c**).

Samples obtained by different methods from various tissues were compared by the percentage of sequenced reads aligned to mtDNA reference genome (hereafter referred to as mtDNA alignment, **Fig. 4.2**, summarized in **Table 4.1**). First, to enable mtDNA variant detection from sample types showing poor mtDNA enrichment, the suitability of amplification-based methods was investigated. Similar to study by Ni et al. (2015), RCA was optimized using mouse mtDNA specific primer sets according to REPLI-g® Mitochondrial DNA kit's instructions (QIAGEN). Despite the extensive experimentation with the primer mix composition, primer concentration, amplification duration and temperature as well as the source of the template DNA, amplification efficiency was poor (data not shown). Finally, the human mt-primer mix provided within the kit surprisingly showed superior amplification efficiency for the mouse mtDNA and was used to enrich mtDNA from an isolated embryo mtDNA and liver gDNA samples (similar observation was also made by Marquis et al. [2017]). Amplification of embryo mtDNA was relatively successful (71 % mtDNA alignment). Despite an apparently equally efficient reaction, amplification of liver mtDNA from gDNA template was not specific, which was noted by positive *PolgA*-PCR (data not shown) and only <1 % mtDNA alignment (**Fig. 4.2**). The difference in amplification efficiencies could originate from the DNA extraction methods. The embryo sample was poorly enriched for mtDNA by mtDNA-seq protocol, during which the contaminating nDNA was likely highly fragmented by the DNase I. This could enable efficient mtDNA amplification by REPLI-g relative to the

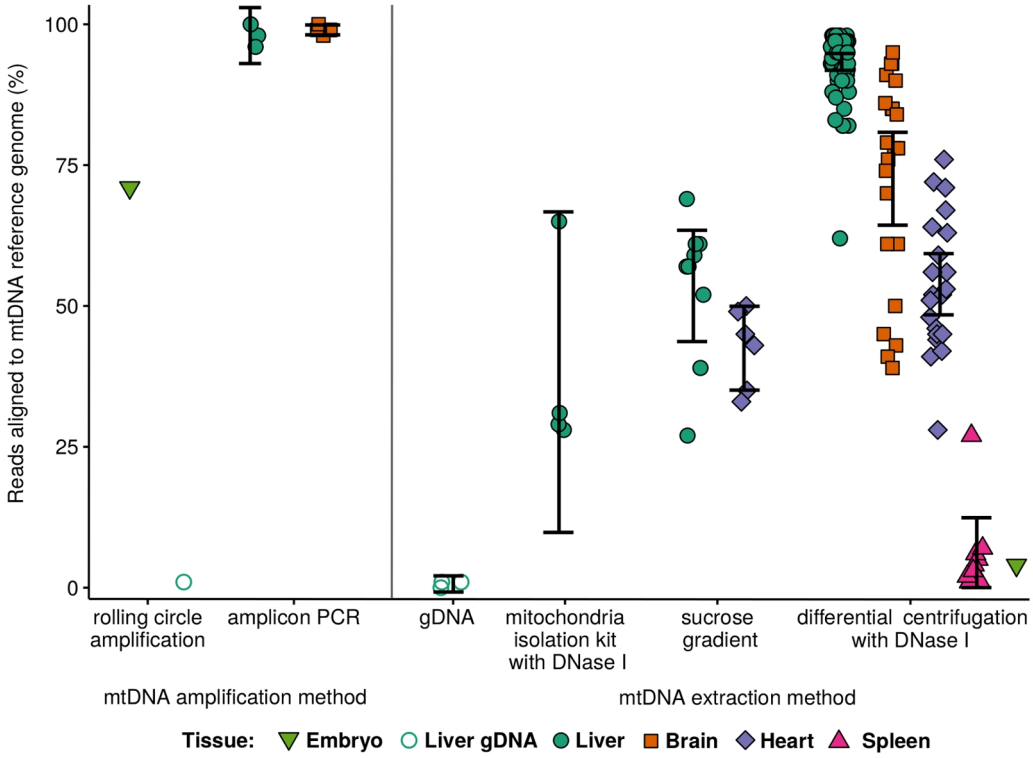

Reads aligned to mtDNA reference genome (%)
100
75
50
25
0
rolling circle amplification
amplicon PCR
gDNA
mitochondria isolation kit with DNase I
sucrose gradient
differential centrifugation with DNase I
mtDNA amplification method
mtDNA extraction method
Tissue: Embryo Liver gDNA Liver Brain Heart Spleen

Figure 4.2. Comparison of different mtDNA extraction methods by the percent of sequencing reads aligning to mtDNA reference genome. *In these thesis projects and our other studies, mtDNA samples from different mouse tissues have been obtained by different methods and sequenced with Illumina HiSeq or SOLiD sequencing (Ameur et al. 2011). To understand the level of mtDNA enrichment, the percent of sequencing reads uniquely aligned to mtDNA reference genome (mtDNA alignment) was compared between the mtDNA extraction methods. Poorly enriched samples were used as templates for rolling circle and PCR amplification. Enriched embryo mtDNA (N1 n = 1) and liver gDNA (WT n = 1) samples were used in rolling circle amplification, yielding 71 and 1 % mtDNA alignment, respectively. Amplicon PCR samples (liver MKO n = 1, WT n = 1, N2 n = 1, brain N1 n = 5) showed the highest mtDNA alignment – 99 % – of all of the evaluated methods. As expected, sequencing of genomic DNA (gDNA) showed <1 % mtDNA alignment (WT n = 3). Mitochondria isolation from liver by Miltenyi Biotec kit, followed by DNase I treatment, reached median of 30% mtDNA alignment (N1, n = 4). A traditionally used mitochondria enrichment method, sucrose gradient, showed median mtDNA alignment 57 % for liver (data obtained from Ameur et al. 2011, n = 9) and 44 % for heart samples (samples extracted and sequenced by Stanka Matic during her PhD projects, n = 6). The method optimized in this thesis (simple differential centrifugation combined with DNase I treatment) showed consistently high mtDNA enrichment for liver mtDNA samples and also brain mtDNA samples performed relatively well with median mtDNA alignments 95 and 78 %, respectively (liver MKO n = 6, WT n = 4, N1 n = 8, N2 n = 11, N3 n = 15, brain MKO n = 6, WT n = 4, N2 n = 5, N3 n = 7). However, heart mtDNA samples (samples extracted and sequenced by Johanna Kauppila during her PhD projects, n = 20) performed equally to sucrose gradient samples; median mtDNA alignment was 52 %, but the high-yield heart mtDNA samples (outliers) were obtained from mice with cardiomyopathy phenotype potentially affecting the mtDNA yield. This was despite equal amount (mg) of tissue was used for the mtDNA extraction. The optimized mtDNA-seq method was not successful in enriching spleen or embryo mtDNA, as the median mtDNA alignment was 4 % (MKO spleen n = 5, WT spleen n = 4, N2 embryo n = 1). Error bars represent 95 % confidence intervals.*

Table 4.1. Summary of different mitochondrial DNA extraction methods and their performance. *Different mitochondria isolation methods were tested on multiple mice (Samples tested). The method was evaluated based on estimated time consumed after mouse dissection (Hands-on time/total time), total yield of mtDNA obtained and amount of nuclear DNA (nDNA) contamination detected by PolgA-PCR. Good quality samples were sequenced (Seq samples) and the performance was judged by the fraction of reads aligning to mtDNA reference genome (mtDNA alignment).*

Method	Samples tested (n)	Hands-on time/total time (h)	mtDNA total yield (ng)	nDNA (*PolgA*-PCR)	RNase A treatment	Seq samples (n)	mtDNA alignment (%)
CsCl-gradient	8	4/90	50–850ND	–/++	–	–	–
Sucrose gradient	18	3/22	25–750ND	+	–	–	–
Percoll gradient	9	3/24	250^{Q}–8000ND	–	–	–	–
Rolling circle amplification (gDNA/E mtDNA)	54	2/27	n.a.	++/(+)	+	2	<1 / 71
Amplicon PCR (gDNA/mtDNA)	8	2.5/22	n.a.	n.a.	+	8	99
Mitochondria isolation kit (beads) and DNase I	11	2/24	39 –100^{Q}	++	+	–	–
Mitochondria isolation kit with DNase I (L/B/H)	40	1.5/22	150–900^{Q}	–	+	4	30
Differential centrifugation with DNase I (S/E)	20	3/24	600–1200^{Q}	++	+	10	4
Differential centrifugation with DNase I (L/B/H)	139	3/24	70–3500^{Q}	–/(+)/+	+	53	78–98

gDNA = genomic DNA as a template, S = spleen, E = embryo, L = liver, B = brain, H = heart, ND = NanoDrop (spectrophotometric), Q = Qubit (fluorometric), n.a. = not analyzed nuclear DNA (nDNA) contamination by PolgA-PCR: – = not detectable, (+) = barely detectable, + = clear contamination, ++ = highly contaminated RNase A treatment: – = no additional treatment, + = treatment of extracted DNA, Seq samples/mtDNA alignment: – = samples not sequenced

nDNA fragments in contrast to the liver gDNA sample where intact nDNA might serve as a template for the amplification. Due to time and financial constraints, no more rolling circle amplified samples were sequenced, as the method did not seem promising for low-frequency mtDNA variant detection (discussed further in **Chapter 4.2.2**).

Another amplification-based method, similar to commonly used long-range PCR method, was evaluated. mtDNA was amplified in 2–7 kb PCR amplicons either using enriched mtDNA sample or gDNA as a template and the purified amplicons were pooled in equimolar ratio to create a representative sample of the entire mtDNA. As expected, amplicon samples showed 99 % mtDNA alignment, confirming specific and efficient amplification of mtDNA. Furthermore, amplicon sequencing, especially directly from gDNA, significantly reduced the required hands-on time in comparison to mtDNA isolation methods.

Despite this successful enrichment by amplification-based protocols, methods based on mitochondria enrichment would be more favorable for the reasons outlined before (**Chapters 1.4.3** and **4.2.2**). As a reference, sequencing of liver gDNA resulted in <1 % mtDNA alignment (**Fig. 4.2**). Mitochondria isolation kit combined with DNase I treatment enriched mtDNA from nDNA only ~30x, although, the *PolgA*-PCR was negative for these samples. According to the prior-sequencing quality control at the MP-GC, these samples were fragmented, which could suggest incomplete digestion of nDNA, possibly explaining the false-negative PCR result and high fraction of nDNA reads. Others have successfully extracted mtDNA by sucrose gradient (see **Fig. 4.2** legend for details), and sequencing data from those studies was included to the comparison. Gradient-based method seemed slightly better than the mitochondria isolation kit, however, one drawback seemed to be sample-to-sample and person-to-person variation in the sample quality. And yet, the median mtDNA alignment was only 57 and 44 % for liver and heart samples, respectively.

In comparison to other methods, the differential centrifugation combined with DNase I treatment showed superior mtDNA enrichment

– especially for liver and brain mtDNA samples – resulting in up to 98 % mtDNA alignment (**Fig. 4.2**). Kidney mtDNA samples showed equal quality to liver mtDNA samples and would be expected to be highly pure mtDNA, but were not sequenced for these thesis projects. Heart mtDNA samples were consistently less pure than liver or brain mtDNA samples. This was likely due to small amount of difficult-to-homogenize tissue available, inevitably leading to the breakage of mitochondria as well as presence of more broken nuclei during the homogenization, and thus to the higher level of nDNA contamination in the final sample. However, the method was not successful for mtDNA enrichment from spleen or embryo. Despite extensive optimization of the tissue homogenization and nDNA digestion steps, only 4 % mtDNA alignment was obtained (**Fig. 4.2**). In order to understand whether the non-mtDNA reads truly originated from contaminating nDNA and not from for example artefacts in sequencing library preparation (e.g. significant adapter dimer formation), reads were also aligned to the full mouse reference genome with exactly the same alignment parameters as was used for the alignment to mtDNA reference genome. Indeed, >99% of the reads were aligned to the full mouse reference genome confirming that the low proportion of mtDNA reads was caused by failed mtDNA enrichment process, and thus significant nDNA contamination.

Due to consistent and significant purity of liver mtDNA samples, as well as simplest hands-on steps without a potentially error-inducing amplification step, differential centrifugation combined with DNase I treatment was chosen as the most promising method to study low-frequency mtDNA variants (hereafter denoted as **mtDNA-seq**). The key steps to diminish nDNA contamination were mild homogenization and DNase I treatment of the mitochondria. Also, extensive RNase A treatment of the mitochondria, as well as the final mtDNA prep was crucial because significant RNA contamination occurred, and caused overestimation in the spectrophotometric concentration measurement and inhibited *PolgA*-PCR, giving false-negative results regarding the nDNA contamination. As an alternative, amplicon sequencing may be a potential method for difficult-to-enrich samples or as a fast method for

studies which do not require extremely sensitive variant detection (for further discussion see **Chapter 4.2.2**). For the studies presented in this thesis, liver (as a mitotic) and brain (as a post-mitotic) were selected as representative tissues.

4.2 Selection of the sequencing method for low-frequency mitochondrial DNA variant detection

Different sequencing technologies have been used to detect mtDNA mutations and more accurate sequencing approaches are constantly being developed. As Duplex-Sequencing was the most accurate method at the time of starting these thesis projects (Fox et al. 2014), the first sequencing experiments were conducted in collaboration with Finnish Institute for Molecular Medicine (FIMM) Technology Center (Finland) in order to set up a similar, UMI-based method. However, the experimentation quickly showed – a topic of which was also touched by Kennedy et al. (2014) – that the method may often require sample-to-sample optimization, which would always include sequencing. Thus, this approach was assumed to wind up as a very expensive and time-consuming approach, unless one has a direct access to a sequencing platform. As in-house Illumina sequencing was not a possibility for our group, the suitability of less sensitive sequencing approaches was investigated in detail.

First, this chapter focuses on important sequencing data analysis steps which were optimized keeping in mind the characteristics of mtDNA genome. The data analysis workflow was developed along with the mtDNA enrichment method optimization. Next, this chapter presents the evaluation of variant calling results obtained from different mtDNA enrichment experiments in order to choose the most optimal approach in terms of sensitivity, accuracy, hands-on time and costs. Finally, the validation of the optimized protocol utilizing spike-in control samples is presented.

4.2.1 Optimization of the data analysis steps suitable for circular mitochondrial genome

Still today, there is no ”standard” data analysis procedures for analyzing high-throughput sequencing data from mtDNA samples. A survey of the literature reveals that analysis details are often not documented in enough detail to be reproducible, or they neglect the physical characteristics of mtDNA leading to poorer results. Key steps in the reliable low-frequency mtDNA variant detection are:

1. Quality and adapter trimming should be always carefully applied for sequencing reads,
2. mtDNA is circular and nDNA contain even 100 % identical NuMTs,
3. mtDNA is a small genome and duplicate reads are likely to naturally occur, and
4. variant calling thresholds often applied are not able to detect low-frequency variants.

Read trimming

Sequencing read trimming is probably the most important step in the sequencing data analysis. Low-quality bases or sequencing adapter leftovers can appear as false sequences in genome assemblies or cause poor alignment results. For example, the *Cyprinus carpio* reference genome (Xu et al. 2014) appers to contain Illumina TruSeq adapter sequences due to vague trimming steps before the genome assembly – an accidental discovery when analyzing non-mtDNA reads from a failed library preparation consisting mainly of Illumina TruSeq adapter dimers. This observation was further supported by others (Etherington 2014, accessed 08/2017). Despite the clear advantage of read trimming, it is not always an included step, or the reporting of this step is neglected. For example, complete human mtDNA data analysis pipelines, such as MToolBox (Calabrese et al. 2014) or mtDNA-Server (Weissensteiner et al. 2016), do not include the read trimming into their data pre-processing steps. They probably expect the data to be pre-processed by the user, thus, reducing the reproducibility of the results.

Tools and parameters used for read trimming should be carefully determined (Del Fabbro et al. 2013). Flexbar (Dodt et al. 2012) was chosen for the projects presented in this thesis. In addition to adapter and quality trimming of the reads, it is one of the rare tools capable of removing barcodes (Jiang et al. 2014), which was a required process for the amplicon sequencing approach used in these thesis projects. Trimming parameters, however, required optimization since the recommended default parameters lead to fuzzy results in mtDNA read trimming (**Fig. 4.3**).

A key reason for the fuzzy results was the required minimum overlap of three nucleotides with the adapter sequence (similar default paramters also in another common tool, Cutadapt [Martin 2011]), and moreover, flexbar allows for even three mismatches per ten nucleotides to trim an adapter. As illustrated in **Figure 4.3** (red line), this leads to the trimming of true mtDNA sequences from the reads, in comparison to the much smoother coverage obtained with more stringent trimming parameters (grey line). Thus, the final parameters for mtDNA-seq data analysis were minimum adapter overlap of ten nucleotides, allowing only a single mismatch, and additionally, in order to increase the reliability of alignment and variant calling, reads were trimmed for minimum length of fifty nucleotides with a minimum base call quality Phred score 28. Many tools, and thus also their default parameters, have been developed with a focus on nuclear genome. These results emphasize that, despite the suggestions in manuals of many bioinformatic tools, the default parameters should not be carelessly applied – at least not to mtDNA data analysis.

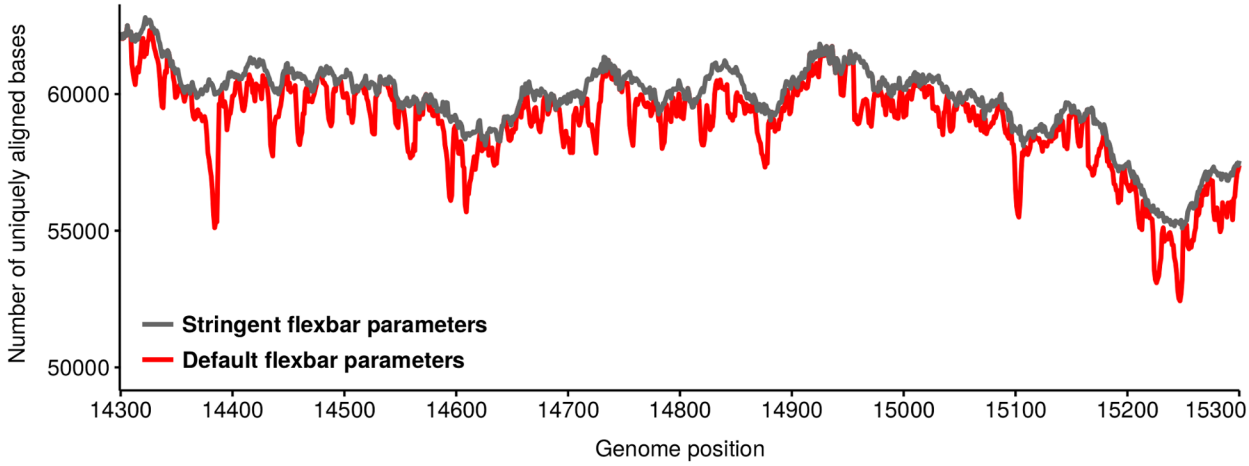


Figure 4.3. Effect of read trimming parameters on coverage. *Although bioinformatic tools often recommend the use of default parameters, the suitability of those for the data set in question should be carefully addressed. For example, the default trimming parameters of flexbar require only a 3-nt overlap and allow 3 mismatches per 10-nt overlap with the adapter sequence in order to remove it from the read. This would lead to artefactual drops in coverage (red line) in comparison to more stringent requirement of minimum 10-nt overlap allowing 1 mismatch (grey line) used in the projects presented in this thesis. Only a short part of the genome coverage is illustrated as the trend was similar for the entire mtDNA genome. The maximum per base coverage difference was ~5000 bases (i.e. 8 % less) and in total two times more bases were trimmed off with the default parameters in comparison to stringent parameters. In both trimming approaches, adapter sequences were detected from both ends of the reads (`-ae ANY`) and minimum base call Phred score (`-q`) was 28. The stringent parameter set additionally required minimum length of 50 nt for the read whereas default is 18 nt. The example data is from a WT liver mtDNA sample.*

Alignment

Genome circularity. Most read aligners are designed for linear genomes and the alignment performance is poor at the edges of the linear genome leading to a drastic coverage drop (**Fig. 4.4**). For most applications, the interest does not focus on the chromosome ends; however, for a circular genome such as mtDNA, this means poor coverage and difficult variant calling at the genome junction region potentially leading to false conclusions. Some studies (e.g. Kennedy et al. 2014, Ni et al. 2015, Zhang et al. 2016) have not considered the mtDNA circularity in their alignment steps. Whereas, in many publications, the circularity is solved by different approaches, such as adding certain amount of bases from the beginning of the genome to the end of the genome (e.g. Ameur et al. 2011; Li et al. 2015; Vellarikkal et al. 2015) or by a ”dual approach” such that the reads are aligned to the normal reference genome and to a split reference genome, in which the first half of the genome sequence is transferred to the end of the genome (e.g. Ding et al. 2015). However, the former approach could lead to biased alignment since the added genome region is present twice in the reference genome and reads may align to multiple places.

The easiest solution for circular genome alignment is to align the reads with a splice-aware aligner, e.g. `bwa mem` (Li 2013), and to optimize the parameters such that read splicing is not heavily penalized, i.e. allow reads to align almost equally likely to the genome junction region as to the other parts of the genome. **Figure 4.4** illustrates the difference of optimized `bwa mem` splice-aware alignment (black, parameters `-T 19 -L 5,4`) in comparison to `bwa aln` non-splicing alignment (red, default parameters) commonly used for DNA reads. Results produced with another non-splicing aligner, `bowtie2` (local alignment mode with default parameters, Langmead & Salzberg 2012), were highly similar to `bwa aln`. Even though the coverage was well-rescued with `bwa mem`, the alignment was further optimized as variant calling failed to detect variants in the first and last five bases of the genome, probably due to the proximity of the read end decreasing the reliability of the variant call.

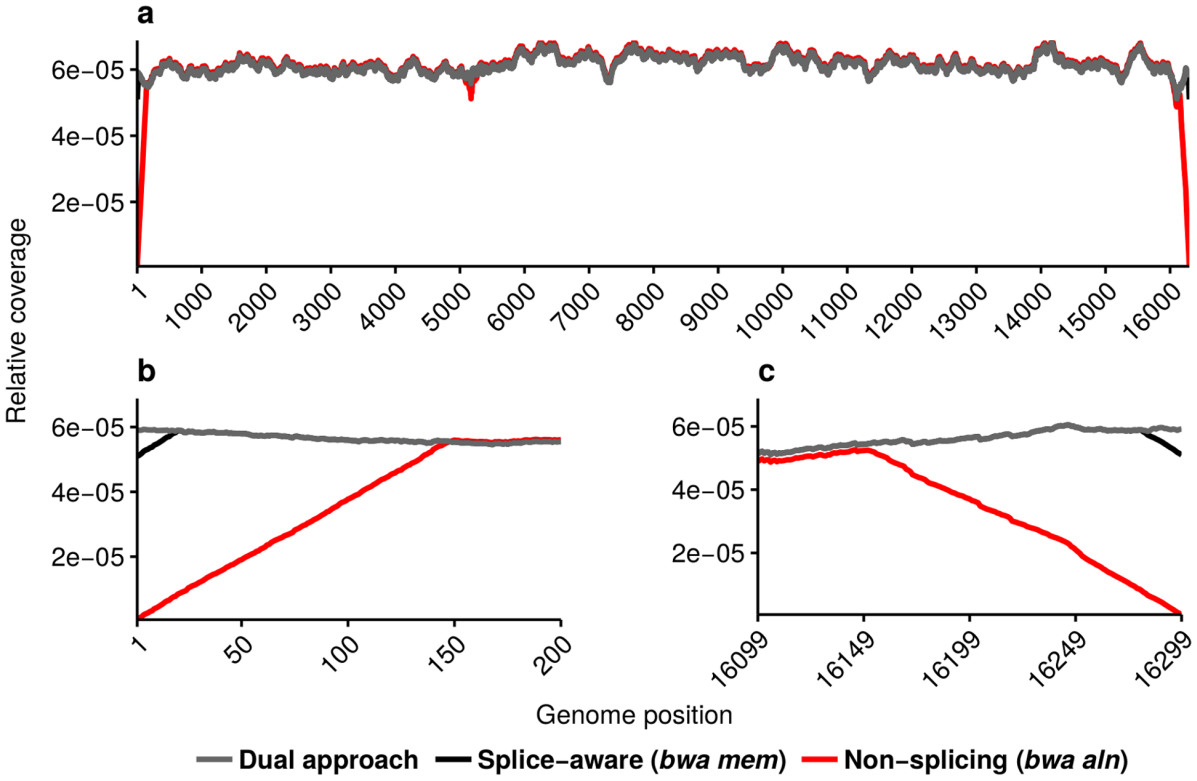

a
b
c
Relative coverage
6e−05
4e−05
2e−05
1
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
50
100
150
200
16099
16149
16199
16249
16299
Genome position
Dual approach
Splice−aware (bwa mem)
Non−splicing (bwa aln)

Figure 4.4. Effect of alignment tool and strategy on coverage. *Relative coverage (per base coverage/total number of bases aligned to mtDNA reference genome) obtained by different alignment strategies is illustrated* ***a)*** *over entire,* ***b)*** *at the beginning and* ***c)*** *at the end of the mtDNA genome. Non-splicing aligners (e.g.* `bwa aln`*, default parameters, red line), failed to align reads at the edges of the linear reference genome, whereas splice-aware alignment strategy with* `bwa mem` *rescued most of the reads (parameters* `-T 19 -L 5,4`*, black line). Coverage of the first and last five bases of the genome was further rescued by dual alignment strategy: reads were aligned to both the normal and split reference genomes. Coverage results from the alignments were combined such that results comprising ±200 positions around the junction were kept from the split alignment, and results for the rest of the genome positions were from normal reference genome alignment. The data here is from a WT liver mtDNA sample.*

The final approach was to not only to use `bwa mem` optimized alignment (**Fig. 4.4**, black), but to align the reads separately to a split reference genome (first 8150 bp of the genome transferred to the end of the genome, as in Ding et al. 2015). The final resulting coverage and variant calling files were created by combining results from the two separate analyses such that genome positions 200–16099 were from the normal reference genome analysis, and the rest of the genome positions, i.e. the junction region of the normal reference genome, were obtained from the split reference genome analysis. Thus, the resulting coverage at the genome junction region (**Fig. 4.4,** grey) was rescued and, more importantly, variant detection succeeded on the first and last bases for MKO mtDNA samples. This approach is considered less biased than the published ones, as there are no longer artefactual gaps in the coverage and in both alignments the reads are aligned to the full mtDNA genome allowing the use of only uniquely aligned reads in downstream analyses. In addition to Ding et al. (2015), similar alignment strategy was recently applied also by Kelly et al. (2017). The only difference to the published approaches was that they used even ±4 kbp junction region, whereas here only ±200 bp was considered to be enough as the read length was 150 bp at maximum.

Nuclear mitochondrial sequences. Another alignment issue is caused by NuMTs. The full mouse reference genome contains 100 % identical regions in both the chromosomes and the mtDNA genome. Thus, the often applied alignment strategies in which the reads are aligned to full mouse reference genome would cause true mtDNA reads to align to multiple positions or more conservatively, the reads are first aligned to nuclear reference genome and only unmapped reads are aligned to the mtDNA reference genome (Guo et al. 2013, Zhang et al. 2016). As it is advisable to keep only the uniquely aligned reads for more reliable variant calling, in comparison to alignment to mouse mtDNA reference genome alone (**Fig. 4.5**, grey line), a significant proportion of the reads would be filtered out from the data set if the reads were aligned to full mouse reference genome (**Fig. 4.5**, red line). Therefore, for mtDNA

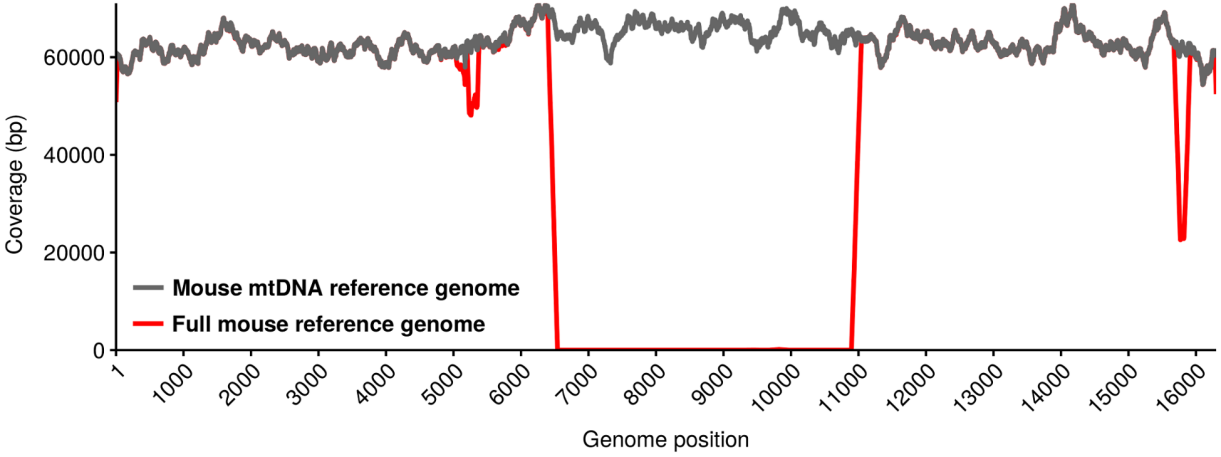


Figure 4.5. Effect of reference genome and alignment strategy on coverage. *Presence of up to 100 % homologous NuMTs in the chromosomes affect the coverage, if full mouse genome is used as a reference for alignment and only uniquely aligned reads (red line) are kept for the downstream analysis. If only the mouse mtDNA genome is used as the reference genome for alignment (grey line), uniquely aligned reads can be kept for downstream analysis without a huge loss of data. The data here is from a WT liver mtDNA sample.*

variant studies, it is recommended to align reads only to the mtDNA reference genome.

De-duplication

Sequencing data analysis steps often include de-duplication, which means removal of duplicate reads. Rationale behind de-duplication is that from a complex DNA sample, e.g. gDNA, it is unlikely to obtain exactly the same DNA fragment multiple times, and thus, duplicate reads are considered likely to originate from the PCR amplification of the library preparation, or as optical duplicates during sequencing (with non-patterned flow-cell type). Duplicate reads could also skew the variant calling results. De-duplication is suggested in the GATK Best practices (van der Auwera et al. 2013), which are often used as golden guidelines for sequencing data analysis. However, de-duplication is a controversial step (Dyer et al. 2015) even for standard deep sequencing data analysis due to difficulties to identify truly artefactual duplicates from natural duplicates (Bainbridge et al. 2010; Niu et al. 2010; Balzer et al. 2013). True duplicate identification would only be enabled by DNA fragment tagging with UMIs, e.g. by Duplex Sequencing (Kennedy et al. 2014), yet, even that may be difficult due to potential artefacts caused by index switching.

Commonly used de-duplication tools are samtools rmdup (Li et al. 2009) and Picard MarkDuplicates (Broad Institution, accessed 08/2017). Both of these tools define a duplicate read based on the 5' end positions of the aligned reads only, with an assumption that duplicated reads have identical starting positions. The 3' end of the read is not considered because that is often trimmed or the base call quality, and thus also alignment quality, is often decreasing towards the end of the read. Therefore, instead of blindly implementing the established guidelines also to the mtDNA sequencing data analysis, careful consideration of mtDNA genome characteristics is needed: When millions of reads are produced from highly pure 16.3-kb mtDNA sample, it is likely that the exactly same sequence occurs more than once in the data set. Roughly

estimated, from 1 Gbase of 150-bp sequencing reads i.e. ~6.7 M reads, there should be ~400 reads starting from each genome position (~$6.7x10^6$/16300 ≈ 400). Hence, de-duplication would remove likely occurring natural duplicates and is not a recommended step for mtDNA or other small-genome data analyses.

Variant calling

Many mtDNA variant studies, especially earlier ones, use relatively high variant calling thresholds for allele frequency (AF, 0.5–10 %) due to low read coverage (often ~100–1000x for human data), or PCR amplification of the DNA, but also because lower detection thresholds have not been considered reliable (e.g. Li et al. 2010; Ameur et al. 2011; Calabrese et al. 2014; Just et al. 2014; Ding et al. 2015; Pyle et al. 2015; Vellarikkal et al. 2015) without special approaches (e.g. PELE-seq reaching reliable detection of AF 0.2 % [Preston et al. 2016]). Highly sensitive detection is rather obtained by much more complicated methods, such as utilizing UMIs (Kennedy et al. 2014) or by circle sequencing (Lou et al. 2014). Since these complicated methods were considered too time-consuming and expensive for the projects presented in this thesis, other means to accomplish sensitive variant detection from standard sequencing were required.

For viral studies, variant calling methods capable of detecting variants even below the sequencing error rate have been developed (Wilm et al. 2012; McElroy et al. 2013; Verbist et al. 2015). The variant caller chosen for these thesis projects, LoFreq* (Wilm et al. 2012), is fast, sensitive and specific (McElroy et al. 2013; Huang et al. 2015). Furthermore, no assumptions on ploidy is included in LoFreq* making it an ideal variant caller also for heteroplasmic mtDNA. Wilm et al. (2012) showed with experimental data that LoFreq* reliably detects variants at AF 0.5 %, moreover, with simulated data, they detected variants even at AF 0.05 %, and the false-positive rate was impressively low – $<5x10^{-7}$.

During these thesis projects, experimentation with LoFreq* variant calling parameters – mainly removal of the hard-coded variant filtering

thresholds – with WT mtDNA-seq samples showed only handful of variants, most of which were likely true heteroplasmies. The average total mutation load was only $7.6x10^{-6}$ mut/bp (SD $1.8x10^{-7}$ mut/bp, WT liver and brain mtDNA-seq samples, n = 8). In contrast, even >17000 variants were detected per MKO mtDNA-seq sample (median variant allele frequency was <0.1 %) and total mutation loads were $1.4x10^{-3}$ mut/bp (SD $1.1x10^{-4}$ mut/bp, n = 6) and $6.7x10^{-4}$ mut/bp (SD $1.2x10^{-4}$ mut/bp, n = 6) for liver and brain mtDNA-seq samples, respectively. These results were very comparable to previous results obtained by post-PCR cloning and sequencing of tail biopsies by Ross et al. (2013): $2.2x10^{-5}$ mut/bp and $6.6x10^{-4}$ mut/bp for WT and MKO mice, respectively. The results for WT and MKO mtDNA-seq samples are discussed in detail in **Chapter 4.3**.

In summary, these promising results suggested that the data analysis protocol, especially LoFreq* and very relaxed variant calling parameters, allow highly sensitive and accurate variant detection – at least for highly pure mtDNA-seq samples. The optimized data analysis protocol was applied to evaluate the different mtDNA enrichment methods (see **Chapter 4.2.2**), and finally validated utilizing spike-in samples (see **Chapter 4.2.3**).

4.2.2 Selection of the mitochondrial DNA enrichment and sequencing method

First in this section, the results from different mtDNA enrichment methods – rolling circle amplification, amplicon PCR, mitochondrial DNA isolation kit with DNase I treatment and differential centrifugation with DNase I treatment (mtDNA-seq) – are evaluated based on the variant results obtained by the optimized data analysis protocol. Then, this section will focus on the improvement of mtDNA-seq method in order to diminish the observed major artefacts and further increase the accuracy of the method.

Comparison of the different mitochondrial DNA enrichment methods

The mtDNA enrichment methods were optimized in parallel yet here, the variant results are compared by using mtDNA-seq samples as ”standards”, because mtDNA-seq was assumed to be the least error-prone method based on high mtDNA enrichment without amplification and the observed very low variant load in WT samples. The key evaluation criterion was the observed variant profile (i.e. proportion and load of different variant types observed) for each sample (**Fig. 4.6**). The variant profile could reveal odd samples such as those likely containing artefactual variants e.g. unexpected load of GC>TA variants (Costello et al. 2012; Chen et al. 2017). Some variation between the different mouse lineages could also arise from clonal expansion of some variants but not the others within a litter or an individual mouse.

Rolling circle amplification. Lou et al. (2014) have successfully utilized RCA approach for circlized short mtDNA fragments to sensitively detect mtDNA variants. However, Ni et al. (2015) and Marquis et al. (2017) have suggested direct use of RCA for mtDNA variant detection – escpecially valuable approach for precious, low-yield samples, since the amplification is possible from <1 ng of DNA. Due to the nature of RCA – i.e. the same circular template molecule is amplified multiple times – polymerase errors should not highly expand and less false-positive variants should be detected in comparison to regular PCR amplification where the polymerase errors may be exponentially amplified. Thus, RCA approach would result in more reliable mutation detection, although, previous studies have used relatively high – 0.3 and 1 % – variant detection thresholds (Ni et al. 2015; Marquis et al. 2017).

Since extraction of highly pure mtDNA from spleen or embryo was not successful, suitability of rolling circle amplification was investigated for these DNA samples. WT liver gDNA and poorly enriched embryo mtDNA were amplified by rolling circle with human mtDNA primer mix. Only the embryo mtDNA sample was highly enriched for mtDNA

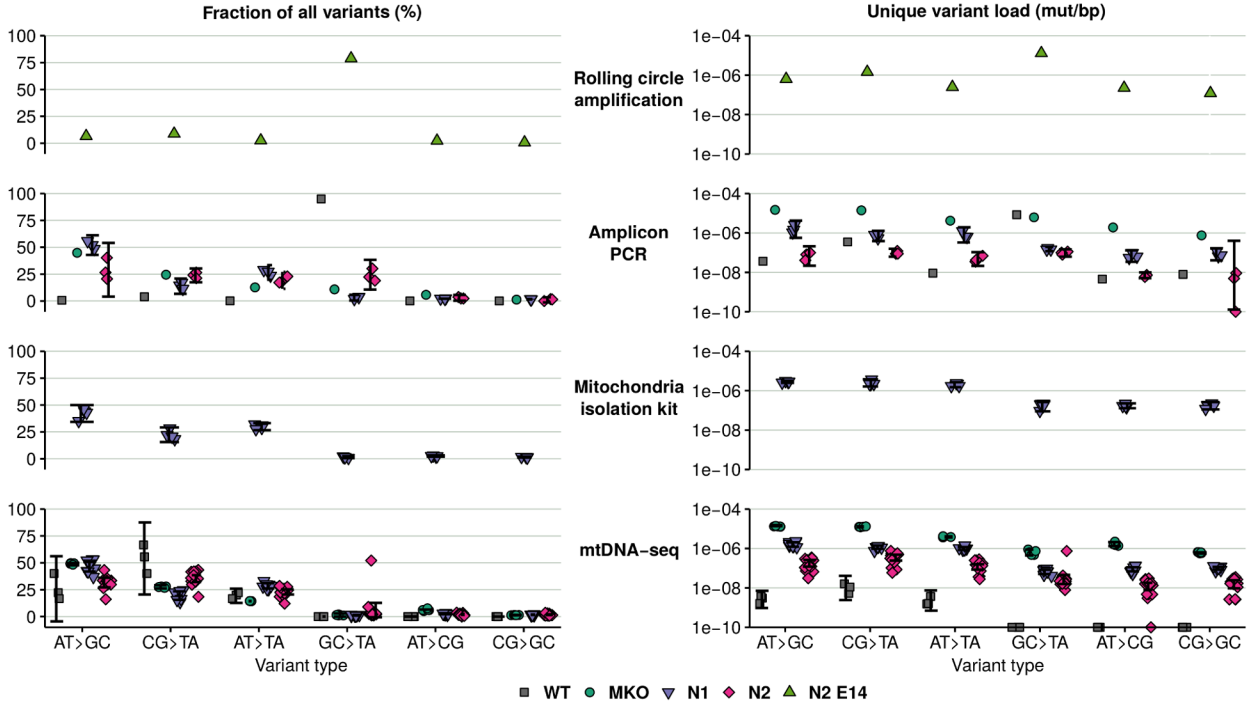

Fraction of all variants (%)
Unique variant load (mut/bp)
Rolling circle amplification
Amplicon PCR
Mitochondria isolation kit
mtDNA−seq
100
75
50
25
0
1e−04
1e−06
1e−08
1e−10
AT>GC
CG>TA
AT>TA
GC>TA
AT>CG
CG>GC
Variant type
WT
MKO
N1
N2
N2 E14

Figure 4.6. Comparison of variant profiles obtained by different mitochondrial DNA enrichment methods. *The variant results obtained by sequencing samples prepared by different mtDNA enrichment methods were compared. The results from mtDNA-seq were considered as "standards", because the method was assumed to be the least error-prone method based on high mtDNA enrichment without amplification and the very low observed variant load in WT samples. When fractions of different mutation types (left panel) are compared, samples containing a high amount of the most likely artefactual GC>TA variants are easily visualized, e.g. embryo sample amplified with rolling circle (N2 E14, light green) or WT (grey) amplicon sample as well as single N2 (pink) mtDNA-seq samples. Moreover, unique variant load (right panel) supports the variant profile by showing more clearly, for example, that MKO amplicon sample carried ~10 times higher load of GC>TA variants than MKO mtDNA-seq samples. The unique GC>TA variant load of N2 amplicon samples was significantly different from N2 mtDNA-seq samples (only brains, p = 0.011, Welch two sample t-test), however, such difference could originate either from the mtDNA enrichment method or due to expansion of certain variants within a mouse lineage (all N2 amplicon samples were from a single litter, whereas N2 mtDNA-seq samples originated from several maternal lineages). Interestingly, N1 mtDNA samples, obtained by any method, always showed similar variant profile, despite the fact that the samples enriched with mitochondria isolation kit were severely nDNA contaminated as illustrated in* ***Figure 4.2****. The number of samples from various tissues: rolling circle N2 E14 n = 1 (embryo), amplicon PCR WT n = 1 (liver gDNA); MKO n = 1 (liver); N1 n = 3 (brain); N2 n = 3 (brain 2, liver 1), mitochondria isolation kit N1 n = 4 (liver), mtDNA-seq WT n = 3 (liver); MKO n = 6 (liver); N1 n = 8 (liver); N2 n = 16 (brain 5, liver 11). Error bars represent 95 % confidence intervals.*

reads (**Fig. 4.2**) and analyzed further. The embryo was obtained from N1 mother, thus carrying maternally transmitted mutations, yet, detection of >2500 GC>TA variants (representing ~80 % of all detected variants, median AF 0.01 %) was surprisingly high (**Fig. 4.6**, left panel). MKO (see **Chapter 4.3** for details) or *PolgA*$^{D275A/D275A}$ (Ameur et al. 2011) mice do not show bias towards transversions, and moreover, in our hands, N1 mtDNA-seq samples showed only median of 29 GC>TA variants with median AF 0.2 % (i.e. median 1.2 % of all variants, n = 8, **Fig. 4.6**). Thus, the embryo mtDNA sample (n = 1) seemed to harbor a significant artefactual GC>TA variant load. Whereas if, similar to Ni et al. (2015), minimum AF was set to 0.3 %, only 30 GC>TA variants were detected, which would be more expected result considering the maternal mtDNA variant load.

Rolling circle amplified DNA is a tree-like, branched structure and during the long amplification step (even 8–16 hours) the DNA may stay single-stranded for a prolonged time potentially exposing the DNA for damage. As noted by Costello et al. (2013), presence of a contaminant in the DNA sample during the sonication step of the library preparation may induce oxidative damage. Such contaminant could be a leftover due to an incomplete purification of the sample after the amplification reaction. Another issue regarding the RCA method is that the amplification is supposedly highly efficient on a circular template but less on a linear template, thus, the method would fail to accurately represent the linear, truncated mtDNA molecules present in MKO mice (e.g. Hämäläinen et al. [2015] and Ni et al. [2015] did not observe mtDNA alignment coverage variations that are diagnostic for the linear mtDNA fragments in homozygote mtDNA mutator mouse tissue or cell culture samples).

Taken together, the level of mtDNA enrichment by rolling circle amplification was not impressive, extremely high load of artefactual GC>TA variants were detected, and accurate representation of various mtDNA molecule types is questionable. These results, although based on a single sample due to time and financial constraints, raised

suspicions towards suitability of rolling circle amplification for low-frequency mtDNA variant detection, so sequencing experiments with RCA method were not continued. It is, however, impossible to conclude based on a single sample whether such burst of artefactual variants was only a random event or more common limitation of the RCA method.

Amplicon PCR. Amplicon PCR would enable sequencing of low-yield mtDNA samples or even simplify the DNA extraction step to easy genomic DNA extraction with easy-to-use kits, directly from dissected or frozen tissue sample. An important concern, however, is the risk of unintended amplification of 100 % homologous NuMTs regions, which is extremely difficult to exclude despite careful primer designs. PCR-based methods are also thought to be prone for early-cycle polymerase errors, which would be indistinguishable from real variants. Payne et al. (2015) stated that, even with extreme-depth sequencing, the variant detection is only possible down to AF 0.1–0.2 % due to the error rates of the sequencing platforms.

Amplicon sequencing was first tested with normal PCR primers using low-yield mtDNA samples obtained during the mtDNA enrichment optimization as templates: N1 brain (n = 3) and N2 liver (n = 1) and brains (n = 3). Not so surprisingly, these samples showed high coverage bias at the primer sites. Thus, the amplicon approach was improved by adding a non-mitochondrial tag sequence to the primers (M13F or M13R, see **Chapter 3.6.2**), allowing efficient removal of the first PCR-cycle reads and normalizing the coverage at the primer sites (**Fig. 4.7**). The amplicon overlaps and lengths were also increased such that only three amplicons were enough to cover the entire mtDNA genome. Templates for the improved approach were WT liver gDNA (n = 1) and MKO liver mtDNA-seq (n = 1) samples.

The mutational profile from N1 and N2 amplicons resembled those obtained by mtDNA-seq, however, GC>TA variants were slightly overrepresented in N2 samples. Only N2 mice from a single litter were used for amplicon PCR, thus, it is not possible to conclude whether the observed increase in GC>TA variant amount was due to the

amplification method or expansion of certain variants inherited from the mother. Similar to RCA, both the WT and MKO amplicon samples showed >2000 GC>TA variants representing 94 and 11 % of all observed variants, respectively. This was clearly an artefact, since for WT, no GC>TA variants were expected. Moreover, the same MKO template DNA was also sequenced by mtDNA-seq in which only 1 % of the detected variants were GC>TA, similar to what was observed for other MKO mtDNA-seq samples (**Fig. 4.6**).

Even if GC>TA variants were ignored, the WT amplicon sample was left with 129 variants (median AF = 0.09 %), most of which were CG>TA transitions – a common error introduced during PCR (Chen et al. 2014). If GC>TA variants indeed arise due to DNA damage, it is impossible to rule out that other observed variants would not be induced by the same damaging factor. Although, the WT total variant load without GC>TA variants was only $1.3x10^{-5}$ mut/bp – similar level to what has been observed with WT mtDNA-seq samples or by other methods (Ross et al. 2013) – the variants occurred at several positions and were of low frequency. This suggested these variants to be false-positive results rather than true heteroplasmies, which would be seen as few, higher frequency variants. If the minimum AF was set to 0.3 %, only 6 variants were detected, which resembled the observation with WT mtDNA-seq samples. However, for the MKO amplicon sample such threshold would leave only 710 variants, whereas mtDNA-seq detected ~17000 variants for the same MKO sample. Such stringent AF threshold apparently improved the variant calling precision at the cost of sensitivity.

Since significant amount of artefactual GC>TA variants were observed only in two amplicon samples (WT and MKO), but not in six other samples (N1 and N2), it cannot be concluded to be caused by the amplicon PCR itself. However, as suggested for RCA, it is possible that sometimes an unknown contaminant (Costello et al. 2013), for example due to incomplete PCR purification, may induce DNA damage during sonication. Based on observations made during these thesis projects, it is

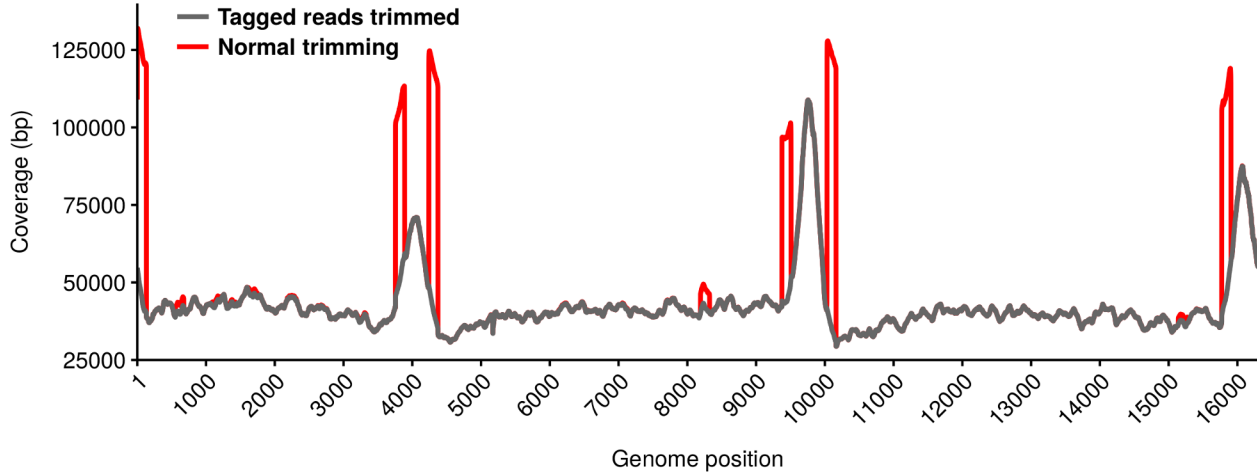


Figure 4.7. Effect of the tagged amplicon read removal on coverage. *In amplicon sequencing, the primer regions are inevitably overrepresented (red peaks) and are difficult to trim off without also removing otherwise normal genome sequences. When the primers included a non-target tag sequence (here M13 forward or reverse primer sequence was added to the 5' end of the amplicon primers), the first-cycle reads could be trimmed off of the data set (grey line) to normalize the coverage to contain only the natural amplicon overlap peaks. Additional tag trimming was conducted with flexbar option to remove barcode sequences (parameters* `-bo 15 -bt 2 -bu -be ANY`*), which specifically separated the tag-containing reads from the rest of the reads. The example data is from WT liver gDNA amplicon sample.*

tempting to suggest that amplification-based methods would be more prone to the occurrence of such artefacts. Presence of other artefactual variants induced by the method itself can neither be confirmed nor excluded based on a single WT amplicon sample. Until more control experiments with amplicon PCR are made, the only putative conclusion here is that the amplicon PCR is a suitable mtDNA enrichment method for various samples, but the variant detection threshold has to be set at minimum AF of ~0.3–0.5 %. Thus, extremely low-frequency mtDNA variant detection with amplification-based methods does not seem possible with standard Illumina HiSeq sequencing. It can be further hypothesized, that the limitation of the detection threshold in amplicon sequencing is due to artefacts induced during the sample processing steps.

Mitochondrial isolation kit. As illustrated in **Figure 4.6**, the N1 samples enriched by mitochondrial isolation kit and treated with DNase I, did not differ from N1 samples enriched by other methods. Despite the substantially higher nDNA contamination, these samples did not show more potential NuMT variants than other N1 samples: This was tested by a BLAST search (Altschul et al. 1990) of mtDNA against *Mus musculus.* A large, >3-kb NuMT region was identified in chromosome 2 showing 72 variants in comparison to mtDNA genome. Of these NuMT variants, 5–12 were detected in the N1 samples, all samples showing similar results despite the extraction method. Thus, these samples were used along with the other N1 samples (**Chapter 4.3.2**) although they were likely fragmented (as discussed in **Chapter 4.1**).

The reason for the DNA fragmentation remained unclear. One possible factor could be that samples prepared with this method were always dissolved into nuclease-free water, and it is known, that the pH of water may sometimes be too acidic for stable, long-term DNA storage. However, during the course of the mtDNA enrichment method optimization, other samples were also stored in water and poor sample quality occurred only with the mitochondrial isolation kit. It can be speculated that another factor, e.g. a contaminant leftover from the kit

reagents, also contributed to the fragmentation. Since the sample quality could not be improved, experimentation with this method was stopped. However, the variant profile comparison suggested that the method has potential for low-frequency mtDNA variant detection. Especially difficult-to-enrich samples, such as heart mtDNA, may benefit from the kit enrichment. The method still requires optimization, and testing with WT control samples is highly recommended.

mtDNA-seq. As briefly mentioned before, mtDNA-seq was considered the most reliable method, despite the fact that standard Illumina HiSeq sequencing is said to have up to a 1 % error rate. The accuracy of mtDNA-seq was further supported by the single-end sequencing results (discussed in detail in **Chapter 4.3**), which showed very precise variant detection from WT samples, and also highly sensitive variant detection from MKO samples (**Fig. 4.6**). Together these data suggest that the mtDNA-seq method indeed is very accurate, and the sensitivity is satisfactory for the research questions addressed within these thesis projects. The precision and sensitivity of the mtDNA-seq was validated by spike-in samples as discussed in **Chapter 4.2.3**. As a short conclusion, mtDNA-seq was chosen as the optimal method for low-frequency mtDNA variant detection studies, however, further improvements to the method were required due to a sudden introduction of artefactual variants as discussed next.

Improved method to diminish the occurrence of artefactual variants

After tens of successfully sequenced mtDNA-seq samples, an increased load of GC>TA variants was noted, similar to RCA and amplicon PCR samples. The number of artefactual variants varied from handful to hundreds or thousands (e.g. **Fig. 4.6**, N2 outlier sample). Simultaneously, between-library cross-contamination reached an intolerable level: instead of maximum a handful of likely cross-contaminating variants in the earlier successful samples, >100 cross-contaminating variants were detected in some samples. A variant was considered a cross-contamination if it was present at low-frequency in a sample in which it was not expected (e.g. WT), and it was present at

high-frequency in another sample included into the same sequencing run.

In order to understand the source of of GC>TA variants, a series of WT mtDNA-seq samples were re-sequenced by varying the input DNA amount (50 ng or 100 ng) with old or upgraded library preparation kit (NEBNext® Ultra™ and NEBNext® Ultra™ II Library Prep Kit for Illumina, New England Biolabs, Inc.), since these were the only changes in the library preparation that had taken place between the successful and artefactual samples at the sequencing facility responsible of the library preparation and sequencing (MP-GC). Furthermore, since single-end libraries were always prepared with dual multiplexing primers although they were sequenced in a single-end run mode. Now, an additional test was to sequence the samples in paired-end run mode, which enabled the read de-multiplexing based on both indices. Dual de-multiplexing enables the removal of cross-contaminating reads originating from mix up of indices during sequencing – more likely to happen with a single index, but much less likely with two indices (already suggested by Kircher et al. [2012]). Additionally, 1 mM of EDTA was added to the samples for sonication step, suggested by Costello et al. (2013) as a potentially effective measure to reduce oxidative damage to the DNA. In order to detect potential between-library cross-contamination, one sample harboring known high-frequency mutations was included into the otherwise WT sample set.

To resemble the single-end sequencing approach of the successful samples, only read 1 (R1) de-multiplexed with single index was used for the analysis. All of the tested library preparation conditions showed similar results, and intriguingly, no artefactual GC>TA variants were detected in this control experiment. In order to understand the difference between the successful and artefactual samples (e.g. if the number of GC>TA variant reads was just below the filter threshold in the successful samples and slightly above the threshold in the artefactual samples), non-filtered variant results were compared in addition to

normally filtered variant results (**Fig. 4.8**). The comparison revealed inherent difference between the GC>TA variant profiles in the successful and artefactual samples. This difference does explain why the variants were detected in the artefactual samples but filtered from the successful samples. Why the identically extracted mtDNA samples showed such different variant profiles is not understood. Moreover, one of the successful control samples was simply a technical library preparation replicate from the exact same mtDNA sample aliquot showing artefactual results earlier (**Fig. 4.8**, 2508.B and 2399.I), illustrating that the occurrence of the artefact is not explained by the DNA extraction process or biological sample-to-sample variation but rather the causing factor lies in the library preparation or sequencing processes.

Technically, the only potential explanatory differences between the successful and artefactual samples were paired-end run mode and library preparation on another day, which included EDTA addition. Since paired-end sequencing uses the same chemistry as single-end sequencing, it should not be the explanatory factor. Addition of EDTA could be a logical reason why all these control experiment samples were clean from the artefactual GC>TA variants, however, it does not explain why the earlier successful samples showed equally good variant profiles without the presence of EDTA. It seems that the artefact could be explained by the batch of reagents used or the laboratory personnel preparing the library, or some other random variable not considered here could equally likely be the artefact contributing factor.

Similar to artefactual GC>TA variants, between-library cross-contamination was reduced back to tolerable levels in the control experiment: only 1–4 out of 10 known high-frequency variants from the mutated sample were detected as low-frequency variants in WT mtDNA-seq samples, if the data was analyzed using single index de-multiplexed R1 (**Table 4.2**). Furthermore, one of the samples was exactly the same DNA aliquot that had shown >100 likely cross-contaminating variants (of which one known cross-contamination was

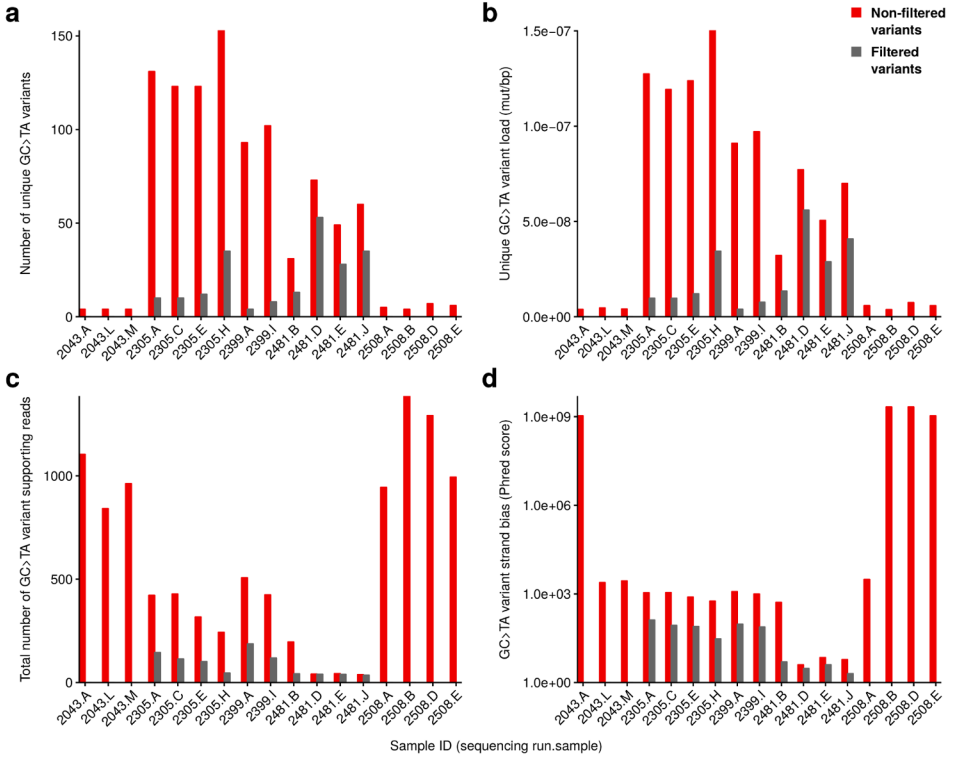

a
Number of unique GC>TA variants
150
100
50
0
b
Unique GC>TA variant load (mut/bp)
1.5e−07
1.0e−07
5.0e−08
0.0e+00
Non-filtered variants
Filtered variants
c
Total number of GC>TA variant supporting reads
1000
500
0
d
GC>TA variant strand bias (Phred score)
1.0e+09
1.0e+06
1.0e+03
1.0e+00
2043.A
2043.L
2043.M
2305.A
2305.C
2305.E
2305.H
2399.A
2399.I
2481.B
2481.D
2481.E
2481.J
2508.A
2508.B
2508.D
2508.E
Sample ID (sequencing run.sample)

Figure 4.8. Comparison of GC>TA variant profiles between successful and artefactual samples. *GC>TA variant profiles of different WT or WT-like (i.e. biologically the sample should contain no GC>TA variants) liver mtDNA-seq samples were compared in order to understand the nature of the artefact.* ***a)*** *The number of observed unique GC>TA variants,* ***b)*** *the unique GC>TA variant load (mut/bp),* ***c)*** *the total number of GC>TA variant supporting reads, and* ***d)*** *the GC>TA variant strand bias Phred score results were obtained from non-filtered (red) and filtered (grey) variant calling results and compared between the successful (Sample ID 2043.x and 2508.x) and artefactual (Sample ID 2305.x, 2399.x, and 2481.x) samples. The control experiment (Sample IDs 2508.x) aimed to test for the effect of library preparation variables on the number of artefactual GC>TA variants included, whereas all the other samples were regular mtDNA-seq samples prepared within these thesis projects or by Johanna Kauppila (Sample IDs 2305.x and 2481.x). The samples are presented in the order of sequencing over a ~10-month period of time. Furthermore, Sample IDs 2399.I and 2508.B were the exactly same DNA aliquot used for the two different library preparations and sequencing runs. All samples were analyzed as if they were single-end sequencing samples i.e. for Sample IDs 2508.x only R1 de-multiplexed by a single index was analyzed. Together these data show that the artefactual samples had high numbers of unique GC>TA variants already in non-filtered variant results (****a*** *and* ***b****), whereas the total number of GC>TA variant supporting reads (****c****) and GC>TA variant strand bias Phred score (****d****) were lower in the artefactual samples than in the successful samples. These different characteristics of the GC>TA variants explain why the variant filtering step is effective for successful samples but not for artefactual samples. Sequencing depths as well as median variant allele frequencies were at similar levels for all samples and thus do not explain the observed differences.*

Table 4.2 Effect of single or dual de-multiplexing strategy on artefactual variant detection. *In the control experiment, three WT samples (2508.A, D and E) were sequenced together with a sample containing ten known high-frequency variants (known, potential cross-contamination source) in paired-end mode allowing read de-multiplexing based on one (single) or two (dual) indices and comparison of the variant results obtained by read 1 (R1) or read 2 (R2) alone (single-end analysis) or in combination (R1 + R2, paired-end analysis).*

Read	**R1**	**R1**	**R2**	**R2**	**R1 + R2**	**R1 + R2**
De-multiplexing strategy	**Single**	**Dual**	**Single**	**Dual**	**Single**	**Dual**
2508.A						
GC>TA variants	0	0	11	10	10	8
Known cross-contamination	1	0	0	0	2	0
Other low-frequency variants	0	0	4	3	3	3
2508.D						
GC>TA variants	0	0	16	12	18	14
Known cross-contamination	4	0	3	0	0	0
Other low-frequency variants	0	0	9	9	12	8
2508.E						
GC>TA variants	0	0	3	2	4	3
Known cross-contamination	1	0	1	0	0	0
Other low-frequency variants	1	1	2	1	0	0

present even at AF 2.6 %) in the first sequencing run. Yet now, with the new library preparation and sequencing run, this sample showed only one clear cross-contamination and three other unexpected variants. Based on a single sample, these results suggest that the unexpected variants were not originally present in the DNA sample, but they were rather cross-contaminations either from library preparation or from the sequencing step. Moreover, the cross-contamination was not detectable, if the dual de-multiplexed R1 was used for the analysis (**Table 4.2**).

Despite the low number of cross-contaminating variants, the data suggest that the dual de-multiplexing strategy is effectively removing between sample cross-contamination. The idea of dual de-multiplexing to increase accuracy has also been supported by others (Kircher et al. 2012, Preston et al. 2016).

To ensure that the change to use the improved mtDNA-seq in the middle of the projects was appropriate, two WT and MKO liver mtDNA samples successfully sequenced with single-end sequencing approach were re-sequenced with the improved dual approach and the variant results were compared (**Fig. 4.9**). The number of detected variants (**Fig. 4.9a,d**) as well as unique (**Fig. 4.9b,e**) and total (**Fig. 4.9c,f**) mutation loads were highly similar between the approaches. Even more precise comparison would include technical replicates from both approaches in order to understand whether the slight variability is simply due to re-sampling the pool of mtDNA molecules and not necessarily by the different approach itself. Either way, the results clearly show that the change of the sequencing approach does not drastically affect the final variant results.

In conclusion, the mtDNA-seq method was modified such that 1 mM EDTA was added to the sonication step, sequencing run was in paired-end mode, reads were dual de-multiplexed and only R1 used for the data analysis. The dual de-multiplexing approach is expected to slightly improve the accuracy of the mtDNA-seq method in comparison to single-end run mode, since even the tolerably low level of cross-contaminating variants are removed. On the downside, sequencing costs are doubled since now two Gbases of sequences are required for equally high mtDNA coverage as obtained by the single-end approach. However, one advantage of the paired-end sequencing is the potential multiuse of the data: in addition to low-frequency variant analysis with R1, both R1 and R2 together can be utilized in e.g. breakpoint or deletion detection.

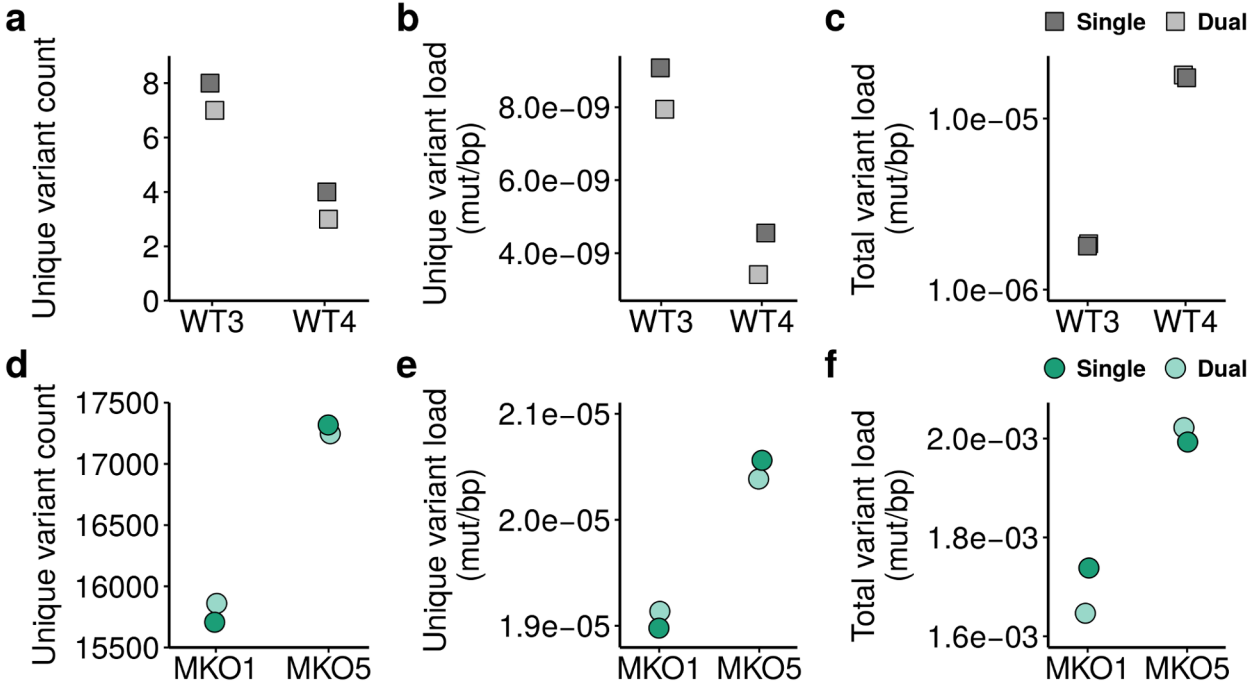

a
Unique variant count
8
6
4
2
0
WT3
WT4
b
Unique variant load
(mut/bp)
8.0e−09
6.0e−09
4.0e−09
WT3
WT4
c
Single
Dual
Total variant load
(mut/bp)
1.0e−05
1.0e−06
WT3
WT4
d
Unique variant count
17500
17000
16500
16000
15500
MKO1
MKO5
e
Unique variant load
(mut/bp)
2.1e−05
2.0e−05
1.9e−05
MKO1
MKO5
f
Single
Dual
Total variant load
(mut/bp)
2.0e−03
1.8e−03
1.6e−03
MKO1
MKO5

Figure 4.9. Effect of single or dual de-multiplexing strategy on variant detection. *Two WT (grey) and MKO (green) mtDNA-seq samples were sequenced and analyzed by the original mtDNA-seq method (single) as well as by the improved mtDNA-seq method including paired-end sequencing with dual de-multiplexing strategy (dual). The variant results obtained by the two methods were compared by* ***a,d****) unique variant count,* ***b,e)*** *unique mutation load (mut/bp), and* ***c,f)*** *total mutation load (mut/bp). By the dual method, on average 4.2 % less variants were detected in MKO samples than by the single method. The average coverages of the total obtained data sets were systematically higher for single than dual data sets (~60000x and ~50000x), thus, to guarantee the most accurate comparison between the two approaches, unique alignment files (.bam) from the single data sets were randomly downsampled to similar coverage levels (~52000x) as the corresponding dual data sets (with command samtools view, parameter -s 0.897, 0.85, 0.8 and 0.86 for WT3, WT4, MKO1 and MKO5, respectively) before the analysis without junction fix for the single data sets.*

4.2.3 Validation of the method for low-frequency mtDNA variant detection

The improved mtDNA-seq approach was validated by sequencing spike-in samples: the background sample was a plasmid containing full mtDNA genome (pAM1, harboring five known variants) and the spike-in additions were titration of a known concentration of mtDNA from NZB mouse strain harboring 88 variants in comparison to the reference sequence (list of variants is given in **Appendix 1**). It was preferable to use pAM1 as a background, because any unexpected low-frequency variant could be considered as an error, whereas if WT mtDNA was used as a background, the possibility of naturally occurring heteroplasmies could not be ruled out complicating the error rate determination. However, it is important to keep in mind that the use of plasmid background is confoundingly different from mtDNA-seq samples because majority of the sample DNA was then prepared by miniprep and not by mtDNA-seq protocol.

In total, two spike-in samples representing NZB mtDNA variants at AF 0.05 % and single spike-in samples representing AF 0.1, 0.2 and 0.5 % were sequenced and the reads were subsampled to represent different coverage levels (10000–60000x). The accuracy of the method was measured by counting true positive (total 88 NZB variant positions) and negative variants (total 16206, the rest of the mtDNA genome positions), and false positive and negative variants as unique variant counts. Precision (positivie predictive value, PPV) and sensitivity (or recall, true positive rate, TPR) were calculated (**Chapter 3.10.2**) and plotted against the variable coverage levels (**Fig 4.10a**). The comparison showed that variant detection is already efficient at 40000x coverage. However, in the spike-in sample representing AF 0.05 % only ~25 % of the NZB mtDNA variants were detected even at 60000x coverage. Furthermore, slightly more false positive variants were detected at higher coverage levels as the precision started to decrease. A relatively low total number of false positive variants were detected at 60000x coverage level. Over all of the spike-in samples, median of five false

positive variants were observed per sample, and all but one were GC>TA variants (**Table 4.3**).

Since higher coverage increased the sensitivity without markedly decreasing precision, samples with 60000x coverage were used as representative control samples for the experiments (**Fig. 4.10b, Table 4.3**). Also, this was the coverage level obtained for mtDNA-seq liver samples. In addition to precision and sensitivity, the total variant read counts were utilized to calculate total true and false positive loads (**Fig. 4.10b**). The true positive variant loads were as expected for all spike-in samples except for the lowest AF 0.05 %. The false positive variant load was low (median $1.37x10^{-7}$ mut/bp), and varied sample-by-sample, mostly indicating a variable number of artefactual GC>TA variants. Presence of the artefact could be due to use of pAM1 as background DNA, yet, to definitely address this, WT mtDNA-seq samples could be sequenced along with the spike-in samples in the future experiments.

To roughly compare the spike-in samples to mtDNA-seq samples, two WT mtDNA-seq samples re-sequenced earlier with the paired-end approach (**Fig. 4.9**) were utilized by considering the variants detected by both, single-end and paired-end approaches as true positive results and variants detected by a single method only as false positive results. Single and dual approach showed average PPV values 0.68 and 0.86, respectively, whereas false positive variant loads were $9.20x10^{-8}$ and $6.18x10^{-8}$ mut/bp (**Table 4.3**). The artefactual GC>TA variants were not present in the dual mtDNA-seq sequencing run, thus expectedly, these WT samples showed slightly lower false positive variant load and similar precision as NZB spike-in samples.

These results show that the data analysis optimized within these thesis projects leads to highly accurate and relatively sensitive variant detection. Furthermore, if the presence of low levels of artefactual GC>TA were indeed caused by the pAM1 plasmid preparation, it is possible that mtDNA-seq samples actually show even lower false positive loads, similar as estimated with the single-end and paired-end re-sequenced WT samples (**Table 4.3**). At AF 0.05 % (simulated

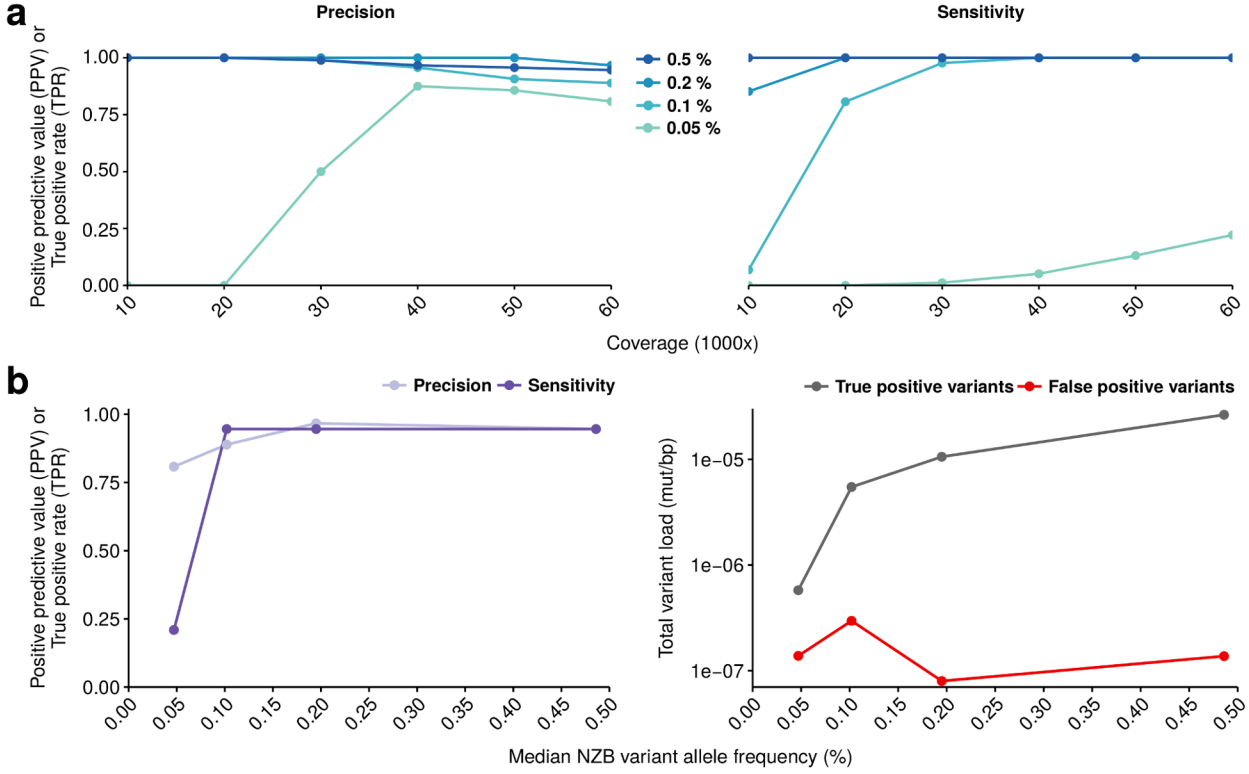

a
Precision
Sensitivity
Positive predictive value (PPV) or
True positive rate (TPR)
1.00
0.75
0.50
0.25
0.00
10
20
30
40
50
60
0.5 %
0.2 %
0.1 %
0.05 %
Coverage (1000x)
b
Precision
Sensitivity
True positive variants
False positive variants
Total variant load (mut/bp)
1e−05
1e−06
1e−07
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
Median NZB variant allele frequency (%)

Figure 4.10. Variant detection accuracy determined by spike-in samples. *NZB mouse mtDNA was used as spike-in DNA in pAM1 plasmid background to determine the variant detection accuracy. Median NZB variant allele frequency (AF) varied from 0.05 % to 0.5 % in the tested samples, results for AF 0.05 % are average of two replicate samples whereas results for other samples are from sequencing of single samples.* ***a)*** *Comparison of precision (PPV) and sensitivity (recall, true positive rate, TPR) over series of coverage subsets (10000x to 60000x) from spike-in samples representing AF levels 0.05–0.5 %. Most variants were detected already at 40000x, however, AF 0.05 % samples detected only ~25 % of the NZB variants even at ~60000x. Slightly more false positive variants were observed at higher coverage levels.* ***b)*** *Coverage level ~60000x was chosen as the representative spike-in sample for precision and sensitivity of mtDNA-seq. Total true positive variant loads were at expected levels, only AF 0.05 % samples were underestimated. For all AF samples, the total false positive variant loads were extremely low. PPV = true positive variants / all detected variants, TPR = true positive variants / all expected variants.*

Table 4.3. Variant detection accuracy. *NZB mouse mtDNA was used as spike-in sample in pAM1 plasmid background to determine the variant detection accuracy. Median NZB variant allele frequency (AF) varied from 0.05 % to 0.5 % in the tested samples. Unique true and false positive (TP and FP, respectively) variant counts were used to calculate positive predictive value (PPV). False positive variants showed mainly the oxidative damage artefact signature (FP GC>TA). True positive rate takes into account the number of false negative variants (non-detected variants out of total 88 NZB variants, TPR) and F1 score is the harmonic mean of PPV and TPR (F1 score). Total TP or FP loads were calculated as the total number of variant reads per the total coverage.*

Sample	AF (%)	TP	FP	FP GC>TA	PPV	TPR	F1 score	Total TP load (mut/bp)	Total FP load (mut/bp)
NZB $5x10^{-4}$ 1	0.05	16	7	6	0.70	0.17	0.29	$4.70x10^{-7}$	$2.17x10^{-7}$
NZB $5x10^{-4}$ 2	0.05	23	2	2	0.92	0.26	0.41	$6.85x10^{-7}$	$5.90x10^{-8}$
NZB $1x10^{-3}$	0.10	88	11	11	0.89	1	0.94	$5.47x10^{-6}$	$2.96x10^{-7}$
NZB $2x10^{-3}$	0.20	88	3	3	0.97	1	0.98	$1.06x10^{-5}$	$7.99x10^{-8}$
NZB $5x10^{-3}$	0.49	88	5	5	0.95	1	0.97	$2.64x10^{-5}$	$1.37x10^{-7}$
WT3 mtDNA-seq liver, **single**	n.a.	5	3	0	0.63	n.a.	n.a.	$1.66x10^{-6}$	$1.40x10^{-7}$
WT3 mtDNA-seq liver, **dual**	n.a.	5	2	0	0.71	n.a.	n.a.	$1.73x10^{-6}$	$1.24x10^{-7}$
WT4 mtDNA-seq liver, **single**	n.a.	3	1	0	0.75	n.a.	n.a.	$1.73x10^{-5}$	$4.43x10^{-8}$
WT4 mtDNA-seq liver, **dual**	n.a.	3	0	0	1	n.a.	n.a.	$1.80x10^{-5}$	0

n.a. = not annotated

LoFreq* detection threshold), only 25 % of the expected variants were detected, which may well be due to difficulties in accurate pipetting (median AF values were 0.047 and 0.0472 %) and thus easily leaving some of the variants just below the LoFreq* detection threhold. To investigate this possibility, also completely non-filtered results were compared. Only two more true positive variants were observed but also 13 additional false positive variants. This suggested that the spike-in

variant frequencies truly were below the detection threshold of LoFreq*. Moreover, these results show that the applied filtering steps indeed are effective in removing false positive variants but not at the cost of sensitivity as hard-coded AF threshold would be. However, if absolute precision would be required rather than higher sensitivity, then the minimum AF threshold should be set at >0.07–0.1 %, as the observed maximum false positive variant AF was 0.07 %.

4.2.4 Discussion

To summarize, mtDNA-seq approach is a cheap, fast and sensitive method to study low-frequency mtDNA mutations over the entire mtDNA genome as an alternative to more labor-intensive or expensive methods (e.g. PCS, UMI methods or circle sequencing). A major drawback is that the extremely rare variant detection may be highly sensitive to artefacts and biases. A clear example of that was the observed artefactual GC>TA variants (**Fig. 4.8**) and between-sample cross-contamination (**Table 4.2**). Since both of these artefacts appeared and disappeared simultaneously, it is tempting to suggest that they would have a common origin. Unfortunately, how such issues suddenly arose, still remains to be clarified.

Intriguingly, it seems that artefactual GC>TA variants showing samples, despite the mtDNA enrichment method, occurred at a certain time frame, suggesting it to be a batch-dependent problem. Ever since the control experiment, the artefact has mainly disappeared. However, all samples have been also sequenced by the improved approach. Thus, to finally conclude whether paired-end run mode or EDTA addition would have an effect, or whether the artefact was present only for a period of time, mtDNA-seq samples should be sequenced again by both run modes simultaneously. The appearance and disappearance of the artefact emphasized that extremely rare variant detection may be susceptible to the slightest changes in the sample preparation or sequencing protocols. Therefore, it is crucial to follow good laboratory practices and, more importantly, good controls with known variant profiles should be included into the sample set for each sequencing run. Cautious data

analysis has to be always applied keeping in mind that unexpected artefacts may arise at any point even with well-established protocols.

Amplicon sequencing of WT sample further highlighted that instead of simply determining the total variant load, it is also essential to investigate the variant profile in detail. Even if the WT sample showed a low total variant load, the number of unique variants was many and they occurred at very low AF levels suggesting them to be artefactual variants. A similarly low total variant load could be obtained by observing only a few higher frequency variants, which more likely represent true heteroplasmic variations than sequencing artefacts. Nevertheless, simply verifying the expected variant load is not enough to conclude successful results.

Sequencing data analysis approaches as well as chosen data analysis tool may have a huge impact on the final variant calling results. For example, variant callers perform very differently and results may even poorly ovelap for the same data set (Pabinger et al. 2013). Chosen data analysis strategy may also have huge impact on the results – e.g. which reference genome to use, usage of single-end or paired-end reads, de-duplication or variant calling thresholds. Zhang et al. (2016), for example, evaluated from exome sequencing and RNA-seq data, how the alignment strategy affects the mtDNA variant detection (Zhang et al. 2016). Here, instead of extensive benchmarking of each data analysis step with different tools, a single tool for each step was carefully chosen and parameters optimized to obtain satisfactory variant calling results. Without a comparison it is impossible to conclude that the chosen strategy is the most optimal combination and also more recently developed tools could even perform better. Nevertheless, the spike-in samples showed very good precision of the applied protocol and no further optimization or updates were considered necessary.

Sequencing errors are often used as an umbrella term for false positive variants, yet clarification of the error source could lead to better corrections. Sequencing with dual indices effectively removed

artefactual errors arising during the sequencing process. Furthermore, a good variant caller, like LoFreq*, was well able to distinguish between an actual sequencing error or a biological variant – thereby efficiently removing artefacts. What was left, were the artefactual variants chemically present in the sample (e.g. mispairing of an oxidative damage adduct during PCR), and thus, such variants are very difficult to exclude by any standard data analysis steps and requires more complex prediction (like in Costello et al. [2013]). Instead of developing more and more ”correction algortihms”, it would be very important to understand at which processing step and why the bias occurs and take all possible measures to minimize that.

Schmitt et al. (2012) investigated the extent of GC>TA variants by Duplex Sequencing, whereas Diegoli et al. (2012) and Lou et al. (2013) successfully included repair enzymes to the sample preparation to reduce artefactual variant results. Furthermore, Costello et al. (2013) investigated the potential source of the damage, and recently Chen et al. (2017) showed how the damage is confounding the somatic variant identification in cancer studies – even up to AF 5 %, thus passing even most of the stringent variant calling thresholds. Although Chen et al. (2017) do have a conflict of interest, with the building evidence and also observations made during this thesis, it is tempting to suggest that a repair step, or other measures to prevent the propagation of the oxidative damage artefacts, should be included into the standard sequencing library preparation – at least for low-frequency variant detection studies, and until the cause can be identified and eliminated. Here, the final mtDNA-seq protocol included EDTA additon as suggested by Costello et al. (2013), however, only additional, well-controlled experiment would show whether EDTA actually affects the occurrence of GC>TA artefact or not.

4.3 Mitochondrial biology research questions addressed by mtDNA-seq

The mtDNA-seq and amplicon sequencing approaches were applied to address various questions in the field of mitochondrial biology. First, the variant profile of the entire mtDNA genome was created to reveal regions critical for mtDNA replication. Second, the developmental timing and mechanism of mtDNA purifying selection was clarified. And third, the effect of mtDNA variants on mitochondrial RNA (mtRNA) processing was studied. The final or preliminary results of each project are presented in this chapter with a brief introduction to each topic.

4.3.1 Creation of variant profile of the entire mitochondrial genome and identification of regions essential for replication and replication-associated transcription

The widely accepted model for mtDNA replication is so called strand-displacement model (as reviewed by Gustafsson et al. [2016], **Fig. 4.11a**). Briefly, the origin of replication of heavy-strand (H-strand), OriH, is located at the control region of mtDNA genome and is the initiation site for the entire mtDNA replication. First, mitochondrial RNA polymerase (POLRMT) forms a short RNA fragment, which is used as a primer by POLG for DNA synthesis of the nascent H-strand. Additionally, the mitochondrial replisome requires replicative mitochondrial helicase (TWINKLE) to unwind the double-stranded DNA (dsDNA), and mitochondrial single-stranded DNA binding protein (mtSSB), which bind to and stabilize the displaced parental H-strand.

Once the unidirectional replication passes OriL, this region becomes single-stranded and is able to form a stem-loop structure. POLRMT uses the structure to produce another short RNA primer, which allows another POLG to initiate synthesis of the nascent L-strand (**Fig. 4.11a**). Now, the replication proceeds simultaneously in both directions. To form two daughter mtDNA molecules, the RNA primers need to be

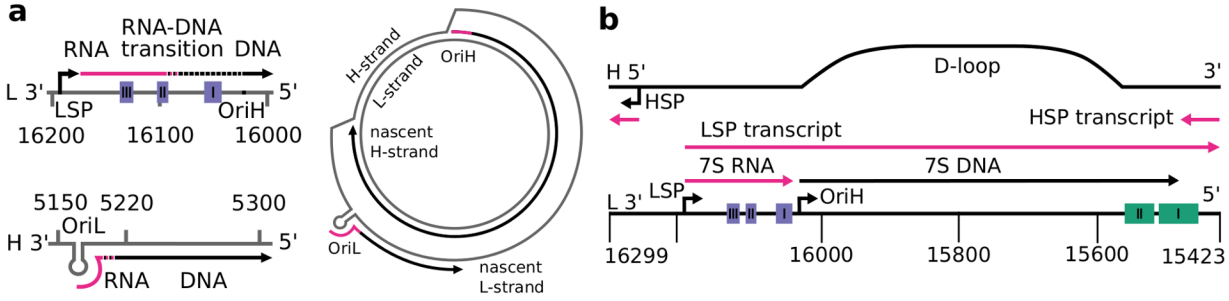


Figure 4.11. Models for mitochondrial DNA replication and transcription. a) *The replication of mtDNA requires a short RNA fragment (pink) formed by mitochondrial RNA polymerase POLRMT upstream of origin of replication of the heavy-strand (OriH, H-strand, RNA-DNA transition site around conserved sequence blocks [CSBs, purple blocks]) to prime the mtDNA synthesis by POLG and other mtDNA replisome components; the replicative mitochondrial helicase TWINKLE and mitochondrial single-stranded DNA binding protein mtSSB. When the replication has proceeded approximately two thirds of the genome, the origin of replication of the light-strand (OriL, L-strand) becomes single-stranded and forms a stem-loop structure. POLRMT synthesizes another RNA primer and the mtDNA synthesis by POLG continues in both directions until the strands are complete.* ***b)*** *The transcription of mtDNA initiates at promoter regions, one for H-strand (HSP) and one for L-strand (LSP). The transcription complex consists of at least POLRMT and mitochondrial transcription factors A (TFAM) and B2 (TFB2M). Transcription from LSP is terminated at mt-tRNA L1, whereas HSP transcription terminates at the extended termination associated sequences (ETASs, green blocks). In addition to full replication or transcription processes, also pre-maturely terminated products, 7S RNA and 7S DNA are formed in significant amounts, however, their functions are not yet fully understood. The processes are reviewed by Gustafsson et al. (2016), and the illustrations are based on the same paper.*

removed, and this RNA-DNA transition site of the H-strand is mapped at the conserved sequence blocks (CSB I–III, **Fig. 4.11a**). Furthermore, newly synthesized DNA strands are ligated by co-operation of DNA ligase III and POLG exonuclease activity. In MKO mice, this process is inefficient due to exonuclease-deficient POLG leading to formation of linear, truncated mtDNA molecules (Uhler & Falkenberg 2015).

Transcription of mtDNA is also initiated at the control region (**Fig. 4.11b**), which harbors two transcription promoters: one for the H-strand (HSP) and one for the L-strand (LSP). Transcription is initiated by mitochondrial transcription factor A (TFAM) binding upstream of the transcription initiation site within the promoter region and changing the DNA structure. This allows interaction with POLRMT and further conformational changes enable binding of mitochondrial transcription factor B2 (TFB2M). The transcription complex then proceeds to produce near-genome-length transcripts. Transcription of the L-strand is terminated at mt-tRNA L1 by mitochondrial transcription termination factor 1 (MTERF1, Terzioglu et al. [2013]), whereas the H-strand transcription is terminated at the extended termination associated sequences (ETAS) by yet-unknown mechanisms (Gustafsson et al. 2016).

The replication and transcription of mtDNA are complex events, which require careful regulation at the busy control region in order to avoid collisions of replication or transcription machineries. In addition to the described basic processes, short 7S RNA and 7S DNA products are formed by pre-mature termination of transcription or replication (**Fig. 4.11b**). The function of these products are not known but speculated to facilitate the transcription and replication processes: 7S RNA, for example, is suggested to secure the delicate ligation process of the DNA synthesis by moving the ligation event further away from the trancriptively active regions. On the other hand, 7S DNA is forming a triple-stranded structure – D-loop – at the control region, and this D-loop, among the many other hypotheses (reviewed by Nicholls & Minczuk 2014), may function as a regulator to avoid replication fork

collisions.

A long-standing issue for studies deciphering the mechanisms of mtDNA replication and transcription is that mammalian mitochondrion cannot be transfected (Patananan et al. 2016). Thus, mtDNA mutator mouse model has been used as a saturation mutagenesis model to clarify these key biological events. For example, Wanrooij S. et al. (2012) showed selection against variants at OriL, whereas multiple studies have observed also low variant load at the mtDNA control region (e.g. Trifunovic et al. 2004; Rovio 2006; Ameur et al. 2011). Such observations highlight the essential nature of those regions for mtDNA maintenance. In these studies, however, it was not possible or simply not in the focus of the study to create a variant profile of the entire mtDNA genome. Detailed information of the mtDNA mutational characteristics, especially at the control region, would aid the research focusing on the mechanisms of mtDNA replication and transcription. Here, with the highly sensitive mtDNA-seq, the aim was to generate a detailed variant profile of the entire mtDNA genome and to identify regions essential for replication and replication-associated transcription. The produced data set is considered to be a highly valuable resource for other researchers; the data can be used, for example, as a starting point to form new hypotheses or to narrow down the target sites in a search of novel proteins and protein-binding sites.

Highly sensitive variant detection from high-coverage sequencing data

With the single (and dual) mtDNA-seq approaches, uniform coverage over the entire mtDNA genome was obtained. Moreover, the linear deletions (brain and liver) as well as control region multimers (brain, as in Williams et al. 2010) of MKO mice were detectable in the coverage profiles (**Fig. 4.12**). These coverage patterns further confirmed the validity of the mtDNA-seq, as information of the linear mtDNA molecules would be easily lost by PCR-based methods. Due to higher levels of nDNA contamination in brain mtDNA-seq samples, the median

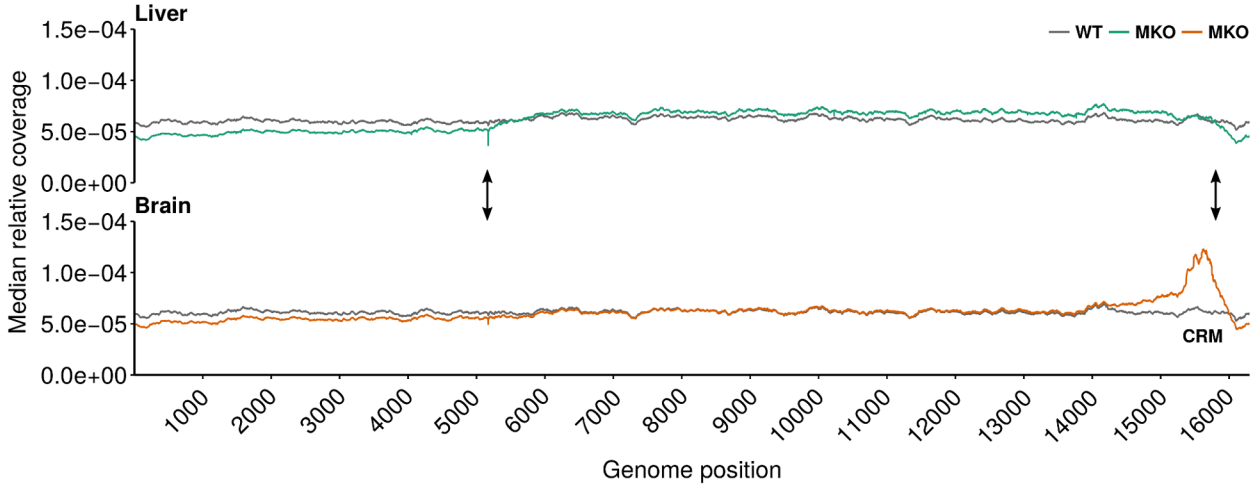


Figure 4.12. Median relative coverage of liver and brain WT and MKO mtDNA-seq samples. *Median relative coverage (coverage per position/total number of bases aligned to mtDNA reference genome) over four and six WT and MKO mtDNA-seq samples, respectively. The linear, truncated mtDNA molecule present in MKO mice is clearly seen in liver coverage profile (green, less notably in brain [orange]) as the relative MKO coverage is higher between positions ~5100–16000 (i.e. between the two origins of replications, indicated by black arrows) than in WT samples (grey). Furthermore, the MKO brain samples show high coverage peak at ~15000–16000, which is caused by the control region multimers (CRM, Williams et al. 2010).*

coverage per position of six brain mtDNA-seq samples was 36000x (min 28000x, max 58000x), whereas results for liver were less variable, median 61000x (min 60000x, max 63000x).

As mentioned in **Chapter 4.2**, mtDNA-seq enabled reliable detection of extremely low-frequency mtDNA variants, yet, despite the extremely low detection threshold, only few variants were observed in WT samples (**Fig. 4.13a–c**), in contrast the majority of the ~17000 variants detected per MKO sample were at AF <0.1 % (**Fig. 4.13d**). On average only 0.8 % (SD = 0.1 %) of the variants were observed at high frequency (AF >0.5 %). Other mtDNA variant studies have often been limited to even higher variant detection thresholds, thus, by mtDNA-seq, it was possible to obtain even ten times better sensitivity and observe significantly more variant results per sample. This is not only leading to more precise variant profile but also ultimately reducing the number of required animals per experiment.

The total variant loads were lower in MKO brain samples than in liver samples (**Fig. 4.13f**). This could reflect the fact that brain (post-mitotic tissue) likely has less on-going mtDNA replication than liver (mitotic tissue). However, brain samples also had higher level of nDNA contamination leading to systematically lower coverages than what was obtained for liver samples and the sequencing depth might easily affect the extremely low-frequency variant detection sensitivity.

To further address the issue of variable coverages, the MKO liver and brain alignment files were subselected (`samtools view -s`) to represent different average coverages per position (range from 10000x to 60000x), and variant calling steps were repeated for these subsets. The comparisons showed that sequencing coverage above 30000x is, especially in liver samples, already reaching plateau in the number of variants detected (**Fig. 4.14a**). Thus, only more of the same mutational events that have independently taken place, or clonal expansion of existing variants are detected with higher sequencing depth as the total variant read count kept increasing with the increasing sequencing depth (**Fig. 4.14b**). Furthermore, a single brain sample reached 58000x

coverage, yet, the variant count was still well below the liver variant counts. Together these data would suggest that the difference between the two tissues, indeed, is biological, but definite conclusions would require sequencing of more high-coverage brain samples.

To further support the hypothesis that mtDNA-seq with ~30000–60000x coverage was sensitively detecting variants at near-saturation level, Venn diagrams of variant positions observed in five of the total six MKO liver and brain samples were compared (**Fig. 4.15a**). Also, variant position results obtained from the liver and brain of a single mouse were compared, pairwise. Venn diagrams showed, that the mutational profiles of the mice were highly homogeneous, showing mostly shared variant positions not only between the different tissues of a single mouse but also between the mice. Moreover, the variant allele frequencies of common variants observed in liver and brain from a single mouse showed high correlation (data not shown). Intriguingly, only a median of 127 and 102 unique variant positions were observed in each independent liver and brain samples, respectively. However, the brain sample showing 58000x coverage (MKO5) harbored 1723 unique variant positions in comparison to other brain samples.

To better visualize the contribution of each mouse in terms of number of detected variant positions, the variant position results were also represented as a cumulative plot (**Fig. 4.15a**). A single mouse liver sample carried variants at ~80 % of all mtDNA genome positions. Further liver samples only slightly increased the number of detected variant positions. With six mouse livers, >90 % of the mtDNA genome positions were observed to be variable. Brain samples showed less saturation than liver samples – <70 % of all mtDNA genome positions carried variants, yet, highly pure MKO5 brain sample had a relatively large impact on the number of observed variant positions, whereas other five brain samples showed more similar behavior.

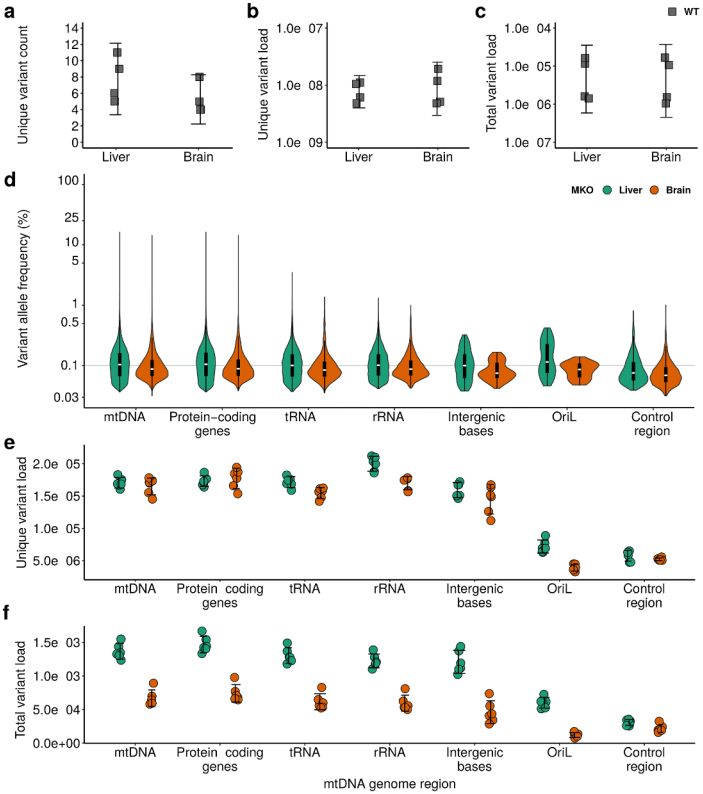


Figure 4.13. Variant loads and frequencies observed in WT and MKO mtDNA-seq samples. *Unique variant counts (**a**) were very low for WT liver and brain samples (grey), thus also unique (**b**) and total (**c**) variant loads were at extremely low levels. Whereas for MKO liver (green) and brain (orange) mtDNA-seq samples, highly sensitive variant detection showed majority of variants at allele frequency (AF) <0.1 % (**d**) (total number of observed mtDNA variants for all liver and brain samples were 102500 and 59622, respectively). Similar trend was observed throughout the genome regions (n in plotting order for liver: 76507, 9122, 14989, 152, 71, 1659, and for brain: 44615, 5119, 8405, 81, 26, 1376). Inside the violin, the white bar corresbonds to the median allele frequency and the black box to the 25th (bottom) and 75th (top) percentile values. Unique (**e**) and total (**f**) variant loads for MKO showed equal distribution of the variants throughout the genome regions – except at OriL and the control region. The total variant loads in MKO brain samples were consistently lower than in liver samples, which could reflect the fact tha brain is a post-mitotic tissue whereas liver is a mitotic tissue. Error bars represent 95 % confidence intervals.*

Taken together, these data suggest that the two organs – representing mitotic and post-mitotic tissues – show highly similar variant profiles. Although, brain samples would slightly benefit from higher coverage, brain mtDNA does harbor fewer variants than liver mtDNA. In conclusion, the results and discussion is focused mainly on liver samples for the sake of clarity. Moreover, it can be concluded that sequencing more mouse samples would not add a considerable value to the results, thus, the variant detection from MKO mice by mtDNA-seq seem to be close to saturation level.

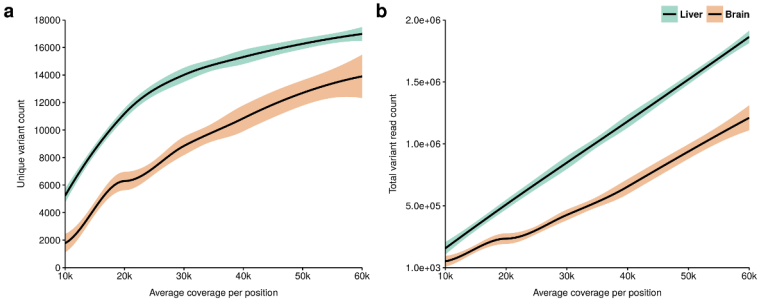


Figure 4.14. Effect of sequencing depth on variant detection sensitivity. *In order to compare the effect of coverage to variant detection sensitivity, sequenced MKO liver and brain mtDNA-seq samples were subselected to different average coverage levels (10000x—60000x, labelled as 10k to 60k) where possible. As seen in* ***a)*** *the unique variant count started to reach plateau already at 30000x, especially in liver samples. Whereas the total variant read count* ***b)*** *showed a linear relationship with the coverage. Together these data suggest, that increasing sequencing depth to over one Gbase (used here, single-end mtDNA-seq), will not significantly increase the number of detected variants. Rather the same mutational event, or clonal expansions are detected. Liver samples had median coverage of 61000x, whereas brain samples were more variable (median coverage 36000x, minimum 28000x and maximum 58000x). Thus, six samples were used for liver (green) subsets, whereas six, three and one sample subsets were available for brain (orange) coverages of 10k–30k, 40k and 50k–60k, respectively. The curves were fitted with method loess and the shaded area represents 95 % confidence interval.*

Non-uniform distribution of variants over the mitochondrial genome regions

In total, from the six MKO brain and liver mtDNA samples, 21713 unique variants were detected at 14898 different genome positions – 91.4 % of the entire mtDNA genome positions were observed to harbor a variant at least once. As shown earlier (**Fig. 4.13e,f**), the variants were distributed equally throughout the mtDNA genome – except at OriL and control regions (**Fig. 4.16**). The extremely low-frequency variant detection enabled the observation of very saturated variant frequency profile over the majority of the mtDNA genome (**Fig. 4.16**, first track). If more stringent variant detection thresholds (AF >=0.5 %) were applied, like in earlier studies, only 229 variants would have been detected (**Fig. 4.16**, second track). Interestingly, some positions were recurrently hypervariable (positions which carried not only a transition but also the two transversions at the same position at least in three liver samples, **Fig. 4.16**, third track), whereas, in addition to OriL and control region, some sites were mutational coldspots (average variant frequency was zero over three or more consecutive positions, **Fig. 4.16**, fourth track).

Previously, Wanrooij S. et al. (2012) showed by PCS of ~1-kb mtDNA region from homozygote mtDNA mutator mice that OriL region is a mutational coldspot. With *in vitro* experiments, they confirmed that variation at OriL region lead to poor mtDNA replication. Thus, they hypothesized that mtDNA molecules harboring variants at regions critical for mtDNA maintenance will not be efficiently replicated. By comparing the data obtained by Wanrooij S. et al. (2012) to this analysis – a comparison of two completely different technologies – strikingly similar variant frequency profiles were observed (**Fig. 4.17a**), confirming that mtDNA-seq is a reliable method. Moreover, these new results – with extremely sensitive variant detection – supported the earlier hypothesis that variants hampering mtDNA maintenance are not expanded to such high-frequency levels that they would be detectable with mtDNA-seq.

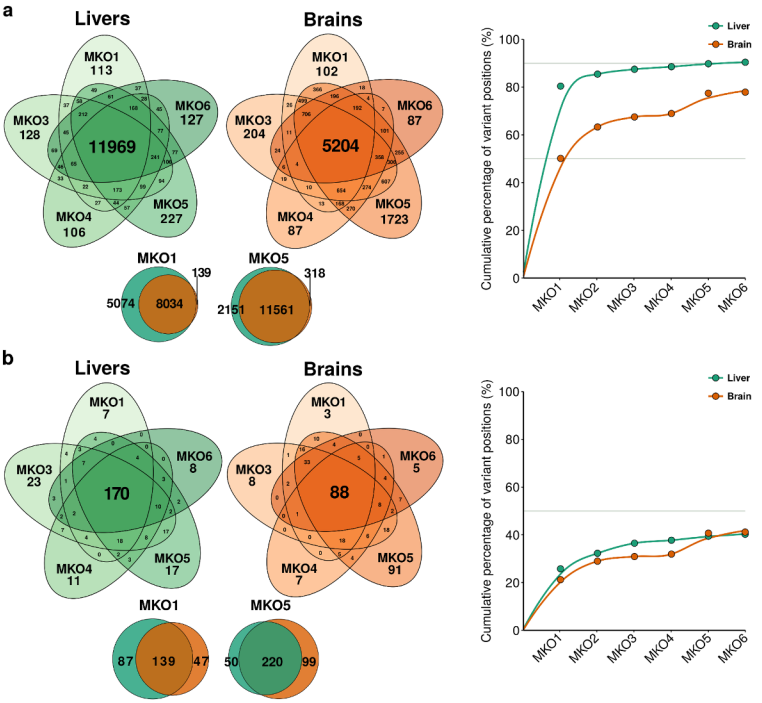


Figure 4.15. Comparison of the variant positions between mice and tissues. *The observed variant profiles between different mice and tissues were compared by Venn diagrams and cumulative plots for* ***a)*** *the entire mtDNA genome and* ***b)*** *separately for the control region. Both tissues showed significant overlap in terms of variant positions over the entire mtDNA genome and over the control region. Each liver and brain sample harbored only ~100 or ~10 unique variant positions over the entire mtDNA genome and the control region, respectively. The total number of mutated positions in liver and brain samples were 14740 and 12697 of the entire mtDNA genome (total 16299 positions), whereas only 353 and 361 mutated positions were observed on the control region (total 877 positions). MKO5 brain sample showed more variant positions due to much less nDNA contamination present in the sample, and thus almost double coverage, in comparison to other brain samples. Pairwise comparisons of the variant positions between the liver and brain from a single mouse showed that majority of the variant positions were common to both tissues. Cumulative plots further illustrated the value added by each sample – most of the variant positions were already observed in a single sample and the curves reached plateau already after 2–3 samples. Over the entire mtDNA genome, >90 % of the genome positions were mutated, whereas only ~40 % of the control region positions harbored a variant. Venn diagrams between mice were constructed using only*

five MKO liver and brain mtDNA-seq samples, because Venn diagram for six samples is not feasible (MKO2 was excluded as it showed highly similar results as most of the other samples). Pairwise comparisons between liver and brain of a single mouse are shown only for two representatives: MKO1 was similar with most of the other samples, whereas MKO5 was the most different one. The cumulative curves were fitted with method loess, and horizontal lines represent 90 % and 50 % levels of the total positions.

Since the OriL region has been already well characterized, here, the focus was on the control region. Although control region has been already earlier noted to be a significant mutational coldspot (e.g. Trifunovic et al. 2004, Rovio 2006, Ameur et al. 2011), none of the studies characterized the variant profile of the control region in detail. In total, six MKO liver and brain samples showed 528 variants at 436 positions – 49.7 % of the control region positions were observed to harbor a variant at least once. Similarly as was observed for the entire mtDNA genome (**Fig. 4.15a**), the mice showed quite homogeneous variant position profiles (**Fig. 4.15b**). However, as is observed in the cumulative plots, a single liver sample contributed to ~89 % of the total variant position results over the entire mtDNA genome, whereas only ~64 % of the variant position results over the control region were observed in a single liver sample. The contribution of the last sample represented only 0.7 percentage points (pp) more variant sites over the entire mtDNA genome, and 2 pp over the control region. Sequencing even more samples was not considered necessary or very cost-efficient.

The control region is known to be highly conserved between rodents, mainly at CSB I–III, at central domain (commonly known as D-loop region) and at ETAS I–II (Sbisà et al. 1997). To compare the evolutionary conservation to the variant frequency profile, mtDNA control region sequences of mouse species (mice), mouse and rat species (mice/rats) and various rodents (rodents) were aligned (**Fig. 4.17b**). The comparison of the variant and invariant sites from the alignment revealed highly evolutionarily conserved regions with mutational coldspots (**Fig. 4.17c**). This observation supported the hypothesis that those regions are highly important for mtDNA replication efficiency.

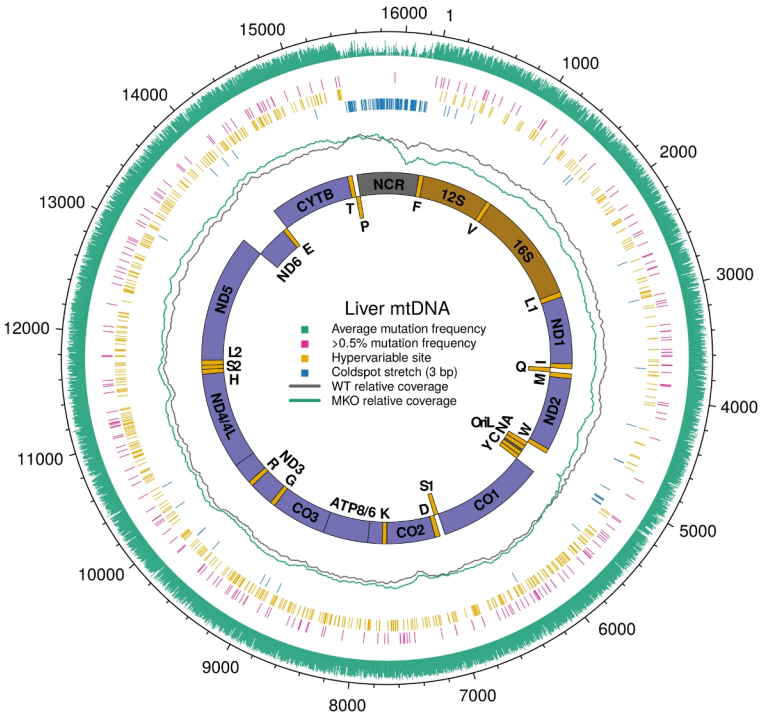


Figure 4.16. Variant profile of the entire mitochondrial genome. *Variant profile over the entire mtDNA genome of the MKO liver samples. Track descriptions from outside to inside: The outer-most, green track shows average mutation frequency per genome position of six MKO mice (y-axis is 1/-log10 of average mutation frequency, 5-bp bins, values ranging from 0 to 0.44). The second, magenta track shows genome positions carrying high-frequency mutations (allele frequency >0.5 %, total 229 positions). The third, yellow track shows recurrent hypervariable genome positions (three different mutations i.e. one transition and two transversions, observed at least in three mice, total 576 positions). The fourth, blue track shows at least 3-bp long coldspot stretches i.e. average mutation frequency was 0 at least for three adjacent genome positions (total 139 stretches). The fifth, line track shows average coverage per genome position for four wild-type mice (WT, grey line) and for six MKO mice (green line, y-axis is 1/-log10 of read depth, 5-bp bins, values ranging from 0.23 to 0.24). The sixth track shows genes encoded on the H- (outward boxes) and L-strands (inward boxes); mRNAs are purple, tRNAs yellow, rRNAs brown and non-coding regions (NCR i.e. control region and OriL) are grey. The plot was constructed with R package circlize (Gu et al. 2014).*

To further characterize the variant profile at the control region, distribution of the variant types were compared. The variant profile was created separately for control region and for the entire mtDNA genome, and the proportion of each variant type, e.g. A>G, was calculated from unique and total loads of that base change in question (i.e. the unique or total variant count of A>G variants were divided by the coverage on As, **Table 4.4**). Thus, the observed proportions were normalized to the mtDNA base composition of the reference strand (L-strand). If no selection and truly random POLG replication errors were assumed, each base would be expected to harbor ~25 % of the observed variants. The assumption was true over the entire mtDNA genome, whereas in the control region the distribution of unique and total variant loads were significantly different ($p = 7.3x10^{-4}$ and $6.6x10^{-4}$, respectively, Chi-squared test). Both A and T bases harbored >30 % of the variants and larger proportions of the variants were either T>C or A>G.

Approximately half of the G>A variants were observed in the control region in comparison to the entire mtDNA (**Table 4.4**). Similarly, the distribution of highly conserved bases (from an alignment of all rodent mtDNA sequences) was different from the base distribution of the control region ($p = 0.046$, Chi-squared test); Gs and Cs were overrepresented, whereas As were underrepresented among the conserved sites. Together these results suggest that the control region GC-content is important for mtDNA maintenance.

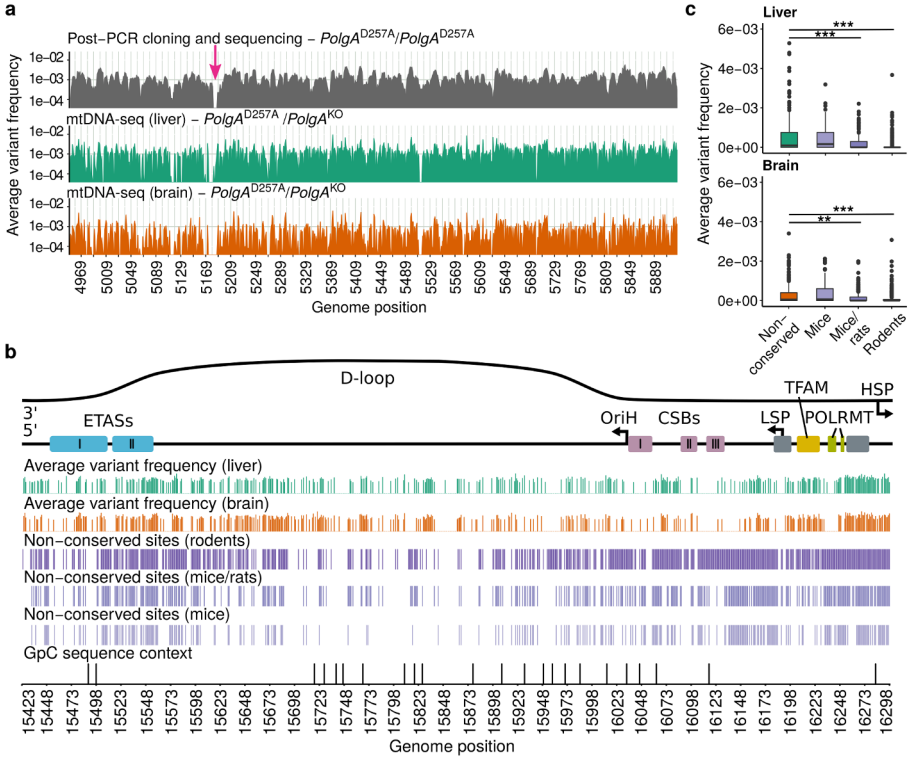

a
Post−PCR cloning and sequencing − PolgA^D257A/PolgA^D257A
mtDNA-seq (liver) − PolgA^D257A /PolgA^KO
mtDNA-seq (brain) − PolgA^D257A/PolgA^KO
Average variant frequency
1e−02
1e−03
1e−04
4969 5009 5049 5089 5129 5169 5209 5249 5289 5329 5369 5409 5449 5489 5529 5569 5609 5649 5689 5729 5769 5809 5849 5889
Genome position
c
Liver
Brain
Average variant frequency
6e−03
4e−03
2e−03
0e+00
***
**
Non−conserved
Mice
Mice/rats
Rodents
b
D-loop
3'
5'
ETASs
OriH
CSBs
LSP
TFAM
POLRMT
HSP
Average variant frequency (liver)
Average variant frequency (brain)
Non−conserved sites (rodents)
Non−conserved sites (mice/rats)
Non−conserved sites (mice)
GpC sequence context
15423 15448 15473 15498 15523 15548 15573 15598 15623 15648 15673 15698 15723 15748 15773 15798 15823 15848 15873 15898 15923 15948 15973 15998 16023 16048 16073 16098 16123 16148 16173 16198 16223 16248 16273 16298
Genome position

Figure 4.17. Variant profile at non-coding regions of the mitochondrial genome. a) *To further confirm the realiability of mtDNA-seq, the observed average variant frequencies from six MKO liver (green) and brain (orange) samples were compared to the results obtained by Wanrooij S. et al. (2012) by PCS of homozygote mtDNA mutator mouse tissues (grey, 5-bp sliding window). Despite the completely different technologies applied, strikingly similar variant frequency profiles were obtained. The pink arrow points to OriL region, which was further characterized in in vitro studies by Wanrooij S. et al. (2012), revealing that variants at that region affect mtDNA replication efficiency.* ***b)*** *In addition to OriL, the control region was also a significant mutational coldspot. The variant frequency profiles showed long stretches where variants were never observed in any of the six MKO liver or brain samples. Intriguingly, these mutational coldspots aligned with the evolutionarily conserved regions (purple tracks): bars represent non-conserved and gaps conserved sites of control region sequences from 26 representative rodents, which were subset to 21 mouse and rats or 8 mouse species (total number of conserved sites for each track were 336, 494 and 603). Additionally, some functionally annotated sites, like POLRMT binding sites, were also mutational colspots. The last track indicates all potential methylation sites i.e. GpC sequence context (n = 23), which can be utilized to potentially determine novel protein-binding sites. ETASs = extended termination associated sequences, OriH = origin of replication of H-strand, CSBs = conserved sequence blocks, LSP = L-strand transcription promoter, TFAM = mitochondrial transcription factor A, POLRMT = mitochondrial RNA polymerase, HSP = H-strand transcription promoter.* ***c)*** *The average variant frequency of liver or brain samples was grouped according to the level of conservation of a site: non-conserved sites were the ones harboring colored bar in all three classes in* ***b)*** *(n = 274), whereas the other groups represent specific gap sites of each class in* ***b)****, i.e. conserved sites only in rodents (n = 336), conserved sites in mice/rats which were not already in rodents (n = 158) and those sites showing conservation only in mice (n = 109). Significantly lower average variant frequency was observed in both liver and brain samples at the most conserved sites i.e. gap regions in rodents or mice/rats in* ***b)****. ** p<0.01, *** p<0.001, post-hoc Tukey Honest Significant Difference of 1-way ANOVA.*

Table 4.4. Comparison of variant distribution over the entire mtDNA genome and control region. *Percentages of each variant type were calculated from unique and total variant loads which were calculated by dividing the unique or total variant read count by the number of aligned bases in question i.e. variant loads were normalized to the base composition of the region. Each row represents the reference base and columns the variant base, total column shows the total distribution of variants on each reference base. Transitions are highlighted with grey background.*

Unique variant load proportions, mtDNA

→	A	C	G	T	Tot.
A	-	2.81	18.17	7.58	28.56
C	0.89	-	0.60	19.45	20.94
G	18.77	1.63	-	1.08	21.48
T	4.00	22.65	2.38	-	29.03

Total variant load proportions, mtDNA

→	A	C	G	T	Tot.
A	-	3.30	15.89	6.71	25.09
C	0.60	-	0.39	22.06	23.05
G	19.64	1.12	-	0.71	21.47
T	3.04	24.71	1.83	-	29.58

Unique variant load proportions, control region

→	A	C	G	T	Tot.
A	-	2.46	27.01	5.64	35.11
C	0.14	-	0.06	17.64	17.84
G	11.59	0.13	-	0.28	12.00
T	4.31	30.28	0.45	-	35.04

Total variant load proportions, control region

→	A	C	G	T	Tot.
A	-	3.25	25.15	4.20	32.60
C	0.11	-	0.04	19.83	19.98
G	10.42	0.08	-	0.14	10.64
T	4.70	31.51	0.59	-	36.80

Discussion

To summarize, mtDNA-seq was successfully applied to produce detailed variant profile of the entire mtDNA genome of MKO mice. The detailed characterization showed, that MKO mice indeed are a true *in vivo* mtDNA saturation mutagenesis model, and for the first time, the variant detection was possible at high sensitivity. Moreover, comparison of the new results to previous PCS results not only further confirmed the reliability of mtDNA-seq, but also extended the previous study by showing that despite the extremely low-frequency detection threshold, the same mutational coldspots were still observed at OriL. Similarly, as noted before with less sensitive methods, also control region harbored major mutational coldspots.

Detailed control region variant profile revealed alignment of highly conserved or functionally important regions to mutational coldspots. This supported the hypothesis that regions essential for mtDNA replication and replication-associated transcription are sensitive to variation as it likely affects the replication efficiency, thus mutated mtDNA molecules do not proliferate to the levels above the variant detection threshold. For example, POLRMT binding site (Posse et al. 2014), although not evolutionarily very conserved except at LSP site, was fully a mutational coldspot supporting that the POLRMT interacts sequence specifically with the DNA at the transcription initiation site (Gaspari et al. 2004; Posse et al. 2014). On the other hand, TFAM binding site seem to be dispensable, although it is suggested to bind sequence specifically to activate promoter-specific transcription (Fisher & Clayton 1988; Fisher et al. 1992; Posse et al. 2014). However, as TFAM is capable of binding mtDNA also in a non-sequence-specific manner, this observation is not that surprising. Probably, specific POLRMT binding without sequence-specific TFAM binding is crucial for replication initiation by providing primer for POLG.

The functionally annotated ETAS sites, of which ETAS I also highly conserved, were also mutational coldspots (also noted by Rovio 2006). Doda et al. (1981) suggested that certain sequence motifs identified near the end points of various 7S DNA molecules would be important signals for termination of mouse mtDNA synthesis. They proposed that the primary sequence or a secondary structure arrests the replication process (Doda et al. 1981). Only two out of the four predicted motifs were also mutational coldspots, whereas the mapped 3' ends of 7S DNA molecules were coldspots. More recently, ETAS I and CSB I regions, both harboring a sequence motif $ATGN_9CAT$ (in human and in mouse mtDNA, partially overlapping with the findings of Doda et al. 1981), were suggested to be important for transcription termination from HSP and LSP, respectively, and the ETAS region also for formation of 7S DNA in humans (Jemt et al. 2015). However, the mouse mtDNA control region harbors in total four of such motifs and only the one on CSB I was a significant mutational coldspot. Thus, it is likely that more precise

or shorter sequence motifs are required, and as suggested also by Jemt et al. (2015), it is likely that other factors than these sequence motifs alone are involved in these key processes. A 48-kDa yet-to-be-isolated protein is known to bind at ETAS region in cows, probably involved in replication termination and formation of D-loop structure (Madsen et al. 1993), however, such protein has not yet been identified in human nor in mice (Gustafsson et al. 2016).

Variant and conservation profiles also seemed to retain the GC-content of the control region. The longest coldspots and conserved regions were on D-loop, which is also more G-rich in comparison to other control region sites (Sbisà et al. 1997; Larizza et al. 2002). It has been suggested, that D-loop region, and especially G-rich motifs, would be important in anchoring the mtDNA to the mitochondrial membrane (Jackson et al. 1996; Larizza et al. 2002). Moreover, although triple-stranded structure formation i.e. D-loop region is well-known, its function still requires clarification: it is suggested to prime mtDNA replication, serve as a dNTP pool or a regulator to avoid replication fork collisions (reviewed by Nicholls & Minczuk 2014).

Another G-rich region is located at CSB II, which was also a mutational coldspot. It is a known site for formation of stable RNA-DNA hybrid, which is suggested to be involved in mtDNA transcription and replication as well as D-loop stabilization via formation of G-quadruplex structures (Xu & Clayton 1996; Wanrooij P.H. et al. 2012). Recent *in vitro* study of human mtDNA suggested that length heterogeneity of the G-tracts changes stability of the G-quadruplex and thus modulates the transcription termination and replication initiation efficiencies (Tan et al. 2016). Furthermore, it was shown in another human and mouse mtDNA *in vitro* study that POLRMT pauses at several sites, especially at G-quadruplex structure, and mitochondrial transcription elongation factor (TEFM) is required to enhance the transcription elongation, thus, TEFM was suggested as a potential regulator of mtDNA replication (Posse et al. 2015).

The above-mentioned cases are just a few examples on how the variant

profile of the entire mtDNA genome may be utilized in further research. The data set produced in this project enables formation of new hypotheses or may represent *in vivo* results to support existing hypotheses, and it may serve as a resource for the search of novel protein-binding sites or other functional elements. In the end, it may provide crucial information of mtDNA replication and replication-associated transcription mechanisms.

4.3.2 Clarification of developmental stage and mechanism of purifying selection of mitochondrial DNA

In mammals, mitochondria are transmitted maternally as, upon fertilization, the sperm mtDNA is eliminated (Shitara et al. 2000, Luo et al. 2013; Pyle et al. 2015). During the early stages of embryogenesis, mtDNA is not replicated, but instead the ~10^5 maternal mtDNA molecules are efficiently diluted during cell divisions (reviewed by Stewart & Larsson 2014) (**Fig. 4.18**, pre-natal bottleneck). Only a tiny subset of the maternal pool of mtDNA molecules is segregated to form the primordial germ cells (PGC), and once the mtDNA replication is re-initiated at the embryonic day ~7.5 (E7.5), mtDNA molecules will be multiplied exponentially. As a result of this mitochondrial bottleneck, and potentially rapid proliferation of some mtDNA molecules but not the others, the proportion of different mtDNA molecules can fluctuate drastically between the cells. However, as reviewed by Stewart & Larsson (2014) there is still no consensus when the mitochondrial bottleneck exactly takes place – some groups argue that the amount of mtDNA is extremely low at early PGCs explaining the bottleneck, others claim the amount of mtDNA to be much higher and that only subset of the mtDNA molecules would be replicated, yet another group suggests the occurrence of another replicative burst after birth to form the mature oocytes (post-natal bottleneck).

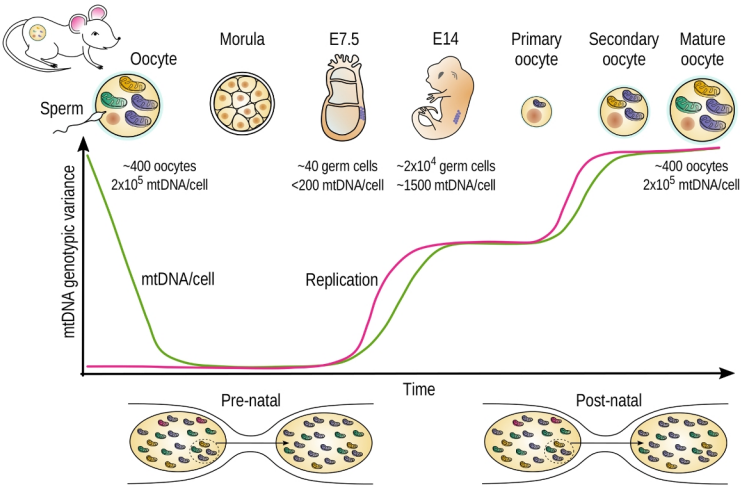


Figure 4.18. Mitochondrial bottleneck. *During embryogenesis, the number of mtDNA molecules per cell (green line) is diluted because very minimal mtDNA replication (pink line) takes place while cells are dividing until embryonic day ~7.5 (E7.5). By E7.5, primordial germ cells are segregated (purple dots) and mtDNA replication is re-initiated. A very small number of mtDNA molecules are extensively replicated (pre-natal bottleneck) in the developing embryo to form oogonia (purple dots). After birth, another replicative burst takes place (post-natal bottleneck) to form the mature oocytes and, through atresia, only small amount of mature oocytes are left. The illustration is based on Wai et al. (2008), Cree et al. (2009), Poulton et al. (2010) and Stewart & Larsson (2014).*

Purifying selection has been shown to affect the mtDNA transmission. Stewart et al. (2008a) utilized the homozygote mtDNA mutator mice as founders for female lineages. By Sanger sequencing of 190 N2 to N6 generation mice, they observed a rapid purifying selection against deleterious variants. Under a neutral model, one would expect to see an equal distribution of the variants at any mtDNA site in early generations. Indeed, transmission of variants on tRNA and rRNA, as well as on third codon position of protein coding sites (i.e. likely neutral variants) followed this expectation. However, the variant loads on the first and

second codon positions (i.e. potentially amino acid changing, deleterious mutations) were significantly lower in comparison to the variant load on the third codon position. Such a variant profile is considered as a hallmark of purifying selection. Their observations suggested that during the formation of subsequent generations the deleterious variants are selectively lost, even if they would be rare and below any functional threshold level (Stewart et al. 2008a). The exact molecular mechanism and developmental stage when this selective pressure occurs are still not understood.

In this project, MKO mice were utilized as founders (F1) for female mouse lineages (N1–N3, **Fig. 4.19**). This breeding scheme allows only variants introduced by the founder mother to be transmitted to the next generations without a background variant load. Furthermore, the previous study utilized insensitive Sanger sequencing, and thus, only the detection of high-frequency variants from N2 generation on was possible. Here, the sensitive mtDNA-seq approach enables, for the first time, efficient detection and following of transmitted variants through the generations, moreover, the variants are also detectable from N1 generation mice. Thus, this project aims to clarify the developmental stage and mechanisms of germ line mtDNA purifying selection.

Mitochondrial bottleneck is the most effective factor in mtDNA transmission

The preliminary samples were obtained from two female lineages (two F1 mothers). Per N1–N3 generation, a mother with a littermate was dissected to obtain in total four liver samples per generation of which N1s were obtained by the mitochondrial extraction kit method (see discussion in **Chapter 4.2.2**) and the other samples with mtDNA-seq (single-end). To increase the power, mice from two more lineages were sequenced with the improved dual mtDNA-seq approach, however, the results presented here, are based only on the preliminary samples as the sequencing of the other samples was delayed due to the previously discussed artefacts (**Chapter 4.2.2**).

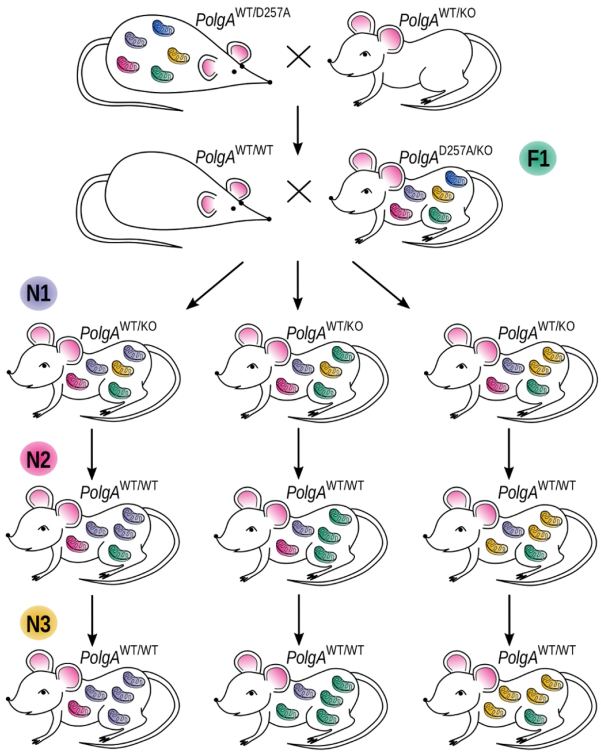


Figure 4.19. Breeding scheme to segregate maternally transmitted mtDNA variants into female mouse lineages. *The breeding scheme illustrates generation of female mouse lineages carrying maternally transmitted mtDNA variants. The founder MKO (F1, PolgA$^{D257A/KO}$) mouse is generated by crossing a heterozygote mtDNA mutator mouse (PolgA$^{WT/D257A}$) male and a hemizygote WT female (PolgA$^{WT/KO}$). Thus, F1 does not inherit mtDNA variants, but is introducing them for the first time. To study the transmission of mtDNA, mtDNA variants can be segregated to female lineages by breeding F1 mice with WT males and selecting for WT (PolgA$^{WT/KO}$) female offspring (N1). These mice inherit subset of the maternal pool of variable mtDNA molecules, but do not introduce more mtDNA variants. Further breedings with WT males produce N2 and N3 generation females (PolgA$^{WT/KO}$ or PolgA$^{WT/WT}$). Each offspring generation is expected to harbor the same total load of maternally transmitted mtDNA variants (illustrated by colored mitochondria) – only the number of different mtDNA variants decreases, but the same variants are observed at higher frequencies, i.e. clonally expanded, in the later generations. The illustration is based on Stewart et al. (2008a).*

The N1 generation mice harbored median of 1836 variants, which was approximately 10x less than what had been observed for MKO mice (**Chapter 4.3.1**, same genotype as founder F1 mice, but the specific F1 mice forming the female lineages used in this project were not sequenced). In comparison to N1 generation mice, approximately 8x and 16x less variants were observed in N2 and N3 generation mice (249 and 112, respectively, **Fig. 4.20a,b**). As expected, the total variant loads were at similar levels in all generations (**Fig. 4.20c**). This was due to clonal expansion of the maternally transmitted mtDNA variant subset, also noted as increasing median allele frequencies in the later generations (**Fig. 4.20d**).

In order to compare the mtDNA-seq data to the previous results obtained by Sanger sequencing (Stewart et al. 2008a), a similar figure revealing the hallmark of purifying selection was plotted. First the varian counts were normalized to the genome element length and then plotted as proportion of all variants in question (**Fig. 4.21**). The figure represents neutral variants i.e. tRNA and rRNA (RNA) and variants on each codon position (CP) of protein-coding regions, which were further separated into two classes according to the variant effect i.e. synonymous or non-synonymous variants. Variants on other regions (intergenic bases, OriL or control region) were excluded from this analysis as non-informative. For comparison, the same plot was also produced from MKO mice presented in **Chapter 4.3.1**.

Expectedly, the variant profile of MKO mice showed uniform distribution of the variants over the different genome sites (**Fig. 4.21**, left-most plot). Similarly, when considering the mtDNA-seq results of N1 to N3 generation mice without any hard-coded AF thresholds, the variants were quite uniformly distributed over the different genome sites (**Fig. 4.21**, middle column of plots). However, there was a mild trend of purifying selection notable by the N3 generation mice: the amount of likely deleterious, amino-acid changing variants (mainly non-synonymous variants on CP1 and CP2) was less than the amount of neutral variants (RNA or synonymous variants on CP3) in comparison

to MKO, N1 or N2 generation mice. These were slightly surprising results, because the previous study showed a strong purifying selection already by N2 generation mice (Stewart et al. 2008a).

A likely explanation for the difference between mtDNA-seq and the earlier results (Stewart et al. 2008a) is that in Sanger sequencing the variant detection threshold is very high (AF ~30 %, Hancock et al. 2005), thus it detects only the highly clonally expanded variants, whereas mtDNA-seq is far more sensitive method. For better comparability to the earlier study (Stewart et al. 2008a), the variant results obtained here by mtDNA-seq were filtered by setting minimum AF to 5 % (**Fig. 4.21**, right-most column of plots). Using any higher AF thresholds was not meaningful, since the number of high-frequency variants was already very low at AF 5 % threshold in N1 generation mice (**Fig. 4.20d**). Now with the minimum AF 5 % threshold, the hallmark of purifying selection became observable even in N1 generation mice and was even more emphasized by N3 generation mice.

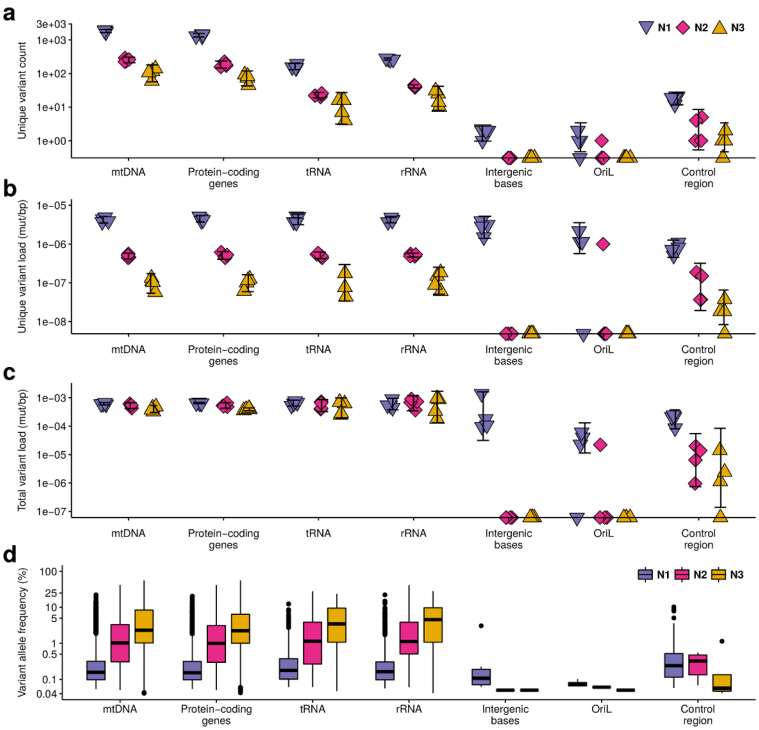


Figure 4.20. Variant loads and frequencies in female lineages carrying maternally transmitted mtDNA variants. *Variant counts (**a**) as well as unique variant loads (**b**) decreased from N1 to N3 generation of mice, whereas total variant load (**c**) was almost at equal level in each mouse generation because the maternally transmitted mtDNA variants were clonally expanded as also noted by increased median allele frequencies from N1 to N3 generation mice (**d**). Samples are from two lineages, two littermates from each generation of both lineages. In **d)**, total number of variants for each generation in plotting order were N1: 7376, 5556, 650, 1086, 7, 4, 73; N2: 1015, 745, 92, 166, 0, 1, 11; and N3: 423, 296, 43, 80, 0, 0, 4.*

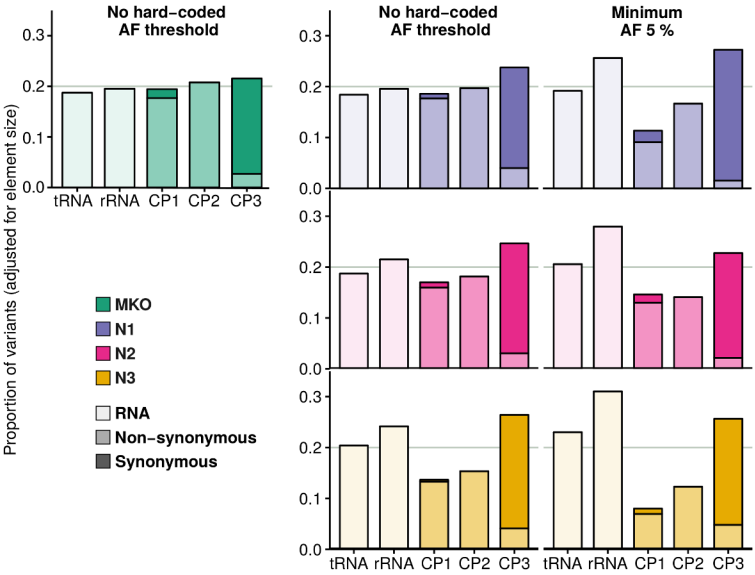


Figure 4.21. Proportion of mutations on different genome elements in different mice generations. *The proportions of each variant type was compared between the N1 to N3 generations and MKO mice (same genotype as female lineage founders F1, however, F1s were not sequenced) was plotted for comparison. The hallmark of purifying selection, i.e. decreased number of variants on first and second codon positions (CP1 and CP2) in comparison to neutral mutations on RNAs or on third codon position (CP3) was mildly visible in N2 and N3 generations (column-wise, germ line) when all mtDNA-seq results were considered (no hard-coded AF threshold). If hard-coded minimum allele frequency (AF) thresholds was set to AF >= 5 %, the decrease in number of deleterious variants in comparison to neutral variants became more visible between generations (column-wise). When the results were considered within the generation (row-wise, somatic), the hallmark of purifying selection was observed strongly already in N1 generation. This suggested that purifying selection takes place during the embryogenesis, whereas random drift is the most effective factor in germ line mtDNA variant transmission. The plot was produced using preliminary data from only two mouse lineages (each generation N1 to N3 n = 4) and the MKO mice data from* ***Chapter 4.3.1*** *(MKO n = 6). The unique variant counts per element were first divided by the element size (tRNA = 1501 bp, rRNA = 2357 bp, CP1 = 3803 bp, CP2 and 3 = 3800 bp) and then proportion of all variants in question were compared. Total number of observed variants were in order tRNA, rRNA and CPs non-synonymous and synonymous: for MKO 9161, 14989, 21914, 2187, 25717, 0, 3357, 23332, for N1 no threshold 650, 1086, 1584, 81, 1764, 0, 359, 1768 and AF 5 % 10, 21, 12, 3, 22, 0, 2, 34, for N2 no threshold 92, 166, 199, 13, 226, 0, 38, 269 and AF 5 % 15, 32, 24, 3, 26, 0, 4, 38, and for N3 no threshold 43, 80, 71, 2, 82, 0, 22, 119 and AF 5 % 17, 36, 13, 2, 23, 0, 9, 39.*

Discussion

One of the major challenges in the field of mitochondrial biology is to understand how exactly the heteroplasmic mtDNA variants are transmitted to the offspring and, in the case of pathogenic mtDNA mutations, how the transmission could be prevented. It is well known, that the mtDNA variant allele frequencies may rapidly shift within just a few generations (Hauswirth & Laipist 1982). Often an asymptomatic mother, not even known to be a carrier of a pathogenic mtDNA mutation, may have an affected child (Kang et al. 2016). Furthermore, the study by Kang et al. (2016) is only one of the many examples, that some pathogenic mtDNA mutations may often be present at very high levels (AF ~70 %), yet the symptoms only occur when the pathogenic mutation is expanded to even much higher levels. This behavior of mtDNA transmission is well explained by the random genetic drift model (Wonnapinij et al. 2008) and the mitochondrial bottleneck. In addition to the mitochondrial bottleneck, Stewart et al. (2008a) showed a strong purifying selection of mtDNA, and also Fan et al. 2008 have shown how severe mtDNA mutation was eliminated within four mice generations whereas less pathogenic mutation was retained. These results indicated that mtDNA purifying selection takes place in the germ line.

Several reviews over the years (Stewart et al. 2008b; Cree et al. 2009; Poulton et al. 2010; Stewart & Larsson 2014; Stewart & Chinnery 2015) have discussed the possible mechanisms for mtDNA transmission and selection, which may take place on molecule, organelle, cellular or organism level. Jenuth et al. (1997) showed that in some tissues mtDNA variants follow random drift model, whereas in others there is selection of mtDNA. They suggested that this may be due to replicative advantages of one mtDNA molecule over another, different turnover rates or selection at a cellular level by altered respiratory chain function (Jenuth et al. 1997). The mtDNA selection in the germ line has been suggested to be an active process, in which defective components are identified and removed or it can be a competition, in which the fittest

component will contribute the most to the next generation (Stewart et al. 2008b, Poulton et al. 2010).

Here, these preliminary results were in line with the previous study (Stewart et al. 2008a) showing a strong purifying selection in the N3 generation mice, however, similar variant profile was only mildly present in N2 generation mice (**Fig. 4.21**). If mtDNA-seq results without hard-coded AF threshold filtering were considered, the variant profile resembled more neutral model, especially for N1 generation mice. Intriguingly, when only high-frequency variants (AF >=5 %) were considered, the hallmark of purifying selection became visible also in N1 generation mice. These new, more sensitive results suggest that any variant may be transmitted from the mother to the offspring, but only the less deleterious variants are able to clonally expand during the development. Thus, it can be hypothesized that the mitochondrial bottleneck and random drift are the most effective factors considering the germ line mtDNA transmission.

The purifying selection seems to take place during the embryogenesis, observable already at very low AF 5 %. The homozygote mtDNA mutator mouse has been found to show a mosaic respiratory chain deficiency (Trifunovic et al. 2004) i.e. only some cells are highly defected harboring high levels of the deleterious mtDNA mutation whereas other cells are normal. Similar mosaicism can be assumed to be present also in these MKO descendants, thus individual cell may show a high variant load but the total variant load of the tissue is low. The selection could then occur on the cellular or organellar level such that proliferation of a highly defected cell/organelle is hampered. On cellular level, defected cells could be eliminated by apoptosis, whereas on organellar level, defects could lead to changes in mitochondrial membrane potential, which in turn would affect the protein import into mitochondria and eventually the mtDNA replication efficiency (as reviewed by Stewart & Chinnery 2015).

4.3.3 Effects of mitochondrial DNA variants on mitochondrial RNA processing

Mitochondrial RNA processing has peculiar characteristics in comparison to nuclear RNA. As already briefly mentioned (**Fig. 4.11b**), mtDNA is transcribed into two near-genome-length transcripts – one for the H-strand and another for the L-strand. Interestingly, most of the mRNAs are flanked by tRNAs (**Fig. 1.1**), and according to so called tRNA punctuation model (Ojala et al. 1981), precursor mtRNA transcripts are spliced into mRNAs, tRNAs and rRNAs in order to form the mature RNA products. Two main protein complexes have been identified, RNase P and RNase Z, which are responsible of the 5' end and 3' end endonucleolytic cleavage of the polycistronic RNAs, respectively. However, several sites (e.g. 5' end of COI), do not fit into the tRNA punctuation model and also nearby sequence contexts do not seem to function as recognition signals for the cleavage enzymes. Furthermore, all the other involved proteins or exact mechanism details how mature mtRNA products are formed, are still not known (as reviewed by van Haute et al. 2015).

Recently, Stewart et al. (2015) studied human tumor sequencing data and compared the relative variant levels detected in mtDNA and mtRNA (Stewart et al. 2015). Most of the variants occurred, expectedly, at similar allele frequencies in both mtDNA and mtRNA. However, handful of variants showed a clear allelic mismatch. Surprisingly, these outlier variants were mainly on tRNAs. Such observation was intriguing since in normal RNA-seq the poly-A enrichment step or size selection should eliminate the presence of small tRNAs in the sequenced RNA pool and detection of tRNA variants is not expected. Those tRNA regions also showed higher coverage pattern in the variant carrying samples than in other samples. This lead to a hypothesis that a variant is potentially disrupting the secondary structure of the tRNA, which would hamper the mtRNA processing and yield to mRNA products, which are still attached to the flanking tRNA sequences (**Fig. 4.22**).

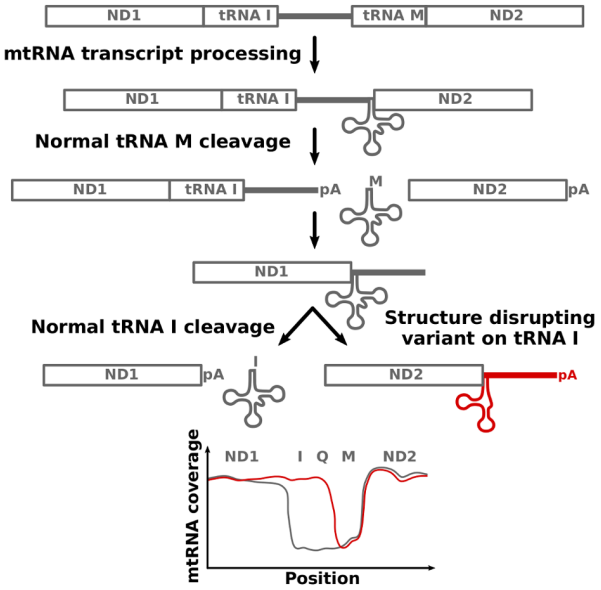


Figure 4.22. Model for mitochondrial RNA precursor transcript processing. *Most of the mitochondrial mRNAs are flanked with tRNAs and near-genome-length precursor transcripts are processed by splicing out the flanking tRNAs to form mature, poly-adenylated (pA) mitochondrial RNA (mtRNA) products. Analysis of variants from human tumor sequencing data of mtDNA and mtRNA (Stewart et al. 2015) revealed mismatches in the allele frequencies; several tRNA variants were observed at high-frequency levels although the variant frequency was much less in the corresponding mtDNA. In poly-A-enriched RNA pool, it was not expected to detect short tRNAs at all, yet, the tRNA variant carrying samples (red) showed high coverage on the tRNA in comparison to non-variant samples (grey). The hypothesis was that a certain variant on tRNA disrupts the tRNA secondary structure in comparison to normal tRNA. Altered tRNA structure hampers the cleavage process leading to accumulation of precursor transcripts, which is detected as high coverage on the variant tRNA region.*

The aim of this project is to continue the human tumor study with the MKO mouse model, utilizing the N1–N3 mouse generations (introduced in **Chapter 4.3.2**), which would harbor high-frequency, clonally expanded variants. Study on mouse samples allows more detailed molecular characterization of the hypothesized mtRNA processing defects. The research idea was tested in a preliminary experiment in which mtDNA from N1 and N2 generation mice were sequenced by amplicon sequencing with non-tagged primers and mtRNA by directional total RNA-seq. Amplicon sequencing is preferred method for this kind of project, since that method is significantly faster regarding the hands-on time and the expected variants are of high-frequency (AF >>0.5 %), thus, reliable detection is straightforward from the amplicon sequencing data. Furthermore, PCR-errors do not pose a great risk of false-positive results since any variant of interest is expected to be observed in both mtDNA and mtRNA. Although, this already eliminates the risk of observing variants originating from the amplicon primers, the use of the tagged primers for future experiments (as discussed in **Chapter 4.2.2**) is still recommended in order to eliminate any unnecessary bias.

With high RNA-seq depth (five Gbases), it was indeed possible to detect ~10–1000x coverage on tRNAs. The final variant results were filtered to contain only those variants shared by the mtDNA and mtRNA within a mouse sample. It was possible to detect two variants showing allelic mismatch between the mtDNA and mtRNA (**Fig. 4.23a**), of which one was on mt-tRNAs (**Fig. 4.23b**). The observed allelic mismatches were, unfortunately, at relatively low levels, e.g. 3902.C>T variant on tRNA M, was observed at AF 4 % and 30 % in mtDNA and mtRNA, respectively. Furthermore, no real alterations in mtRNA coverage of the mutated sample were observed in comparison to other samples which did not harbor the variant in question and were thus considered as controls.

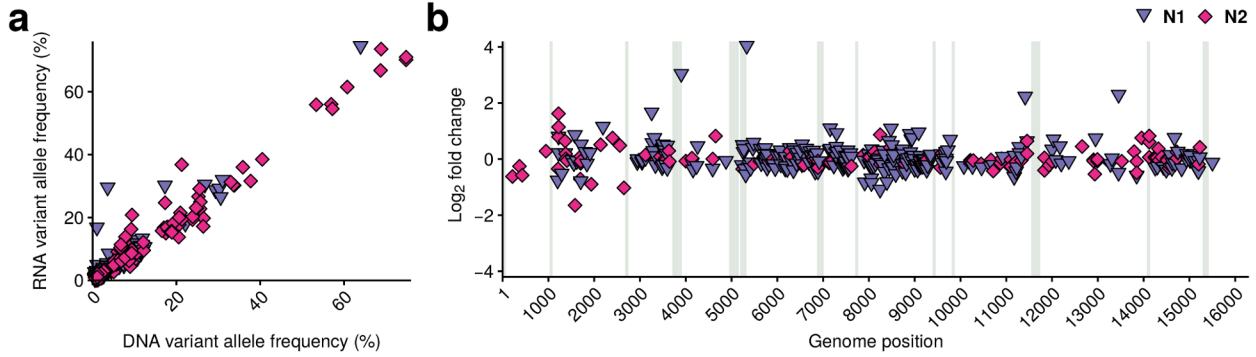


Figure 4.23. Comparison of variant allele frequencies between mitochondrial DNA and RNA. a) *Variant allele frequencies observed for common mtDNA and mtRNA variants were compared by plotting the AF values from both source materials against each other. This revealed mild outliers i.e. variants showing allelic mismatch at lower mtDNA AF levels (<5 %).* ***b)*** *In order to understand the variant site and significance of the allelic mismatch, the allelic mismatch was calculated as log_2-fold-change of mtRNA variant AF over mtDNA variant AF and plotted over the mtDNA genome positions. This comparison did not only reveal the two significant outliers on tRNA M and COI, but also showed that variants – although not showing high allelic mismatch – were detected on many tRNA regions (shaded areas). These were probably normal processing intermediates carrying variants which do not disrupt the mtRNA processing.*

Discussion

Based on these preliminary data, the project might have potential and amplicon sequencing approach seem to fit well for the aim. However, in order to detect more significant allelic mismatches, and thus, potentially also greater alterations in mtRNA coverages, it is advisable to use at least N3 – or even N4 – generation mice. These mice will have fewer but highly clonally expanded variants, which increases the potential of such mutations to show up as allelic mismatches. For example, in the human tumor data, the variants showing allelic mismatches were present in mtDNA at AF ~20–70 % and at AF ~100 % in mtRNA (Stewart et al. 2015), whereas in N1 and N2 generation mice mtDNA the variants showing allelic mismatch were at AF <5 %. In the purifying selection project (**Chapter 4.3.2**), mtDNA-seq samples from N3 generation mice harbored even AF ~80 % variants, whereas maximum AF values were only 55 % and 28 % for N2 and N1 generation mice variants, respectively. Those results suggest, that already N3 generation mice could be more suitable samples for detection of potential allelic mismatch of the mtDNA and mtRNA variants.

For detection of alterations in RNA coverages, wild-type mouse RNA could be used as a better, true control. Now, similar to the human tumor study, the control samples were the other highly mutated N1 and N2 generation mice samples, which did not carry the particular variant in question. However, the high load of neighbouring variants in such control samples could in theory affect the coverage profile, and usage of pure WT samples would be a cleaner approach. Although, with N3 or N4 generation mice, the number of variants is much lower and also their suitability as controls could be investigated. Additionally, with RNA-seq one should consider that the coverage might drastically depend on differences in expression levels between e.g. tissues or age groups.

5 CONCLUSIONS AND FUTURE PROSPECTS

5.1 Optimization of the mitochondrial DNA extraction and sequencing method for extremely low-frequency mitochondrial DNA variant detection

The main findings of the method optimization were:

- Many mitochondria enrichment methods yield in significant nuclear DNA contamination in the extracted mtDNA.
- Treatment of enriched liver and brain mitochondria with DNase I is highly efficient method to diminish nDNA contamination, thus only two Gbase of paired-end sequencing reads are enough to obtain ~40000–60000x depth over the entire mtDNA genome.
- In order not to waste sequencing data or introduce bias, de-duplication should not be included into the data analysis and reads should be aligned directly to mtDNA reference genome with an approach considering the circularity of the genome.
- For good quality mtDNA sample, sensitive variant calling is possible without hard-coded allele frequency thresholds well beyond the sequencing platform error rates, as low as AF 0.05–0.1 %.
- Extremely low-frequency variant detection is susceptible for slight changes in the sample preparation and sequencing protocols, but precision is increased by utilizing dual indices and potentially by adding EDTA to the sonication step or repair enzymes to the library preparation PCR step.

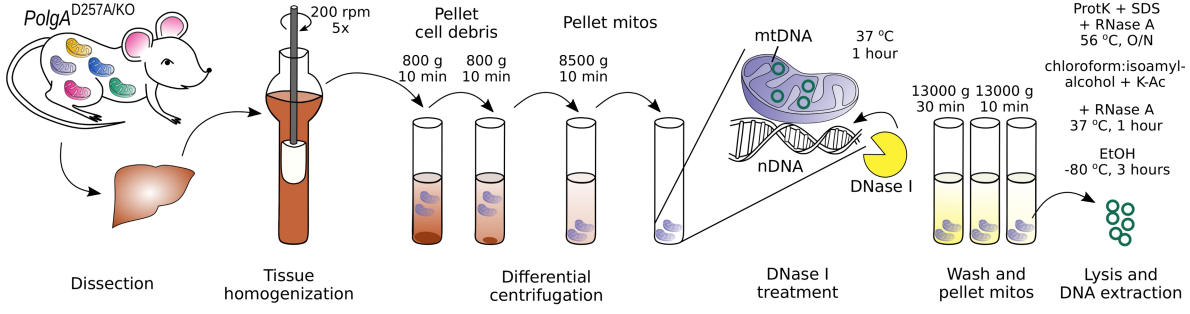
The final optimized workflow to detect rare variants from mtDNA is illustrated in **Figure 5.1**.

For reliable, extremely rare variant detection, it is important to obtain high-quality mtDNA sample pure from nDNA contamination. Although, the experimentation with different methods was not optimally designed due to financial constrains (e.g. such that similar samples or treatments would have been used with all methods), mtDNA-seq (i.e. simple differential centrifugation combined with DNase I treatment of the enriched mitochondria) consistently yielded to very pure, good quality, high-yield mtDNA fractions with minimal hands-on time and low sample preparation costs. Traditionally used gradient-based mitochondria enrichment methods alone also yielded in relatively pure mtDNA fractions when experienced lab personnel performed the enrichment. Furthermore, commercial mitochondria isolation kit might be useful for difficult or precious tissue samples, such as heart.

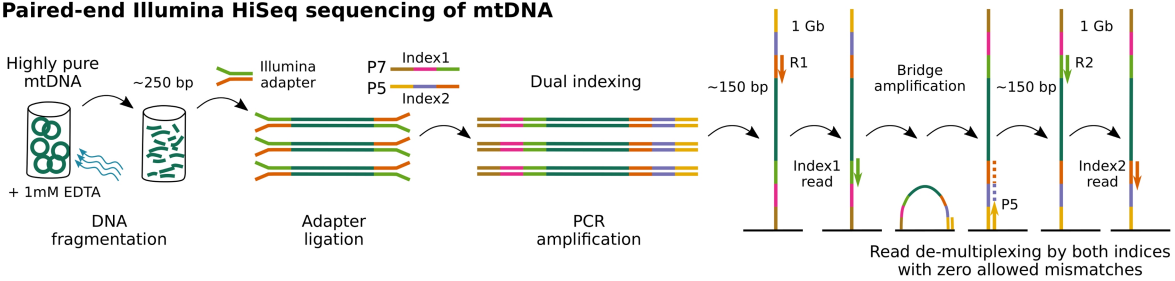
Here, the projects utilized only big tissues – liver and brain – which easily lead to high-yield mtDNA samples. Originally, the required minimum DNA amount for standard Illumina library preparation was 100 ng (Max Planck Genome Centre Cologne). The DNA input requirement is a prohibitive factor in applying the method for e.g. human tissue biopsies, where the sample amount may be very limited.

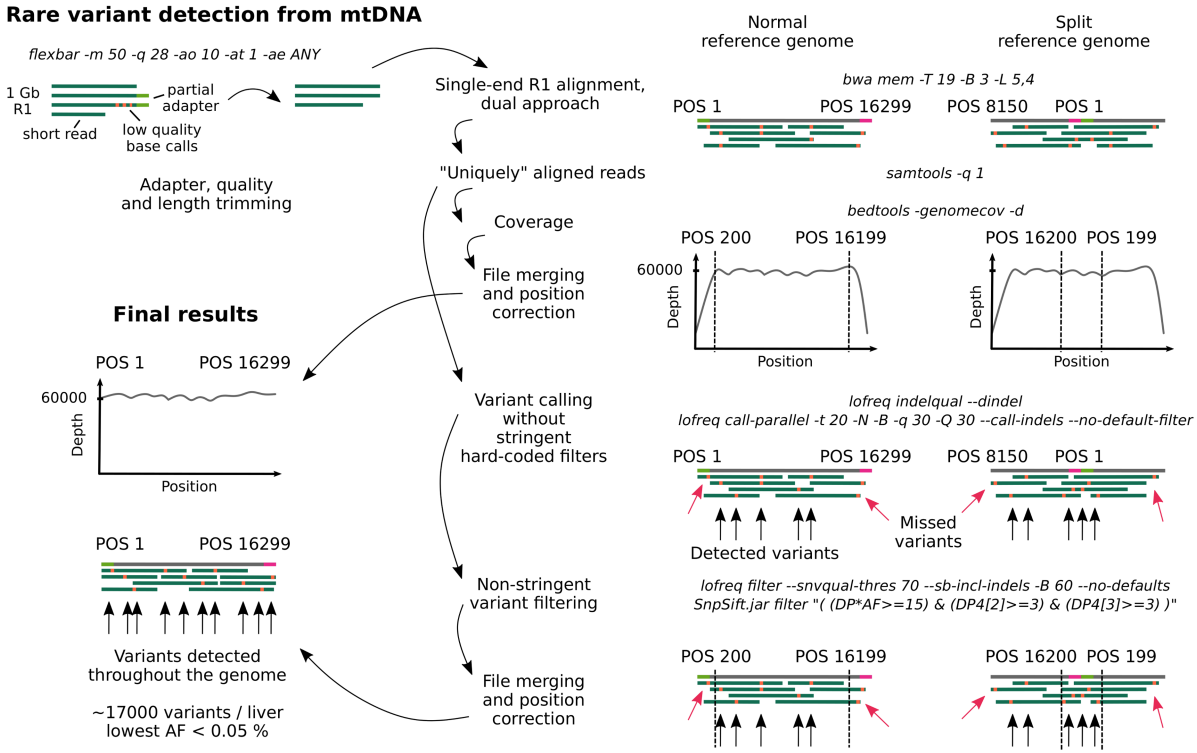
Figure 5.1. Final optimized workflow for extraction of highly pure mtDNA and detection of extremely rare variants.
The protocol for mtDNA extraction is based on simple differential centrifugation enrichment of mitochondria (mitos), which are treated with DNase I in order to remove almost all contaminating nuclear DNA (nDNA). DNA is extracted with chloroform (without phenol to avoid unnecessary risk of introducing DNA damage) and treated extensively with RNase A. Highly pure mtDNA is fragmented in the presence of EDTA in order to diminish the risk of introducing oxidative damage to DNA. DNA fragments are ligated to adapters with dual indexing approach and sequenced in paired-end mode to enable de-multiplexing based on both indices in order to exclude between-sample cross-contamination. Reads are trimmed for high quality and aligned to normal and split reference genomes in order to rescue coverage and variant detection at the junction region of the circular genome. Variant calling does not include stringent filtering and the most effective filter is the strand-bias filtering allowing reliable variant detection even below allele frequency (AF) of 0.05 %. (Figure is on the next page.)

Extraction of highly pure mtDNA



Paired-end Illumina HiSeq sequencing of mtDNA




Rare variant detection from mtDNA
flexbar -m 50 -q 28 -ao 10 -at 1 -ae ANY
1 Gb R1
partial adapter
short read
low quality base calls
Adapter, quality and length trimming
Single-end R1 alignment, dual approach
"Uniquely" aligned reads
Coverage
File merging and position correction
Variant calling without stringent hard-coded filters
Non-stringent variant filtering
File merging and position correction
Final results
POS 1
POS 16299
60000
Depth
Position
Variants detected throughout the genome
~17000 variants / liver
lowest AF < 0.05 %
Normal reference genome
Split reference genome
bwa mem -T 19 -B 3 -L 5,4
POS 1
POS 16299
POS 8150
POS 1
samtools -q 1
bedtools -genomecov -d
POS 200
POS 16199
POS 16200
POS 199
60000
Depth
Position
lofreq indelqual --dindel
lofreq call-parallel -t 20 -N -B -q 30 -Q 30 --call-indels --no-default-filter
Detected variants
Missed variants
lofreq filter --snvqual-thres 70 --sb-incl-indels -B 60 --no-defaults
SnpSift.jar filter "((DP*AF>=15) & (DP4[2]>=3) & (DP4[3]>=3))"

However, as the sequencing library preparation kits are constantly being developed as well as the practices in the sequencing centers, already during these thesis projects, the required DNA input was decreased to 50 ng. This allowed also sequencing of the smaller yield samples, such as heart, and the input DNA amounts will probably be improved to even lower levels in the future. Yet, this development should be cautiously followed, since Costello et al. (2013) suggested that decrease in input amount from 3 µg to 100 ng without proper adaptation of the sonication parameters caused significant increase in oxidative damage of the DNA (Costello et al. 2013).

Another possibility for low-yield samples could be a low-input library preparation kit, which is based on simultaneous enzymatic fragmentation and adapter ligation. This should also be carefully tested with controls, since the fragmentation method is completely different and could cause unexpected artefacts in extremely rare variant detection. Moreover, with the ever decreasing DNA input amounts, one should consider the fact that the sequencing library is always representing only a subsample of the original pool of mtDNA molecules. If the original mtDNA was already a low-yield sample, and only extremely low amount of that is used for the library preparation, even two subsampling bottlenecks will be applied and any effect of a random artefact may become magnified and seen as unintended bias in extremely rare variant detection results.

On the other hand, PCR-free library preparation currently requires 2.5 µg of input DNA (MP-GC), which is reachable with liver mtDNA-seq samples. It would be an interesting comparison to mtDNA-seq variant results, although the PCR-free library preparation includes four cycles of PCR to normalize the pool of ssDNA and dsDNA molecules to dsDNA. However, the starting pool of mtDNA molecules would be 50x larger and PCR cycles are only half of what is applied to standard library preparation. This should at least diminish the detection of library preparation artefacts; however, as indicated by Illumina, it could actually increase the index hopping and thus lead to detection of cross-

contaminating variants (Illumina 2017).

Different mtDNA enrichment methods were compared by their variant profiles. Although, the experimental design was suboptimal (due to the relatively high costs, only single samples or no WT controls were tested by some methods), the results seemed to favor non-amplification based enrichment. Together with the sample preparation optimization results, mtDNA-seq was concluded as the most optimal in comparison to other methods. However, major artefact observed during the experiments was GC>TA variants – a known signature of oxidative damage (Shibutani et al. 1991). Simultaenously with artefactual GC>TA variants, also between-sample cross-contamination suddenly increased to intolerable levels. The simple solution was to switch to paired-end sequencing mode and de-multiplex the reads based on both indices instead of just one, as suggested already in 2012 by Kircher et al. Furthermore, it was shown that trimming or alignment strategy can easily introduce unintended bias to the coverage profile. Thus, a dual alignment and variant calling strategy was applied in order to enhance the variant detection at the junction region of the circular mtDNA genome – a highly important approach for projects focusing on mtDNA control region variant profile.

The key finding of the data analysis optimization for extremely rare mtDNA variant detection was that the removal of any hard-coded variant allele frequency thresholds from the variant calling with LoFreq* did not lead to significant amounts of false positive results. Almost all studies utilizing standard Illumina sequencing for variant detection set a hard-coded AF threshold because of relatively high ”sequencing platform error rate”. Here, it was shown with the spike-in mtDNA-seq samples, that from a high-quality, high-coverage mtDNA sample, LoFreq* is capable of detecting variants with high precision as low as AF <0.05 %. These results clearly stress the impact of the DNA sample quality to the variant detection accuracy. Nevertheless, a major drawback of mtDNA-seq is that the extremely rare variant detection may be highly sensitive to even slight changes in the sample preparation

protocol or in the sequencing platform. Thus, it is advisable to carefully follow the protocol and always include WT control samples in each sequencing run in order to quickly notice unexpected artefacts. It also seems, that with a high-quality DNA sample, it is possible to reliably detect variants well beyond the sequencing platform error rate. Here, only LoFreq* was used. However, it would be interesting to utilize the spike-in samples to compare different variant calling models and their performance in comparison to LoFreq*, especially the recently developed VarDict, which was recently suggested to outperform LoFreq* (Sandmann et al. 2017).

As predicated by Lou et al. (2013), when the aim is to develop even more sensitive variant detection methods, potentially more artefacts will be discovered and these issues need to be tackled. Yet, the development of the data analysis approaches is relatively slow in comparison to how fast the technology evolves. The above-mentioned oxidative damage likely induced during standard sequencing library preparation is a good sample case. Moreover, the switch to patterned flow cell usage has been recently noted to increase index hopping (Illumina 2017, Sinha et al. 2017) supporting the usage of dual indices as a standard method for sensitive applications like low-frequency variant detection. Yet, another recently arised concern is increased duplicate read formation in the patterned flow cells (Wingett 2017, accessed 07/2017). These are just a few examples of how updates and technology advancements may bring along unintended bias. Now that high-throughput sequencing is already a routine method and the hype has mostly calmed down, more solid and properly controlled, peer-reviewed research is required to address and acknowledge the existing and newly introduced caveats and find solutions – before rushing into new applications with stretched sensitivity thresholds e.g. extremely low-frequency variant detection or single-cell sequencing.

5.2 Mitochondrial biology research questions addressed by mtDNA-seq

The optimized mtDNA-seq protocol was succesfully applied to address several mitochondrial biology research questions. The new results obtained within these thesis projects extend the previous studies regarding the variant profile of the entire mtDNA genome, purifying selection of mtDNA as well as processing of the polycistronic mtRNA transcripts.

5.2.1 Creation of variant profile of the entire mitochondrial genome and identification of regions essential for replication and replication-associated transcription

The main findings of the project were:

- The mtDNA mutator mouse is a true saturation mutagenesis model and, with AF <0.05 % variant detection threshold, show >90 % of the genome positions to harbor a variant.
- Liver and brain mtDNA samples show similar trend in their variant profiles.
- Control region harbor significant mutational coldspots, which align with evolutionarily conserved regions.
- Detailed variant profile serves as a valuable resource for studies focusing on mtDNA replication and transcription mechanisms.

Because no methods to transfect mammalian mtDNA exists, mtDNA mutator mouse is one of the only *in vivo* models for mtDNA mutagenesis. With mtDNA-seq it was possible to obtain uniform coverage over the entire mtDNA genome and even to represent the linear, truncated mtDNA molecules and control region multimers (Trifunovic et al. 2004, Williams et al. 2010). The variant profiles were highly similar between the liver and brain mtDNA samples. The only

difference was that brain mtDNA seemed to harbor fewer variants, which can be explained by the fact that more replication is ongoing in the mitotic liver than in the post-mitotic brain tissue. The results were well in line with the previous PCS results, despite the completely different technologies used. Also similar to previous studies, but now with extremely sensitive variant detection, significant mutational coldspots were observed at OriL and control regions, suggesting their importance on mtDNA replication process. This hypothesis was further supported by the alignment of the evolutionarily highly conserved sites with the coldspots.

The data set presented here is an extremely valuable for mtDNA biology research, especially for studies focusing on mechanisms of mtDNA replication and replication-associated transcription or identification and characterization of yet-unknown proteins involved in mtDNA maintenance. The variant profile will be complemented with an *in vivo* methylation assay (as in Rebelo et al. 2009 and Terzioğlu et al. 2013) to reveal potential protein-binding sites by observing protection from methylation at GpC sequence context. Non-methylated bases are converted to Ts in bisulfite treatment and will be detected as C>T variants by pyrosequencing of short amplicons over the GpC-sites at the control region (**Fig. 4.15b**, lowest track). Furthermore, some of the coldspots and variant positions located in the CSB II and III (**Fig. 5.2**) will be tested in *in vitro* transcription assay. In such experiments, synthetic DNA strands, each containing a single variant, are used as a template for transcription and treated with RNase A. Those transcripts resistant to the RNase A treatment, likely form stable RNA-DNA hybrid required for replication initiation (Wanrooij P. et al. 2012). Differences in hybrid stability could reveal those bases essential for replication-associated transcription and support the hypothesis that coldspots are required for efficient replication of the mtDNA. The publication is aimed as a resource for other researchers to create or narrow down hypotheses and to better target their studies, e.g. the data set may be utilized as a source material to form bait sequences for the search of ETAS-binding proteins.

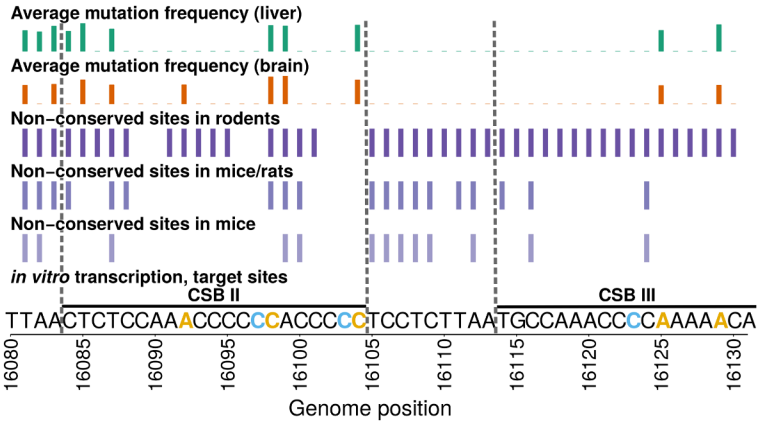


Figure 5.2. Sites to be tested in an in vitro transcription assay. *The lowest track illustrates the sequence content of the conserved sequence blocks (CSB II and III) and the target sites to be tested in an in vitro transcription assay. In this assay, DNA strands are synthesized to contain a single variant each (target bases are marked with colors: yellow = observed variant sites and blue = observed coldspots) and used as templates for transcription reactions, which are then treated with RNase A. If the sample is resistant to the RNase A treatment, it would indicate formation of a stable RNA-DNA hybrid. Differences in the RNA-DNA hybrid stability between the different variant templates could reveal those bases essential for the hybrid formation, thus, essential for the mtDNA replication efficiency. Here, the hypothesis is that DNA templates harboring a variant at the sites observed to be variable (yellow) should form stable RNA-DNA hybrids, whereas DNA templates harboring a variant at the observed coldspots (blue) should show poor RNA-DNA hybrid stability. The experimental design and the actual experiments are carried out by our collaborators, PhD Viktor Posse and Prof. Claes Gustafsson at the University of Gothenburg, Sweden. The dashed grey lines highlight the CSB regions over the other tracks. Other tracks are as in* ***Figure 4.15b****.*

5.2.2 Clarification of developmental stage and mechanism of purifying selection of mitochondrial DNA

The main findings of the preliminary results:

- Strong purifying selection takes place already in N1 generation mice when considering only high-frequency variants.
- Low levels of potentially deleterious mtDNA mutations are transmitted even to N3 generation mice.

Only mild purifying selection was observed when rare mtDNA variants were considered, but the hallmark of purifying selection became visible already in N1 generation mice if only high-frequency variants were considered. Furthermore, if considering all results (without a minimum AF threshold), the hallmark of purifying selection was mild even in N3 generation mice. These data suggest that even deleterious mutations are transmitted to the offspring, however, such mutations are not highly clonally expanded.

One original aim of this project was to apply the mtDNA-seq also to different stage embryos or even to oocytes in order to gain deeper understanding on the developmental stage of the purifying selection. However, all attempts to obtain highly pure mtDNA from N2 E14 embryos (in addition to mtDNA-seq, also ExoV treatment or extraction of the mtDNA bands from agarose gel was experimented but not discussed within this thesis) failed. Furthermore, RCA did not turn out as a successful method for extremely low-frequence variant detection neither did amplicon sequencing. Thus, the future challenge to establish an effective mtDNA enrichment method for embryos remains to be solved. On the other hand, the results showed only mild difference between N1 and N2 generation mice, thus sequencing of different stage embryos in between these generations would not add up much information, if any. More interesting would be variant profiles of N1 E7.5 and E14 embryos.

For the final conclusions, the additional data from two more lineages should be included into the analysis as here only two lineages were analyzed. Furthermore, in addition to simply plotting the relative amounts of variants on different genome elements, there are multitude of other analysis options: e.g. the distribution of variants on certain amino acids (e.g. hydrophobic vs. hydrophilic) could be compared to reveal whether certain type of mutations are more tolerated than others, detailed characterization of the variant distribution on different codon bases to detect whether certain codon compositions are preferred, follow up of individual variants within the lineage to observe the variant frequency fluctuations along the line or between litters and whether it follows any pattern, e.g. in case of tRNA variants. Moreover, here, the data were analyzed such that two samples per family was considered. However, as a lot more mice have been sequenced than what was included into such analysis, more power can be obtained by including all sequenced samples and the confounding problem of variable number of sequenced littermates can be eliminated by counting each variant only once per relative mice per generation.

5.2.3 Effects of mitochondrial DNA variants on mitochondrial RNA processing

The main findings of the preliminary results:

- Amplicon sequencing is suitable approach for a study focusing on high-frequency (AF >1 %) mtDNA variants.
- N1 and N2 generation mice do not show high alleleic mismatch between mtDNA and mtRNA variants.

The preliminary results suggested that when searching for variants potentially causing mtRNA processing defects visible by RNA-seq, the variant likely has to be present in relatively high fraction of the mtDNA

(AF >20 %). Thus, it would be advisable to continue the project by utilizing later generation of mice – at least N3 or even N4 generation. As these mice carry less variants, the number of mice required for the project is likely substantially increased. As the same mouse breeding scheme was utilized in the purifying selection project, mtDNA-seq data from N2 and N3 generation mice could be utilized to screen for prevalence of high-frequency tRNA variants and their transmission. During the purifying selection project, also tissue samples were stored, thus the mtDNA-seq results can be used to select interesting N3 generation tissue samples for a pilot RNA-seq experiment.

Moreover, recently Kuznetsova et al. (2017) utilized circular RNA sequencing (Chu et al. 2015) to characterize unprocessed RNA molecules (Kuznetsova et al. 2017). In that method, the RNA molecules are first circlized by intramolecular ligation and only then reverse transcribed with adapter-containing random primers. This way the 5' and 3' end information of the RNA molecule is preserved in comparison to traditionally prepared RNA-seq libarary (Kuznetsova et al. 2017). Furthermore, this method allows also simultaneous sequencing of small RNA products, e.g. mature tRNAs carrying post-transcriptionally added CCA at 3' end (Kuznetsova et al. 2017).

Although the computational steps of circular RNA-seq are more complex than in standard RNA-seq, it seems preferable method for this project. The human tumor study was limited to poly-A-enriched RNAs, whereas here only ribosome depleted RNA was used. Thus, circular RNA-seq would provide more complete picture of the RNA pool. Moreover, detection of CCAs may provide highly valuable information on whether the variant harboring tRNAs are actually ever processed to the mature form or not. This data may aid in planning the molecular characterization experiments or interpreting *in silico* tRNA secondary structure predictions.

5.3 Summary

Within this thesis, mtDNA extraction from mouse liver and brain tissues was optimized such that extremely pure mtDNA can be obtained in comparison to traditional mitochondria enrichment methods. High-quality mtDNA samples were used to optimize cheap and fast sequencing protocol for extremely rare mtDNA variant detection. The analysis steps were improved to account for the characteristics of the small, circular mtDNA genome, and the final validation of the protocol was shown with spike-in samples. Observation and elimination of artefactual variants raised recommendations that standard sequencing library preparation practices should include a repair step and sequencing in paired-end mode with dual indices, although, only R1 should be utilized for the extremely low-frequency variant detection.

The mtDNA-seq approach was successfully applied to address different research questions in mitochondrial biology: A detailed mtDNA variant profile was created to aid the research focusing on mtDNA replication and transcription mechanisms. The method enabled sensitive detection of mtDNA variants in early mouse generations and will help to clarify the timing and mechanism of mtDNA purifying selection. Finally, the variant data sets can be combined with RNA-seq experiments of the collected tissues in order to study mitochondrial RNA processing. Detailed knowledge on the key processes involved in mtDNA maintenance are in the end the basis for development of treatments for mitochondrial disorders as well as measures to prevent transmission of pathogenic mtDNA mutations.

6 REFERENCES

Adey, A. et al., 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.*, 11(12), p.R119.

Altmann, A. et al., 2012. A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Hum Genet.*, 131, pp.1541–1554.

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), pp.403–410.

Ameur, A. et al., 2011. Ultra-deep sequencing of mouse mitochondrial DNA: Mutational patterns and their origins. *PLoS Genet.*, 7(3), pp.18–21.

Arbeithuber, B., Makova, K.D. & Tiemann-Boege, I., 2016. Artifactual mutations resulting from DNA lesions limit detection levels in ultrasensitive sequencing applications. *DNA Res.*, 23(6), pp.547–559.

van der Auwera, G.A. et al., 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr Protoc Bioinformatics*, 11(1110), pp.11.10.1–11.10.33.

Bainbridge, M.N. et al., 2010. Whole exome capture in solution with 3 Gbp of data. *Genome Biol.*, 11(6), p.R62.

Baines, H.L. et al., 2014. Similar patterns of clonally expanded somatic mtDNA mutations in the colon of heterozygous mtDNA mutator mice and ageing humans. *Mech Ageing Dev.*, 139(1), pp.22–30.

Balzer, S. et al., 2013. Filtering duplicate reads from 454 pyrosequencing data. *Bioinformatics*, 29(7), pp.830–836.

Boore, J., Macey, M. & Medina, N., 2005. Sequencing and Comparing Whole Mitochondrial Genomes of Animals. *Methods in Enzymol.*, 395, 311–318.

Bradnam, K., 2015. More madness with MAPQ scores (a.k.a. why bioinformaticians hate poor and incomplete soft-ware documentation). *ACGT*. Available at: http://www.acgt.me/blog/2015/3/17/more-madness-with-mapq-scores-aka-why-bioinformaticians-hate-poor-and-incomplete-software-documentation [Accessed August 5, 2017].

Broad Institution. MarkDuplicates tool documentation GitHub pages. Available at: https://broadinstitute.github.io/ picard/command-line-overview.html#MarkDuplicates [Accessed August 29, 2017].

Burrows, M. & Wheeler, D., 1994. A Block-sorting Lossless Data Compression Algorithm, Research Report, Digital Systems Research Center, Palo Alto, California.

Cabral Neto, J.B. et al., 1992. Mutation Spectrum of Heat-induced Abasic Sites on a Single-stranded Shuttle Vector Replicated in Mammalian Cells. *J Biol Chem.*, 267(27), pp.19718–19723.

Calabrese, C. et al., 2014. MToolBox: A highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinformatics*, 30(21), pp.3115–3117.

Calabrese, F.M., Simone, D. & Attimonelli, M., 2012. Primates and mouse NumtS in the UCSC Genome Browser. *BMC Bioinformatics*, 13, p.S15.

Chen, G. et al., 2014. Cytosine deamination is a major cause of baseline noise in next-generation sequencing. *Mol Diagn Ther.*, 18(5), pp.587–593.

Chen, L. et al., 2017. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, 355, pp.752–756.

Chinnery, P.F. & Samuels, D.C., 1999. Relaxed Replication of mtDNA: A Model with Implications for the Expression of Disease. *Am J Hum Genet.*, 64(4), pp.1158–1165.

Chinnery, P.F. et al., 2014. The Challenges of Mitochondrial Replacement. *PLoS Genet.*, 10(4), pp.3–4.

Chu, Y. et al., 2015. Intramolecular circularization increases efficiency of RNA sequencing and enables CLIP-Seq of nuclear RNA from human cells. *Nucleic Acids Res.*, 43(11), p.e75.

Cingolani, P. et al., 2012a. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, 6(2), pp.80–92.

Cingolani, P. et al., 2012b. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front Genet.*, 3(35), doi: 10.3389/fgene.2012.00035.

Claycamp, H.G., 1992. Phenol sensitization of dna to subsequent oxidative damage in 8-hydroxyguanine assays. *Carcinogenesis*, 13(7), pp.1289–1292.

Costello, M. et al., 2013. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.*, 41(6), pp.1–12.

Cree, L.M., Samuels, D.C. & Chinnery, P.F., 2009. The inheritance of pathogenic mitochondrial DNA mutations. *Biochim Biophys Acta*, 1792(12), pp.1097–1102.

Cui, H. et al., 2013. Comprehensive next-generation sequence analyses of the entire mitochondrial genome reveal new insights into the molecular diagnosis of mitochondrial DNA disorders. *Genet Med.*, 15(5), pp.388–394.

Dames, S. et al., 2013. The Development of Next-Generation Sequencing Assays for the Mitochondrial Genome and 108 Nuclear Genes Associated with Mitochondrial Disorders. *J Mol Diagn.*, 15(4), pp.526–534.

Dean, F.B. et al., 2001. Polymerase and Multiply-Primed Rolling Circle Amplification Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification from Colonies or Plaques. *Methods*, 11, pp.1095–1099.

DePristo, M.A. et al., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.*, 43(5), pp.491–498.

Diegoli, T.M. et al., 2012. An optimized protocol for forensic application of the PreCR™ Repair Mix to multiplex STR amplification of UV-damaged DNA. *Forensic Sci Int Genet.*, 6(4), pp.498–503.

Ding, J. et al., 2015. Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of ~2,000 Sardinians Using Tailored Sequencing Analysis Tools. *PLoS Genet.* 11(7), p.e1005306.

Dobin, A. et al., 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), pp.15–21.

Doda, J.N., Wright, C.T. & Clayton, D.A., 1981. Elongation of displacement-loop strands in human and mouse mitochondrial DNA is arrested near specific template sequences. *Proc Natl Acad Sci.*, 78(10), pp.6116–6120.

Dodt, M. et al., 2012. FLEXBAR-Flexible Barcode and Adapter Processing for Next-

Generation Sequencing Platforms. *Biology*, 1(3), pp.895–905.

Dressman, D. et al., 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci.*, 100(15), pp.8817–8822.

Durham, S.E. et al., 2007. Normal levels of wild-type mitochondrial DNA maintain cytochrome c oxidase activity for two pathogenic mitochondrial DNA mutations but not for m.3243A->G. *Am J Hum Genet.*, 81(1), pp.189–195.

Dyer, N., Young, L. & Ott, S., 2015. Artifacts in the data of Hu et al. *Nat Genet.*, 48(1), pp.2–3.

Eckert, K.A. & Kunkel, T.A., 1991. DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl.*, 1(1), pp.17–24.

Edgar, D. & Trifunovic, A., 2009. The mtDNA mutator mouse: Dissecting mitochondrial involvement in aging. *Aging*, 1(12), pp.1028–1032.

Elliott, H.R. et al., 2008. Pathogenic Mitochondrial DNA Mutations Are Common in the General Population. *Am J Hum Genet.*, 83(2), pp.254–260.

Esteban, J.A., Salas, M. & Blanco, L., 1993. Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *J Biol Chem.*, 268(4), pp.2719–2726.

Etherington, G., 2014. Bioinformatics Bits and Bobs: Why you should QC your reads AND your assembly. Available at: http://grahametherington.blogspot.fi/2014/09/why-you-should-qc-your-reads-and-your.html [Accessed August 8, 2017].

Ewing, B. et al., 1998. Base-calling of automated sequencer traces using Phred I Accuracy assessment. *Genome Res.*, 8(3), pp.175–185.

Del Fabbro, C. et al., 2013. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLoS ONE*, 8(12), p.e85024.

Fan, W. et al., 2008. A Mouse Model of Mitochondrial Disease Reveals Germline Selection Against Severe mtDNA Mutations. *Science*, 319, pp.958–963.

Farge, G. et al., 2014. In Vitro-Reconstituted Nucleoids Can Block Mitochondrial DNA Replication and Transcription. *Cell Rep.*, 8(1), pp.66–74.

Fisher, R.P. et al., 1992. DNA wrapping and bending by a mitochondrial high mobility group-like transcriptional activator protein. *J Biol Chem.*, 267(5), pp.3358–3367.

Fisher, R.P. & Clayton, D.A., 1988. Purification and characterization of human mitochondrial transcription factor 1. *Mol Cell Biol.*, 8(8), pp.3496–509.

Fonseca, HTS Mappers. Available at: http://www.ebi.ac.uk/~nf/hts_mappers/ [Accessed August 5, 2017].

Fox, E.J. et al., 2014. Accuracy of Next Generation Sequencing Platforms. *Next Gener Seq Appl.*, 1(1), p.1000106.

Franko, A. et al., 2013. Efficient isolation of pure and functional mitochondria from mouse tissues using automated tissue disruption and enrichment with anti-TOM22 magnetic beads. *PLoS One*, 8(12), p.e82392.

Frezza, C., Cipolat, S. & Scorrano, L., 2007. Organelle isolation: functional mitochondria from mouse liver, muscle and cultured fibroblasts. *Nat Protoc.*, 2(2), pp.287–295.

Fuller, C. et al., 2009. The challanges of sequencing by synthesis. *Nat Biotechnol.*,

27(11), pp.1013–1023.

Gardner, K. et al., 2015. Use of stereotypical mutational motifs to define resolution limits for the ultra-deep resequencing of mitochondrial DNA. *Eur J Hum Genet.*, 23(3), pp.413–45.

Gaspari, M., Larsson, N.G. & Gustafsson, C.M., 2004. The transcription machinery in mammalian mitochondria. *Biochim Biophys Acta*, 1659(2–3), pp.148–152.

Gilkerson, R.W., 2009. Mitochondrial DNA nucleoids determine mitochondrial genetics and dysfunction. *International Journal of Biochemistry and Cell Biol.*, 41(10), pp.1899–1906.

van Goethem, G. et al., 2001. Mutation of POLG is associated with progressive external ophthalmoplegia characterized by mtDNA deletions. *Nat Genet.*, 28(3), pp.211–212.

Goodwin, S., McPherson, J.D. & McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet.*, 17(6), pp.333–351.

Gorman, G.S. et al., 2015. Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. *Ann Neurol.*, 77(5), pp.753–759.

Gould, M.P. et al., 2015. PCR-free enrichment of mitochondrial DNA from human blood and cell lines for high quality next-generation DNA sequencing. *PLoS ONE*, 10(10), p.e0139253.

Gray, M.W., 2012. Mitochondrial evolution. *Cold Spring Harb Perspect Biol.*, 4(9), p.a011403.

Greaves, L.C. et al., 2009. Quantification of mitochondrial DNA mutation load. *Aging Cell*, 8(5), pp.566–572.

Greaves, L.C. et al., 2014. Clonal Expansion of Early to Mid-Life Mitochondrial DNA Point Mutations Drives Mitochondrial Dysfunction during Human Ageing. *PLoS Genet.*, 10(9), p.e1004620.

Gregory, M.T. et al., 2015. Targeted single molecule mutation detection with massively parallel sequencing. *Nucleic Acids Res.*, 44(3), p.e22.

Griffin, H.R. et al., 2014. Accurate mitochondrial DNA sequencing using off-target reads provides a single test to identify pathogenic point mutations. *Genet Med.*, 16(12), pp.962–971.

Gu, Z. et al., 2014. Circlize implements and enhances circular visualization in R. *Bioinformatics*, 30(19), pp.2811–2812.

Guo, Y. et al., 2013. MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. *Bioinformatics*, 29(9), pp.1210–1211.

Guo, Y. et al., 2012. The use of next generation sequencing technology to study the effect of radiation therapy on mitochondrial DNA mutation. *Mut Res.*, 744, pp.154–160.

Gustafsson, C.M., Falkenberg, M. & Larsson, N.-G., 2016. Maintenance and Expression of Mammalian Mitochondrial DNA. *Annu Rev Biochem.*, 85(1), pp.133–160.

Hadfield, J., 2016. Index mis-assignment between samples on HiSeq 4000 and X-Ten - Enseqlopedia. *Enseqlopedia*. Available at: http://enseqlopedia.com/2016/12/index-mis-assignment-between-samples-on-hiseq-4000-and-x-ten/ [Accessed August 6, 2017].

Hagström, E. et al., 2014. No recombination of mtDNA after heteroplasmy for 50 generations in the mouse maternal germline. *Nucleic Acids Res.*, 42(2), pp.1111–1116.

Hämäläinen, R.H. et al., 2015. mtDNA Mutagenesis Disrupts Pluripotent Stem Cell Function by Altering Redox Signaling. *Cell Rep.*, 11(10), pp.1614–1624.

Hance, N., Ekstrand, M.I. & Trifunovic, A., 2005. Mitochondrial DNA polymerase gamma is essential for mammalian embryogenesis. *Hum Mol Genet.*, 14(13), pp.1775–1783.

Hancock, D.K., Tully, L.A. & Levin, B.C., 2005. A Standard Reference Material to determine the sensitivity of techniques for detecting low-frequency mutations, SNPs, and heteroplasmies in mitochondrial DNA. *Genomics*, 86, pp.446–461.

Hauswirth, W.W. & Laipist, P.J., 1982. Mitochondrial DNA polymorphism in a maternal lineage of Holstein cows. *Genetics*, 79, pp.4686–4690.

van Haute, L. et al., 2015. Mitochondrial transcript maturation and its disorders. *J Inheri Metab Dis.*, 38(4), pp.655–680.

Hazkani-Covo, E., Zeller, R.M. & Martin, W., 2010. Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet.*, 6(2), p.e1000834.

He, Y. et al., 2010. Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature*, 464(25), pp.610–614.

Huang, H.W., Mullikin, J.C. & Hansen, N.F., 2015. Evaluation of variant detection software for pooled next-generation sequence data. *BMC Bioinformatics*, 16(235).

Illumina, 2015. Patterned Flow Cell Technology. Available at: https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/patterned-flow-cell-technology-technical-note-770-2015-010.pdf [Accessed August 6, 2017].

Illumina, 2016. Illumina Two-Channel SBS Sequencing Technology. Available at: https://www.illumina.com/ content/dam/illumina-marketing/documents/products/techspotlights/techspotlight_two-channel_sbs.pdf [Accessed August 6, 2017].

Illumina, 2017. Effects of Index Misassignment on Multiplexing and Downstream Analysis. Available at: https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf?linkId=36607862 [Accessed July 19, 2017].

Jackson, D.A., Bartlett, J. & Cook, P.R., 1996. Sequences attaching loops of nuclear and mitochondrial DNA to underlying structures in human cells: the role of transcription units. *Nucleic Acids Res.*, 24(7), pp.1212–1219.

Jain, M. et al., 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.*, 17(239).

Jayaprakash, A.D. et al., 2015. Stable heteroplasmy at the single-cell level is facilitated by intercellular exchange of mtDNA. *Nucleic Acids Res.*, 43(4), pp.2177–2187.

Jemt, E. et al., 2015. Regulation of DNA replication at the end of the mitochondrial D-loop involves the helicase TWINKLE and a conserved sequence element. *Nucleic Acids Res.*, 43(19), pp.9262–9275.

Jenuth, J.P., Peterson, a C. & Shoubridge, E. a, 1997. Tissue-specific selection for different mtDNA genotypes in heteroplasmic mice. *Nat Genet.*, 16(1), pp.93–95.

Jiang, H. et al., 2014. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, 15(182).

Johne, R. et al., 2009. Rolling-circle amplification of viral DNA genomes using phi29 polymerase. *Trends Microbiol.*, 17(5), pp.205–211.

Jokinen, R. & Battersby, B.J., 2013. Insight into mammalian mitochondrial DNA segregation. *Ann Med.*, 45(2), pp.149–155.

Just, R.S., Irwin, J.A. & Parson, W., 2014. Questioning the prevalence and reliability of human mitochondrial DNA heteroplasmy from massively parallel sequencing data. *Proc Natl Acad Sci.*, 111(43), pp.E4546-4547.

Kang, E. et al., 2016. Mitochondrial replacement in human oocytes carrying pathogenic mitochondrial DNA mutations. *Nature*, 540(218), pp.270–275.

Kasamatsu, H., Robberson, D.L. & Vinograd, J., 1971. A Novel Closed-Circular Mitochondrial DNA with Properties of a Replicating Intermediate, *Proc Natl Acad Sci.*, 68(9), pp.2252–2257.

Katoh, K. & Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol.*, 30(4), pp.772–780.

Kauppila, J.H.K. et al., 2016. A Phenotype-Driven Approach to Generate Mouse Models with Pathogenic mtDNA Mutations Causing Mitochondrial Disease. *Cell Rep.*, 16(11), pp.2980–2990.

Kauppila, J.H.K. & Stewart, J.B., 2015. Mitochondrial DNA: Radically free of free-radical driven mutations. *Biochim Biophys Acta*, 1847(11), pp.1354–1361.

Kelly, P.S. et al., 2017. Ultra-deep next generation mitochondrial genome sequencing reveals widespread heteroplasmy in Chinese hamster ovary cells. *Metab Eng.*, 41, pp.11–22.

Kennedy, S.R. et al., 2014. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc.*, 9, pp.2586–2606.

Kennedy, S.R. et al., 2013. Ultra-Sensitive Sequencing Reveals an Age-Related Increase in Somatic Mitochondrial Mutations That Are Inconsistent with Oxidative Damage. *PLoS Genet.*, 9(9), p.e1003794.

Kinde, I. et al., 2011. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci.*, 108(23), pp.9530–9535.

Kircher, M., Sawyer, S. & Meyer, M., 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.*, 40(1), p.e3.

Koboldt, D.C. et al., 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17), pp.2283–2285.

Kraytsberg, Y. et al., 2008. Single molecule PCR in mtDNA mutational analysis: Genuine mutations vs. damage bypass-derived artifacts. *Methods*, 46(4), pp.269–273.

Kraytsberg, Y. & Khrapko, K., 2005. Single-molecule PCR: an artifact-free PCR approach for the analysis of somatic mutations. *Expert Rev Mol Diagn.*, 5(5), pp.809–815.

Kujoth, G.C., et al., 2005. Mitochondrial DNA Mutations, Oxidative Stress, and Apoptosis in Mammalian Aging. *Science*, 309(5733), pp.481–484.

Kukat, A. et al., 2011. Random mtDNA mutations modulate proliferation capacity in

mouse embryonic fibroblasts. *Biochem Biophys Res Commun.*, 409(3), pp.394–399.

Kukat, C. et al., 2011. Super-resolution microscopy reveals that mammalian mitochondrial nucleoids have a uniform size and frequently contain a single copy of mtDNA. *Proc Natl Acad Sci.*, 108(33), pp.13534–13539.

Kukat, C. et al., 2015. Cross-strand binding of TFAM to a single mtDNA molecule forms the mitochondrial nucleoid. *Proc Natl Acad Sci.*, 112(36), pp.11288–11293.

Kuznetsova, I. et al., 2017. Simultaneous processing and degradation of mitochondrial RNAs revealed by circularized RNA sequencing. *Nucleic Acids Res.*, 45(9), pp.5487–5500.

Labbé, K., Murley, A. & Nunnari, J., 2014. Determinants and functions of mitochondrial behavior. *Annu Rev Cell Dev Biol.*, 30, pp.357–391.

Lai, Z. et al., 2016. VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.*, 44(11), p.e108.

Langmead, B. et al., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3), p.R25.

Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4), pp.357–359.

Larizza, A. et al., 2002. Lineage specificity of the evolutionary dynamics of the mtDNA D-loop region in rodents. *J Mol Evol.*, 54(2), pp.145–155.

Larsson, N.G., 2010. Somatic mitochondrial DNA mutations in mammalian aging. *Annu Rev Biochem.*, 79, pp.683–706.

Levene, M.J. et al., 2003. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science*, 299(5607), pp.682–686.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv 1303.3997*.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078–2079.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), pp.1754–1760.

Li, M. et al., 2010. Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. *Am J Hum Gen.*, 87(2), pp.237–249.

Li, M. et al., 2012. Fidelity of capture-enrichment for mtDNA genome sequencing: Influence of NUMTs. *Nucleic Acids Res.*, 40(18), p.e137.

Li, M. & Stoneking, M., 2012. A new approach for detecting low-level mutations in next-generation sequence data. *Genome Biol.*, 13(5), p.R34.

Li, M. et al., 2015. Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations. *Proc Natl Acad Sci.*, 112(8), pp.2491–2496.

Linck, E., 2017. Right reads, wrong index? Concerns with data from Illumina's HiSeq 4000. *The Molecular Ecologist*. Available at: http://www.molecularecologist.com/2017/04/right-reads-wrong-index-concerns-with-data-from-illuminas-hiseq-4000/ [Accessed August 6, 2017].

Lou, D.I. et al., 2013. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc Natl Acad Sci.*, 110(49), pp.19872–19877.

Lundberg, K.S. et al., 1991. High-fidelity amplification using a thermostable DNA polymerase isolated from Pyrococcus furiosus. *Gene*, 108(1), pp.1–6.

Luo, S.-M. et al., 2013. Unique insights into maternal mitochondrial inheritance in mice. *Proc Natl Acad Sci.*, 110(32), pp.13038–43.

Macao, B. et al., 2015. The exonuclease activity of DNA polymerase γ is required for ligation during mitochondrial DNA replication. *Nat Commun.*, 6, p.7303.

Madsen, C.S., Ghivizzani, S.C. & Hauswirth, W.W., 1993. Protein Binding to a Single Termination-Associated Sequence in the Mitochondrial DNA D-Loop Region. *Mol Cell Biol.*, 13(4), pp.2162–2171.

Malik, A.N., Czajka, A. & Cunningham, P., 2016. Accurate quantification of mouse mitochondrial DNA without co-amplification of nuclear mitochondrial insertion sequences. *Mitochondrion*, 29, pp.59–64.

Maricic, T., Whitten, M. & Pääbo, S., 2010. Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products R. C. Fleischer, ed. *PLoS ONE*, 5(11), p.e14004.

Marquis, J. et al., 2017. MitoRS, a method for high throughput, sensitive, and accurate detection of mitochondrial DNA heteroplasmy. *BMC Genomics*, 18(1), p.326.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), p.10.

Martin, W.F., Garg, S. & Zimorski, V., 2015. Endosymbiotic theories for eukaryote origin. *Philos Trans R Soc Lond B Biol Sci.*, 370(1678), p.20140330.

McElhoe, J.A. et al., 2014. Development and assessment of an optimized next-generation DNA sequencing approach for the mtgenome using the Illumina MiSeq. *Forensic Sci Int Genet.*, 13, pp.20–29.

McElroy, K. et al., 2013. Accurate single nucleotide variant detection in viral populations by combining probabilistic clustering with a statistical test of strand bias. *BMC Genomics*, 14(501).

Moraes, C.T. et al., 2003. Techniques and pitfalls in the detection of pathogenic mitochondrial DNA mutations. *J Mol Diagn.*, 5(4), pp.197–208.

Nesbitt, V. et al., 2013. The UK MRC Mitochondrial Disease Patient Cohort Study: clinical phenotypes associated with the m.3243A>G mutation--implications for diagnosis and management. *J Neurol Neurosurg Psychiatry*, 84(8), pp.936–938.

Neuwirth, E., 2014. RColorBrewer: ColorBrewer Palettes. Available at: http://cran.r-project.org/package=RColorBrewer.

Ngo, H., Kaiser, J. & Chan, D., 2011. The mitochondrial transcription and packaging factor Tfam imposes a U-turn on mitochondrial DNA. *Nat Struct Mol Biol.*, 18(11), pp.1290–1296.

Ni, T. et al., 2015. MitoRCA-seq reveals unbalanced cytocine to thymine transition in Polg mutant mice. *Sci Rep.*, 5, p.12049.

Nicholls, T.J. & Minczuk, M., 2014. In D-loop: 40 years of mitochondrial 7S DNA. *Exp. Gerontol.*, 56, pp.175–181.

Nielsen, R. et al., 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet.*, 12(6), pp.443–451.

Niu, B. et al., 2010. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics*, 11(187).

Nunnari, J. & Suomalainen, A., 2012. Mitochondria: In Sickness and in Health. *Cell*, 148(6), pp.1145–1159.

Ojala, D., Montoya, J. & Attardi, G., 1981. tRNA punctuation model of RNA processing in human mitochondria. *Nature*, 290, pp.470–474.

Otto, C., Stadler, P.F. & Hoffmann, S., 2014. Lacking alignments? The next-generation sequencing mapper segemehl revisited. *Bioinformatics*, 30(13), pp.1837–1843.

Pabinger, S. et al., 2013. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform.*, 15(2), pp.256–278.

Paez, J.G. et al., 2004. Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res.*, 32(9), p.e71.

Patananan, A.N. et al., 2016. Modifying the Mitochondrial Genome. *Cell Metab.*, 23(5), pp.785–796.

Payne, B.A.I. et al., 2011. Mitochondrial aging is accelerated by anti-retroviral therapy through the clonal expansion of mtDNA mutations. *Nat Genet.*, 43(8), pp.806–810.

Payne, B.A.I. et al., 2013. Universal heteroplasmy of human mitochondrial DNA. *Hum Mol Genet.*, 22(2), pp.384–390.

Payne, B.A.I. et al., 2015. Deep Resequencing of Mitochondrial DNA. In *Methods Mol Biol.*, 1264, pp.59–66.

Picardi, E. & Pesole, G., 2012. Mitochondrial genomes gleaned from human whole-exome sequencing. *Nat Methods*, 9(6), pp.523–524.

Poptsova, M.S. et al., 2014. Non-random DNA fragmentation in next-generation sequencing. *Sci Rep.*, 4, p.4532.

Posse, V. et al., 2014. The amino terminal extension of mammalian mitochondrial RNA polymerase ensures promoter specific transcription initiation. *Nucleic Acids Res.*, 42(6), pp.3638–3647.

Posse, V. et al., 2015. TEFM is a potent stimulator of mitochondrial transcription elongation in vitro. *Nucleic Acids Res.*, 43(5), pp.2615–2624.

Poulton, J. & Bredenoord, A.L., 2010. 174th ENMC International Workshop: Applying pre-implantation genetic diagnostic to mtDNA diseases: Implications of scientific advances 19-21 March 2010. *Neuromuscul Disord.* 20, pp.559–563.

Poulton, J. et al., 2010. Transmission of mitochondrial DNA diseases and ways to prevent them. *PLoS Genet.*, 6(8), p.e1001066.

Preston, J.L. et al., 2016. High-specificity detection of rare alleles with Paired-End Low Error Sequencing (PELE-Seq). *BMC Genomics*, 17(464).

Pyle, A. et al., 2007. Depletion of mitochondrial DNA in leucocytes harbouring the 3243A->G mtDNA mutation. *J Med Genet.*, 44(1), pp.69–74.

Pyle, A. et al., 2015. Extreme-Depth Re-sequencing of Mitochondrial DNA Finds No Evidence of Paternal Transmission in Humans. *PLoS Genet.*, 11(5), p.e1005040.

Quinlan, A.R. & Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), pp.841–842.

Quispe-tintaya, W. et al., 2015. Fast mitochondrial DNA isolation from mammalian cells for next-generation sequencing. *Biotechniques*, 55(3), pp.133–136.

Rebelo, A.P., Williams, S.L. & Moraes, C.T., 2009. In vivo methylation of mtDNA reveals the dynamics of protein-mtDNA interactions. *Nucleic Acids Res.*, 37(20),

pp.6701–6715.

Reinert, K. et al., 2015. Alignment of Next-Generation Sequencing Reads. *Annu Rev Genomics Hum Genet.*, 16, pp.133–151.

Rohlin, A. et al., 2009. Parallel sequencing used in detection of mosaic mutations: Comparison with four diagnostic DNA screening techniques. *Hum Mut.*, 30(6), pp.1012–1020.

Ross, J.M. et al., 2013. Germline mitochondrial DNA mutations aggravate ageing and can impair brain development. *Nature*, 501(7476), pp.412–415.

Ross, J.M. et al., 2014. Maternally transmitted mitochondrial DNA mutations can reduce lifespan. *Sci Rep.*, 4, p.6569.

Rothberg, J.M. et al., 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356), pp.348–352.

Rovio, A., 2006. DNA Polymerase Gamma Mutations in Male Infertility and Ageing, Academic dissertation, University of Tampere, Institute of Medical Technology, Finland, *Acta Universitatis Tamperensis* 1159, ISBN 951-44-6664-0.

Rubio-Cosials, A. et al., 2011. Human mitochondrial transcription factor A induces a U-turn structure in the light strand promoter. *Nat Struct Mol Biol.*, 18(11), pp.1281–1289.

Sambrook, J. & Russell, D.W., 2006. Fragmentation of DNA by Nebulization. *CSH Protoc.*, 2006(23), doi:10.1101/pdb.prot4539.

Samuels, D.C. et al., 2013. Finding the lost treasures in exome sequencing data. *Trends Genet.*, 29(10), pp.593–599.

Sandmann, S. et al., 2017. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Sci Rep.*, 7, p.43169.

Sbisà, E. et al., 1997. Mammalian mitochondrial D-loop region structural analysis: Identification of new conserved sequences and their functional and evolutionary implications. *Gene*, 205(1–2), pp.125–140.

Schmitt, M.W. et al., 2012. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci.*, 109(36), pp.14508–14513.

Shibutani, S., Takeshita, M. & Grollman, A.P., 1991. Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. *Nature*, 349(6308), pp.431–434.

Shitara H. et al. 2000. Selective and continuous elimination of mitochondria microinjected into mouse eggs from spermatids, but not from liver cells, occurs throughout embryogenesis. *Genetics*, 156(3), pp.1277–1284.

Shoffner, J.M. et al., 1990. Myoclonic epilepsy and ragged-red fiber disease (MERRF) is associated with a mitochondrial DNA tRNALys mutation. *Cell*, 61(6), pp.931–937.

Sinha, R. et al., 2017. Index Switching Causes “Spreading-Of-Signal” Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing. *bioRxiv*, doi: http://dx.doi.org/10.1101/125724

Sosa, M.X. et al., 2012. Next-Generation Sequencing of Human Mitochondrial Reference Genomes Uncovers High Heteroplasmy Frequency. *PLoS Comput Biol.*, 8(10), p.e1002737.

Stewart, J.B. et al., 2008a. Strong purifying selection in transmission of mammalian mitochondrial DNA. *PLoS Biol.*, 6(1), p.e10.

Stewart, J.B. et al., 2008b. Purifying selection of mtDNA and its implications for understanding evolution and mitochondrial disease. *Nat Rev Genet.*, 9(9), pp.657–662.

Stewart, J.B. & Larsson, N.G., 2014. Keeping mtDNA in Shape between Generations. *PLoS Genet.*, 10(10), p.e1004670.

Stewart, J.B. et al., 2015. Simultaneous DNA and RNA Mapping of Somatic Mitochondrial Mutations across Diverse Human Cancers. *PLoS Genet.*, 11(6), pp.1–15.

Stewart, J.B. & Chinnery, P.F., 2015. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat Rev Genet.*, 16(9), pp.530–542.

Takagi, M. et al., 1997. Characterization of DNA polymerase from Pyrococcus sp. strain KOD1 and its application to PCR. *Appl Environ Microbiol.*, 63(11), pp.4504–4510.

Tan, B.G. et al., 2016. Length heterogeneity at conserved sequence block 2 in human mitochondrial DNA acts as a rheostat for RNA polymerase POLRMT activity. *Nucleic Acids Res.*, 44(16), pp.7817–7829

Terzioglu, M. et al., 2013. MTERF1 Binds mtDNA to prevent transcriptional interference at the light-strand promoter but is dispensable for rRNA gene transcription regulation. *Cell Metab.*, 17(4), pp.618–626.

Travers, K.J. et al., 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.*, 38(15), p.e159.

Trifunovic, A. et al., 2004. Premature ageing in mice expressing defective mitochondrial DNA polymerase. *Nature*, 429, pp.417–423.

Tyynismaa, H. & Suomalainen, A., 2009. Mouse models of mitochondrial DNA defects and their relevance for human disease. *EMBO Rep.*, 10, pp.137–143.

Uhler, J.P. & Falkenberg, M., 2015. Primer removal during mammalian mitochondrial DNA replication. *DNA Repair*, 34, pp.28–38.

Urban, J., 2014. How does bowtie2 assign MAPQ scores? *Biofinysics*. Available at: http://biofinysics.blogspot.fi/2014/05/how-does-bowtie2-assign-mapq-scores.html [Accessed August 5, 2017].

Vellarikkal, S.K. et al., 2015. mit-o-matic: A Comprehensive Computational Pipeline for Clinical Evaluation of Mitochondrial Variations from Next-Generation Sequencing Datasets. *Hum Mutat.*, 36(4), pp.419–424.

Verbist, B.M.P. et al., 2015. VirVarSeq: a low-frequency virus variant detection pipeline for Illumina sequencing using adaptive base-calling accuracy filtering. *Bioinformatics*, 31(1), pp.94–101.

Vermulst, M. et al., 2007. Mitochondrial point mutations do not limit the natural lifespan of mice. *Nat Genet.*, 39(4), pp.540–543.

Voelkerding, K. V., Dames, S.A. & Durtschi, J.D., 2009. Next-Generation Sequencing: From Basic Research to Diagnostics. *Clin Chem.*, 55(4), pp.641–658.

Wai, T., Teoli, D. & Shoubridge, E.A., 2008. The mitochondrial DNA genetic bottleneck results from replication of a subpopulation of genomes. *Nat Genet.*, 40(12), pp.1484–1488.

Wanrooij, P.H. et al., 2012. A hybrid G-quadruplex structure formed between RNA and DNA explains the extraordinary stability of the mitochondrial R-loop. *Nucleic Acids Res.*, 40(20), pp.10334–10344.

Wanrooij, S. et al., 2012. In vivo mutagenesis reveals that OriL is essential for mitochondrial DNA replication. *EMBO Rep.*, 13(12), pp.1130–1137.

Weissensteiner, H. et al., 2016. mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. *Nucleic Acids Res.*, 44, pp.W64–W69.

Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis, *Springer-Verlag New York*. Available at: http://ggplot2.org.

Wieckowski, M.R. et al., 2009. Isolation of mitochondria-associated membranes and mitochondria from animal tissues and cells. *Nat Protoc.*, 4(11), pp.1582–1590.

Williams, S.L. et al., 2010. The mtDNA mutation spectrum of the progeroid polg mutator mouse includes abundant control region multimers. *Cell Metab.*, 12(6), pp.675–682.

Wilm, A. et al., 2012. LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.*, 40(22), pp.11189–11201.

Wingett, S., 2017. Illumina Patterned Flow Cells Generate Duplicated Sequences. *QCFail.com*. Available at: https://academhttps//sequencing.qcfail.com/articles/illumina-patterned-flow-cells-generate-duplicated-sequences/ic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw038 [Accessed July 19, 2017].

Wonnapinij, P., Chinnery, P.F. & Samuels, D.C., 2008. The Distribution of Mitochondrial DNA Heteroplasmy Due to Random Genetic Drift. *Am J Hum Genet.*, 83(5), pp.582–593.

Xu, B. & Clayton, D.A., 1996. RNA-DNA hybrid formation at the human mitochondrial heavy-strand origin ceases at replication start sites: an implication for RNA-DNA hybrids serving as primers. *EMBO J.*, 15(12), pp.3135–3143.

Xu, P. et al., 2014. Genome sequence and genetic diversity of the common carp, Cyprinus carpio. *Nat Genet.*, 46(11), pp.1212-1219.

Yao, Y.-G. et al., 2008. Pseudomitochondrial genome haunts disease studies. *J Med Genet.*, 45, pp.769–772.

Zaragoza, M. V. et al., 2010. Mitochondrial DNA Variant Discovery and Evaluation in Human Cardiomyopathies through Next-Generation Sequencing W. Li, ed. *PLoS ONE*, 5(8), p.e12295.

Zhang, P. et al., 2016. Mitochondria sequence mapping strategies and practicability of mitochondria variant detection from exome and RNA sequencing data. *Brief Bioinform.*, 17(2), pp.224–232.

Zheng, W. et al., 2006. Origins of human mitochondrial point mutations as DNA polymerase gamma-mediated errors. *Mutat Res.*, 599(1–2), pp.11–20.

APPENDIX

Appendix 1. Difference between the standard mouse mitochondrial DNA reference genome (C57Bl/6J, NC_005089.1) and NZB mitochondrial DNA. *The wild-type mouse strain used in these thesis projects (C57Bl/6N) deviates from the reference genome (C57Bl/6J) at positions 4891.T>C and 9461.T>C, and some mice lines also at position 9027.G>A. The table lists positions at which C57Bl/6N mice carrying mtDNA from NZB strain (mtNZB) deviates from the mtDNA reference genome. In total, there are 89 variants, of which one is common (position 9461.T>C) with the wild-type mouse.*

Genome position NC_005089.1	Reference base NC_005089.1	Variant base mtNZB
55	G	A
716	A	G
1353	A	G
1519	G	A
1590	G	A
1822	T	C
2201	T	C
2340	G	A
2525	C	T
2766	A	G
2767	T	C
2798	C	T
2814	T	C
2840	C	T
2934	C	T
3194	T	C
3260	A	G
3422	T	C
3467	T	C
3599	T	C
3692	A	G
3932	G	A
4123	C	T
4276	G	A
4324	T	C
4408	G	A
4706	A	G
4732	C	T
4771	T	C
4885	A	C
4903	T	G
5463	G	A
5552	T	C
5930	G	A
6041	T	C
6407	C	T
6470	A	G
6575	C	T
6620	G	A
6785	G	A
7411	A	G
7870	G	A
8439	A	G
8467	T	C
8568	C	T
8858	T	C
8864	C	T
9137	A	G
9152	T	C
9391	A	G
9461	**T**	**C**
9530	C	T
9581	C	T
9599	A	G
9985	G	A
10547	C	T
10583	A	G
10952	C	A
11843	G	A
11846	C	T
11933	A	C
12353	C	T
12575	T	A
12695	A	G
12835	T	C
12890	A	G
13004	G	A
13444	C	T
13612	T	C
13689	C	T
13781	A	G
13782	T	C
13837	A	G
13983	A	G
14186	T	C
14211	G	A
14363	A	G
14642	G	A
14738	C	T
15499	T	A
15549	C	T
15578	A	T
15588	C	T
15603	C	T
15657	T	C
15917	C	T
16017	A	C
16268	A	G
16272	T	C

Appendix 2. Reference sequence of pAM1 plasmid. *A pAM1 plasmid contains the entire mouse mtDNA (deviating from the mouse mtDNA reference genome [C57BL/6J, NC_005089.1] at positions 4794.C>T, 9348.G>A, 9461.C>T, 10918.A>G and 12048.T>C) restriction digested with HaeII (restriction site RGCGCY, at position 2603), cloned into a 2.5-kb pACYC177 plasmid backbone also restricted with HaeII.*

>pAM1 | pACYC177_mtDNA

GTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGA
AAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACT
TACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGAT
CGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGCAGCAATGGCAACAACGTT
GCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTG
CAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGC
GGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTAT
GGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCAT
ATATACTTTAGATTGATTTAAAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACC
AAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCTTAATAAGATGATCTTCTTGAGATCGTTTTGGTCT
GCGCGTAATCTCTTGCTCTGAAAACGAAAAAACCGCCTTGCAGGGCGGTTTTTCGAAGGTTCTCTGAGCTACCAACTCTT
TGAACCGAGGTAACTGGCTTGGAGGAGCGCAGTCACCAAAACTTGTCCTTTCAGTTTAGCCTTAACCGGCGCATGACTTC
AAGACTAACTCCTCTAAATCAATTACCAGTGGCTGCTGCCAGTGGTGCTTTTGCATGTCTTTCCGGGTTGGACTCAAGAC
GATAGTTACCGGATAAGGCGCAGCGGTCGGACTGAACGGGGGGTTCGTGCATACAGTCCAGCTTGGAGCGAACTGCCTAC
CCGGAACTGAGTGTCAGGCGTGGAATGAGACAAACGCGGCCATAACAGCGGAATGACACCGGTAAACCGAAAGGCAGGAA
CAGGAGAGCGCACGAGGGAGCCGCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCACTGATTT
GAGCGTCAGATTTCGTGATGCTTGTCAGGGGGGCGGAGCCTATGGAAAAACGGCTTTGCCGCGGCCCTCTCACTTCCCTG
TTAAGTATCTTCCTGGCATCTTCCAGGAAATCTCCGCCCCGTTCGTAAGCCATTTCCGCTCGCCGCAGTCGAACGACCGA
GCGTAGCGAGTCAGTGAGCGAGGAAGCGGAATATATCCTGTATCACATATTCTGCTGACGCACCGGTGCAGCCTTTTTTC
TCCTGCCACATGAAGCACTTCACTGACACCCTCATCAGTGCCAACATAGTAAGCCAGTATACACTCCGCTAGCGCTCTCA
ACTTAATTTATGAATAAAATCTAAATAAAATATATACGTACACCCTCTAACCTAGAGAAGGTTATTAGGGTGGCAGAGCC
AGGAAATTGCGTAAGACTTAAAACCTTGTTCCCAGAGGTTCAAATCCTCTCCCTAATAGTGTTCTTTATTAATATCCTAA
CACTCCTCGTCCCCATTCTAATCGCCATAGCCTTCCTAACATTAGTAGAACGCAAAATCTTAGGGTACATACAACTACGA
AAAGGCCCTAACATTGTTGGTCCATACGGCATTTTACAACCATTTGCAGACGCCATAAAATTATTTATAAAAGAACCAAT
ACGCCCTTTAACAACCTCTATATCCTTATTTATTATTGCACCTACCCTATCACTCACACTAGCATTAAGTCTATGAGTTC
CCCTACCAATACCACACCCATTAATTAATTTAAACCTAGGGATTTTATTTATTTTAGCAACATCTAGCCTATCAGTTTAC
TCCATTCTATGATCAGGATGAGCCTCAAACTCCAAATACTCACTATTCGGAGCTTTACGAGCCGTAGCCCAAACAATTTC
ATATGAAGTAACCATAGCTATTATCCTTTTATCAGTTCTATTAATAAATGGATCCTACTCTCTACAAACACTTATTACAA
CCCAAGAACACATATGATTACTTCTGCCAGCCTGACCCATAGCCATAATATGATTTATCTCAACCCTAGCAGAAACAAAC
CGGGCCCCCTTCGACCTGACAGAAGGAGAATCAGAATTAGTATCAGGGTTTAACGTAGAATACGCAGCCGGCCCATTCGC
GTTATTCTTTATAGCAGAGTACACTAACATTATTCTAATAAACGCCCTAACAACTATTATCTTCCTAGGACCCCTATACT
ATATCAATTTACCAGAACTCTACTCAACTAACTTCATAATAGAAGCTCTACTACTATCATCAACATTCCTATGGATCCGA
GCATCTTATCCACGCTTCCGTTACGATCAACTTATACATCTTCTATGAAAAAACTTTCTACCCCTAACACTAGCATTATG
TATGTGACATATTTCTTTACCAATTTTTACAGCGGGAGTACCACCATACATATAGAAATATGTCTGATAAAAGAATTACT
TTGATAGAGTAAATTATAGAGGTTCAAGCCCTCTTATTTCTAGGACAATAGGAATTGAACCTACACTTAAGAATTCAAAA
TTCTCCGTGCTACCTAAACACCTTATCCTAATAGTAAGGTCAGCTAATTAAGCTATCGGGCCCATACCCCGAAAACGTTG
GTTTAAATCCTTCCCGTACTAATAAATCCTATCACCCTTGCCATCATCTACTTCACAATCTTCTTAGGTCCTGTAATCAC
AATATCCAGCACCAACCTAATACTAATATGAGTAGGCCTGGAATTCAGCCTACTAGCAATTATCCCCATACTAATCAACA
AAAAAACCCACGATCAACTGAAGCAGCAACAAAATACTTCGTCACACAAGCAACAGCCTCAATAATTATCCTCCTGGCC
ATCGTACTCAACTATAAACAACTAGGAACATGAATATTTCAACAACAAACAAACGGTCTTATCCTTAACATAACATTAAT
AGCCCTATCCATAAAACTAGGCCTCGCCCCATTCCACTTCTGATTACCAGAAGTAACTCAAGGGATCCCACTGCACATAG
GACTTATTCTTCTTACATGACAAAAAATTGCTCCCCTATCAATTTTAATTCAAATTTACCCGCTACTCAACTCTACTATC
ATTTTAATACTAGCAATTACTTCTATTTTCATAGGGGCATGAGGAGGACTTAACCAAACACAAATACGAAAAATTATAGC
CTATTCATCAATTGCCCACATAGGATGAATATTAGCAATTCTTCCTTACAACCCATCCCTCACTCTACTCAACCTCATAA
TCTATATTATTCTTACAGCCCCTATATTCATAGCACTTATACTAAATAACTCTATAACCATCAACTCAATCTCACTTCTA
TGAAATAAAACTCCAGCAATACTAACTATAATCTCACTGATATTACTATCCCTAGGAGGCCTTCCACCACTAACAGGATT
CTTACCAAAATGAATTATCATCACAGAACTTATAAAAAACAACTGTCTAATTATAGCAACACTCATAGCAATAATAGCTC
TACTAAACCTATTCTTTTATACTCGCCTAATTTATTCCACTTCACTAACAATATTTCCAACCAACAATAACTCAAAAATA
ATAACTCACCAAACAAAAACTAAACCCAACCTAATATTTTCCACCCTAGCTATCATAAGCACAATAACCCTACCCCTAGC
CCCCCAACTAATTACCTAGAAGTTTAGGATATACTAGTCCGCGAGCCTTCAAAGCCCTAAGAAAACACACAAGTTTAACT
TCTGATAAGGACTGTAAGACTTCATCCTACATCTATTGAATGCAAATCAATTGCTTTAATTAAGCTAAGACCTCAACTAG
ATTGGCAGGAATTAAACCTACGAAAATTTAGTTAACAGCTAAATACCCTATTACTGGCTTCAATCTACTTCTACCGCCGA
AAAAAAAAAATGGCGGTAGAAGTCTTAGTAGAGATTTCTCTACACCTTCGAATTTGCAATTCGACATGAATATCACCTTA
AGACCTCTGGTAAAAAGAGGATTTAAACCTCTGTGTTTAGATTTACAGTCTAATGCTTACTCAGCCATTTTACCTATGTT

CATTAATCGTTGATTATTCTCAACCAATCACAAAGATATCGGAACCCTCTATCTACTATTCGGAGCCTGAGCGGGAATAG
TGGGTACTGCACTAAGTATTTTAATTCGAGCAGAATTAGGTCAACCAGGTGCACTTTTAGGAGATGACCAAATTTACAAT
GTTATCGTAACTGCCCATGCTTTTGTTATAATTTTCTTCATAGTAATACCAATAATAATTGGAGGCTTTGGAAACTGACT
TGTCCCACTAATAATCGGAGCCCCAGATATAGCATTCCCACGAATAAATAATATAAGTTTTTGACTCCTACCACCATCAT
TTCTCCTTCTCCTAGCATCATCAATAGTAGAAGCAGGAGCAGGAACAGGATGAACAGTCTACCCACCTCTAGCCGGAAAT
CTAGCCCATGCAGGAGCATCAGTAGACCTAACAATTTTCTCCCTTCATTTAGCTGGAGTGTCATCTATTTTAGGTGCAAT
TAATTTTATTACCACTATTATCAACATGAAACCCCCAGCCATAACACAGTATCAAACTCCACTATTTGTCTGATCCGTAC
TTATTACAGCCGTACTGCTCCTATTATCACTACCAGTGCTAGCCGCAGGCATTACTATACTACTAACAGACCGCAACCTA
AACACAACTTTCTTTGATCCCGCTGGAGGAGGGGACCCAATTCTCTACCAGCATCTGTTCTGATTCTTTGGGCACCCAGA
AGTTTATATTCTTATCCTCCCAGGATTTGGAATTATTTCACATGTAGTTACTTACTACTCCGGAAAAAAAGAACCTTTCG
GCTATATAGGAATAGTATGAGCAATAATGTCTATTGGCTTTCTAGGCTTTATTGTATGAGCCCACCACATATTCACAGTA
GGATTAGATGTAGACACACGAGCTTACTTTACATCAGCCACTATAATTATCGCAATTCCTACCGGTGTCAAAGTATTTAG
CTGACTTGCAACCCTACACGGAGGTAATATTAAATGATCTCCAGCTATACTATGAGCCTTAGGCTTTATTTTCTTATTTA
CAGTTGGTGGTCTAACCGGAATTGTTTTATCCAACTCATCCCTTGACATCGTGCTTCACGATACATACTATGTAGTAGCC
CATTTCCACTATGTTCTATCAATGGGAGCAGTGTTTGCTATCATAGCAGGATTTGTTCACTGATTCCCATTATTTTCAGG
CTTCACCCTAGATGACACATGAGCAAAAGCCCACTTCGCCATCATATTCGTAGGAGTAAACATAACATTCTTCCCTCAAC
ATTTCCTGGGCCTTTCAGGAATACCACGACGCTACTCAGACTACCCAGATGCTTACACCACATGAAACACTGTCTCTTCT
ATAGGATCATTTATTTCACTAACAGCTGTTCTCATCATGATCTTTATAATTTGAGAGGCCTTTGCTTCAAAACGAGAAGT
AATATCAGTATCGTATGCTTCAACAAATTTAGAATGACTTCATGGCTGCCCTCCACCATATCACACATTCGAGGAACCAA
CCTATGTAAAAGTAAAATAAGAAAGGAAGGAATCGAACCCCCTAAAATTGGTTTCAAGCCAATCTCATATCCTATATGTC
TTTCTCAATAAGATATTAGTAAAATCAATTACATAACTTTGTCAAAGTTAAATTATAGATCAATAATCTATATATCTTAT
ATGGCCTACCCATTCCAACTTGGTCTACAAGACGCCACATCCCCTATTATAGAAGAGCTAATAAATTTCCATGATCACAC
ACTAATAATTGTTTTCCTAATTAGCTCCTTAGTCCTCTATATCATCTCGCTAATATTAACAACAAAACTAACACATACAA
GCACAATAGATGCACAAGAAGTTGAAACCATTTGAACTATTCTACCAGCTGTAATCCTTATCATAATTGCTCTCCCCTCT
CTACGCATTCTATATATAATAGACGAAATCAACAACCCCGTATTAACCGTTAAAACCATAGGGCACCAATGATACTGAAG
CTACGAATATACTGACTATGAAGACCTATGCTTTGATTCATATATAATCCCAACAAACGACCTAAAACCTGGTGAACTAC
GACTGCTAGAAGTTGATAACCGAGTCGTTCTGCCAATAGAACTTCCAATCCGTATATTAATTTCATCTGAAGACGTCCTC
CACTCATGAGCAGTCCCCTCCCTAGGACTTAAAACTGATGCCATCCCAGGCCGACTAAATCAAGCAACAGTAACATCAAA
CCGACCAGGGTTATTCTATGGCCAATGCTCTGAAATTTGTGGATCTAACCATAGCTTTATGCCCATTGTCCTAGAAATGG
TTCCACTAAAATATTTCGAAAACTGATCTGCTTCAATAATTTAATTTCACTATGAAGCTAAGAGCGTTAACCTTTTAAGT
TAAAGTTAGAGACCTTAAAATCTCCATAGTGATATGCCACAACTAGATACATCAACATGATTTATCACAATTATCTCATC
AATAATTACCCTATTTATCTTATTTCAACTAAAAGTCTCATCACAAACATTCCCACTGGCACCTTCACCAAAATCACTAA
CAACCATAAAAGTAAAAACCCCTTGAGAATTAAAATGAACGAAAATCTATTTGCCTCATTCATTACCCCAACAATAATAG
GATTCCCAATCGTTGTAGCCATCATTATATTTCCTTCAATCCTATTCCCATCCTCAAAACGCCTAATCAACAACCGTCTC
CATTCTTTCCAACACTGACTAGTTAAACTTATTATCAAACAAATAATGCTAATCCACACACCAAAAGGACGAACATGAAC
CCTAATAATTGTTTCCCTAATCATATTTATTGGATCAACAAATCTCCTAGGCCTTTTACCACATACATTTACACCTACTA
CCCAACTATCCATAAATCTAAGTATAGCCATTCCACTATGAGCTGGAGCCGTAATTACAGGCTTCCGACACAAACTAAAA
AGCTCACTTGCCCACTTCCTTCCACAAGGAACTCCAATTTCACTAATTCCAATACTTATTATTATTGAAACAATTAGCCT
ATTTATTCAACCAATGGCATTAGCAGTCCGGCTTACAGCTAACATTACTGCAGGACACTTATTAATACACCTAATCGGAG
GAGCTACTCTAGTATTAATAAATATTAGCCCACCAACAGCTACCATTACATTTATTATTTTACTTCTACTCACAATTCTA
GAATTTGCAGTAGCATTAATTCAAGCCTACGTATTCACCCTCCTAGTAAGCCTATATCTACATGATAATACATAATGACC
CACCAAACTCATGCATATCACATAGTTAATCCAAGTCCATGACCATTAACTGGAGCCTTTTCAGCCCTCCTTCTAACATC
AGGTCTAGTAATATGATTTCACTATAATTCAATTACACTATTAACCCTTGGCCTACTCACCAATATCCTCACAATATATC
AATGATGACGAGACGTAATTCGTGAAGGAACCTACCAAGGCCACCACACTCCTATTGTACAAAAAGGACTACGATATGGT
ATAATTCTATTCATCGTCTCGGAAGTATTTTTCTTTGCAGGATTCTTCTGAGCGTTCTATCATTCTAGCCTCGTACCAAC
ACATGATCTAGGAGGCTGCTGACCTCCAACAGGAATTTCACCACTTAACCCTCTAGAAGTCCCACTACTTAATACTTCAG
TACTTCTAGCATCAGGTGTTTCAATTACATGAGCTCATCATAGCCTTATAGAAGGTAAACGAAACCACATAAATCAAGCC
CTACTAATTACCATTATACTAGGACTTTACTTCACCATCCTCCAAGCTTCAGAATACTTTGAAACATCATTCTCCATTTC
AGATGGTATCTATGGTTCTACATTCTTCATGGCTACTGGATTCCATGGACTCCATGTAATTATTGGATCAACATTCCTTA
TTGTTTGCCTACTACGACAACTAAAATTTCACTTCACATCAAAACATCACTTCGGATTTGAAGCCGCAGCATGATACTGA
CATTTTGTAGACGTAGTCTGACTTTTCCTATACGTCTCCATTTATTGATGAGGATCTTACTCCCTTAGTATAATTAATAT
AACTGACTTCCAATTAGTAGATTCTGAATAAACCCAGAAGAGAGTAATTAACCTGTACACTGTTATCTTCATTAATATTT
TATTATCCCTAACGCTAATTCTAGTTGCATTCTGACTCCCCCAAATAAATCTGTACTCAGAAAAAGCAAATCCATATGAA
TGCGGATTCGACCCTACAAGCTCTGCACGTCTACCATTCTCAATAAAATTTTTCTTGGTAGCAATTACATTTCTATTATT
TGACCTAGAAATTGCTCTTCTACTTCCACTACCATGAGCAATTCAAACAATTAAAACCTCTACTATAATAATTATAGCCT
TTATTCTAGTCACAATTCTATCTCTAGGCCTAGCATATGAATGAACACAAAAAGGATTAGAATGAACAGAGTAAATGGTA
ATTAGTTTAAAAAAAATTAATGATTTCGACTCATTAGATTATGATGATGTTCATAATTACCAATATGCCATCTACCTTCT
TCAACCTCACCATAGCCTTCTCACTATCACTTCTAGGGACACTTATATTTCGCTCTCACCTAATATCCACATTACTATGC
CTGGAAGGCATAGTATTATCCTTATTTATTATAACTTCAGTAACTTCCCTAAACTCCAACTCCATAAGCTCCATACCAAT
CCCCATCACCATCTTAGTTTTCGCAGCCTGCGAAGCAGCTGTAGGACTAGCCCTACTAGTAAAAGTTTCAAACACGTACG
GAACAGATTACGTCCAAAATCTCAACCTACTACAATGCTAAAAATTATTCTTCCCTCACTAATGCTACTACCACTAACCT
GACTATCAAGCCCTAAAAAAACCTGAACAAACGTAACCTCATATAGTTTTCTAATTAGTTTAACCAGCCTAACACTTCTA
TGACAAACCGACGAAAATTATAAAAACTTTTCAAATATATTCTCCTCAGACCCCCTATCCACACCATTAATTATTTTAAC
AGCCTGATTACTGCCACTAATATTAATAGCTAGCCAAAACCACCTAAAAAAAGATAATAACGTACTACAAAAACTCTACA
TCTCAATACTAATCAGCTTACAAATTCTCCTAATCATAACCTTTTCAGCAACTGAACTAATTATATTTTATATTTTATTT
GAAGCAACCTTAATCCCAACACTTATTATTATTACCCGATGAGGGAACCAAACTGAACGCCTAAACGCAGGGATTTATTT

CCTATTTTATACCCTAATCGGTTCTATTCCACTGCTAATTGCCCTCATCTTAATCCAAAACCATGTAGGAACCCTAAACC
TCATAATTTTATCATTCACAACACACACCTTAGACGCTTCATGATCTAACAACTTACTATGGTTGGCATGCATAATAGCA
TTTCTTATTAAAATACCATTATATGGAGTTCACCTATGACTACCAAAAGCCCATGTTGAAGCTCCAATTGCTGGGTCAAT
AATTCTAGCAGCTATTCTTCTAAAATTAGGTAGTTACGGAATAATTCGCATCTCCATTATTCTAGACCCACTAACAAAAT
ATATAGCATACCCCTTCATCCTTCTCCCTATGAGGAATAATTATAACTAGCTCAATCTGCTTACGCCAAACAGATTTA
AAATCACTAATCGCCTACTCCTCAGTTAGCCACATAGCACTTGTTATTGCATCAATCATAATCCAAACTCCATGAAGCTT
CATAGGAGCAACAATACTAATAATCGCACATGGCCTCACATCATCACTCCTATTCTGCCTAGCAAACTCCAACTACGAAC
GGATCCACAGCCGTACTATAATCATGGCCCGAGGACTTCAAATGGTCTTCCCACTTATAGCCACATGATGACTGATAGCA
AGTCTAGCTAATCTAGCTCTACCCCCTTCAATCAATCTAATAGGAGAATTATTCATTACCATATCATTATTTTCTTGATC
AAACTTTACCATTATTCTTATAGGAATTAACATTATTATTACAGGTATATACTCAATATACATAATTATTACCACCCAAC
GCGGCAAACTAACCAACCATATAATTAACCTCCAACCCTCACACACACGAGAACTAACACTAATAGCCCTTCACATAATT
CCACTTATTCTTCTAACTACCAGTCCAAAACTAATTACAGGCCTGACAATATGTGAATATAGTTTACAAAAAACATTAGA
CTGTGAATCTGACAACAGGAAATAAACCTCCTTATTCACCAAGAAAGATTGCAAGAACTGCTAATTCATGCTTCCATGTT
TAAAAACATGGCTTTCTTACTTTTATAGGATAATAGTAATCCATTGGTCTTAGGAACCAAAAACCTTGGTGCAAATCCAA
ATAAAAGTAATCAATATTTTCACAACCTCAATCTTATTAATCTTCATTCTTCTACTATCCCCAATCCTAATTTCAATATC
AAACCTAATTAAACACATCAACTTCCCACTGTACACCACCACATCAATCAAATTCTCCTTCATTATTAGCCTCTTACCCC
TATTAATATTTTTCCACAATAATATAGAATATATAATTACAACCTGGCACTGAGTCACCATAAATTCAATAGAACTTAAA
ATAAGCTTCAAAACTGACTTTTTCTCTATCCTGTTTACATCTGTAGCCCTTTTTGTCACATGATCAATTATACAATTCTC
TTCATGATATATACACTCAGACCCAAACATCAATCGATTCATTAAATATCTTACACTATTCCTGATTACCATGCTTATCC
TCACCTCAGCCAACAACATATTTCAACTTTTCATTGGCTGAGAAGGGGTGGGAATTATAACTTTCCTACTAATTGGATGA
TGGTACGGACGAACAGACGCAAATACTGCAGCCCTACAAGCAATCCTCTATAACCGCATCGGAGACATCGGATTCATTTT
AGCTATAGTTTGATTTTCCCTAAACATAAACTCATGAGAACTTCAACAGATTATATTCTCCAACAACAACGACAATCTAA
TTCCACTTATAGGCCTATTAATCGCAGCTACAGGAAAATCAGCACAATTTGGCCTCCACCCATGACTACCATCAGCAATA
GAAGGCCCTACACCAGTTTCAGCACTACTACACTCAAGTACAATAGTAGTTGCAGGAATTTTCCTACTGGTCCGATTCCA
CCCCCTCACGACTAATAATAACTTTATTTTAACAACTATACTTTGCCTCGGAGCCCTAACCACATTATTTACAGCTATTT
GTGCTCTCACCCAAAACGACATCAAAAAAATCATTGCCTTCTCTACATCAAGCCAACTAGGCCTGATAATAGTGACGCTA
GGAATAAACCAACCACACCTAGCATTCCTACACATCTGTACCCACGCATTCTTCAAAGCTATACTCTTTATATGCTCTGG
CTCAATCATTCATAGCCTGGCAGACGAACAAGACATCCGAAAAATAGGAAACATCACAAAAATCATACCATTCACATCAT
CATGCCTAGTAATCGGAAGCCTCGCCCTCACAGGAATACCATTCCTAACAGGGTTCTACTCAAAAGACCTAATTATTGAA
GCAATTAATACCTGCAACACCAACGCCTGAGCCCTACTAATTACACTAATCGCCACTTCTATAACAGCTATGTACAGCAT
ACGAATCATTTACTTCGTAACAATAACAAAACCGCGTTTTCCCCCCCTAATCTCCATTAACGAAAATGACCCAGACCTCA
TAAACCCAATCAAACGCCTAGCATTCGGAAGCATCTTTGCAGGATTTGTCATCTCATATAATATTCCACCAACCAGCATT
CCAGTCCTCACAATACCATGATTTTAAAAACCACAGCCCTAATTATTTCAGTATTAGGATTCCTAATCGCACTAGAACT
AAACAACCTAACCATAAAACTATCAATAAATAAAGCAAATCCATATTCATCCTTCTCAACTTTACTGGGGTTTTTCCCAT
CTATTATTCACCGCATTACACCCATAAAATCTCTCAACCTAAGCCTAAAAACATCCCTAACTCTCCTAGACTTGATCTGG
TTAGAAAAAACCATCCCAAAATCCACCTCAACTCTTCACACAAACATAACCACTTTAACAACCAACCAAAAAGGCTTAAT
TAAATTGTACTTTATATCATTCCTAATTAACATCATCTTAATTATTATCTTATACTCAATTAATCTCGAGTAATCTCGAT
AATAATAAAAATACCCGCAAACAAAGATCACCCAGCTACTACCATCATTCAAGTAGCACAACTATATATTGCCGCTACCC
CAATCCCTCCTTCCAACATAACTCCAACATCATCAACCTCATACATCAACCAATCTCCCAAACCATCAAGATTAATTACT
CCAACTTCATCATAATAATTAAGCACACAAATTAAAAAAACCTCTATAATCACCCCCAATACTAAAAAACCCAAAATTAA
TCAGTTAGATCCCCAAGTCTCTGGATATTCCTCAGTAGCTATAGCAGTCGTATATCCAAACACAACCAACATCCCCCCTA
AATAAATTAAAAAAACTATTAAACCTAAAAAACGATCCACCAAACCCTAAAACCATTAAAACAACCAACAAACCCACTAACA
ATTAAACCTAAACCTCCATAAATAGGTGAAGGCTTTAATGCTAACCCAAGACAACCAACCAAAAATAATGAACTTAAAAC
AAAAATATAATTATTCATTATTTCTACACAGCATTCAACTGCGACCAATGACATGAAAAATCATCGTTGTAATTCAACTA
CAGAAACACCTAATGACAAACATACGAAAAACACACCCATTATTTAAAATTATTAACCACTCATTCATTGACCTACCTGC
CCCATCCAACATTTCATCATGATGAAACTTTGGGTCCCTTCTAGGAGTCTGCCTAATAGTCCAAATCATTACAGGTCTTT
TCTTAGCCATACACTACACATCAGATACAATAACAGCCTTTTCATCAGTAACACACATTTGTCGAGACGTAAATTACGGG
TGACTAATCCGATATATACACGCAAACGGAGCCTCAATATTTTTTATTTGCTTATTCCTTCATGTCGGACGAGGCTTATA
TTATGGATCATATACATTTATAGAAACCTGAAACATTGGAGTACTTCTACTGTTCGCAGTCATAGCCACAGCATTTATAG
GCTACGTCCTTCCATGAGGACAAATATCATTCTGAGGTGCCACAGTTATTACAAACCTCCTATCAGCCATCCCATATATT
GGAACAACCCTAGTCGAATGAATTTGAGGGGGCTTCTCAGTAGACAAAGCCACCTTGACCCGATTCTTCGCTTTCCACTT
CATCTTACCATTTATTATCGCGGCCCTAGCAATCGTTCACCTCCTCTTCCTCCACGAAACAGGATCAAACAACCCAACAG
GATTAAACTCAGATGCAGATAAAATTCCATTTCACCCCTACTATACAATCAAAGATATCCTAGGTATCCTAATCATATTC
TTAATTCTCATAACCCTAGTATTATTTTTCCCAGACATACTAGGAGACCCAGACAACTACATACCAGCTAATCCACTAAA
CACCCCACCCCATATTAAACCCGAATGATATTTCCTATTTGCATACGCCATTCTACGCTCAATCCCCAATAAACTAGGAG
GTGTCCTAGCCTTAATCTTATCTATCCTAATTTTAGCCCTAATACCTTTCCTTCATACCTCAAAGCAACGAAGCCTAATA
TTCCGCCCAATCACACAAATTTTGTACTGAATCCTAGTAGCCAACCTACTTATCTTAACCTGAATTGGGGGCCAACCAGT
AGAACACCCATTTATTATCATTGGCCAACTAGCCCTCCATCTCATACTTCTCAATCATCTTAATTCTTATACCAATCTCAG
GAATTATCGAAGACAAAATACTAAAATTATATCCATGTCTTGATAGTATAAACATTACTCTGGTCTTGTAAACCTGAAAT
GAAGATCTTCTCTTCTCAAGACATCAAGAAGAAGGAGCTACTCCCCACCACCAGCACCCAAAGCTGGTATTCTAATTAAA
CTACTTCTTGAGTACATAAATTTACATAGTACAACAGTACATTTATGTATATCGTACATTAAACTATTTTCCCCAAGCAT
ATAAGCTAGTACATTAAATCAATGGTTCAGGTCATAAAATAATCATCAACATAAATCAATATATATACCATGAATATTAT
CTTAAACACATTAAACTAATGTTATAAGGACATATCTGTGTTATCTGACATACACCATACAGTCATAAACTCTTCTCTTC
CATATGACTATCCCCTTCCCCATTTGGTCTATTAATCTACCATCCTCCGTGAAACCAACAACCCGCCCACCAATGCCCCT
CTTCTCGCTCCGGGCCCATTAAACTTGGGGGTAGCTAAACTGAAACCTTATCAGACATCTGGTTCTTACTTCAGGGCCAT
CAAATGCGTTATCGCCCATACGTTCCCCTTAAATAAGACATCTCGATGGTATCGGGTCTAATCAGCCCATGACCAACATA

ACTGTGGTGTCATGCATTTGGTATCTTTTTATTTTGGCCTACTTTCATCAACATAGCCGTCAAGGCATGAAAGGACAGCA
CACAGTCTAGACGCACCTACGGTGAAGAATCATTAGTCCGCAAAACCCAATCACCTAAGGCTAATTATTCATGCTTGTTA
GACATAAATGCTACTCAATACCAAATTTTAACTCTCCAAACCCCCCACCCCCTCCTCTTAATGCCAAACCCCAAAAACAC
TAAGAACTTGAAAGACATATAATATTAACTATCAAACCCTATGTCCTGATCAATTCTAGTAGTTCCCAAAATATGACTTA
TATTTTAGTACTTGTAAAAATTTTACAAAATCATGTTCCGTGAACCAAAACTCTAATCATACTCTATTACGCAATAAACA
TTAACAAGTTAATGTAGCTTAATAACAAAGCAAAGCACTGAAAATGCTTAGATGGATAATTGTATCCCATAAACACAAAG
GTTTGGTCCTGGCCTTATAATTAATTAGAGGTAAAATTACACATGCAAACCTCCATAGACCGGTGTAAAATCCCTTAAAC
ATTTACTTAAAATTTAAGGAGAGGGTATCAAGCACATTAAAATAGCTTAAGACACCTTGCCTAGCCACACCCCCACGGGA
CTCAGCAGTGATAAATATTAAGCAATAAACGAAAGTTTGACTAAGTTATACCTCTTAGGGTTGGTAAATTTCGTGCCAGC
CACCGCGGTCATACGATTAACCCAAACTAATTATCTTCGGCGTAAAACGTGTCAACTATAAATAAATAAATAGAATTAAA
ATCCAACTTATATGTGAAAATTCATTGTTAGGACCTAAACTCAATAACGAAAGTAATTCTAGTCATTTATAATACACGAC
AGCTAAGACCCAAACTGGGATTAGATACCCCACTATGCTTAGCCATAAACCTAAATAATTAAATTTAACAAAACTATTTG
CCAGAGAACTACTAGCCATAGCTTAAAACTCAAAGGACTTGGCGGTACTTTATATCCATCTAGAGGAGCCTGTTCTATAA
TCGATAAACCCCGCTCTACCTCACCATCTCTTGCTAATTCAGCCTATATACCGCCATCTTCAGCAAACCCTAAAAAGGTA
TTAAAGTAAGCAAAAGAATCAAACATAAAAACGTTAGGTCAAGGTGTAGCCAATGAAATGGGAAGAAATGGGCTACATTT
TCTTATAAAAGAACATTACTATACCCTTTATGAAACTAAAGGACTAAGGAGGATTTAGTAGTAAATTAAGAATAGAGAGC
TTAATTGAATTGAGCAATGAAGTACGCACACACCGCCCGTCACCCTCCTCAAATTAAATTAAACTTAACATAATTAATTT
CTAGACATCCGTTTATGAGAGGAGATAAGTCGTAACAAGGTAAGCATACTGGAAAGTGTGCTTGGAATAATCATAGTGTA
GCTTAATATTAAAGCATCTGGCCTACACCCAGAAGATTTCATGACCAATGAACACTCTGAACTAATCCTAGCCCTAGCCC
TACACAAATATAATTATACTATTATATAAATCAAAACATTTATCCTACTAAAAGTATTGGAGAAAGAAATTCGTACATCT
AGGAGCTATAGAACTAGTACCGCAAGGGAAAGATGAAAGACTAATTAAAAGTAAGAACAAGCAAAGATTAAACCTTGTAC
CTTTTGCATAATGAACTAACTAGAAAACTTCTAACTAAAAGAATTACAGCTAGAAACCCCGAAACCAAACGAGCTACCTA
AAAACAATTTTATGAATCAACTCGTCTATGTGGCAAAATAGTGAGAAGATTTTTAGGTAGAGGTGAAAAGCCTAACGAGC
TTGGTGATAGCTGGTTACCCAAAAAATGAATTTAAGTTCAATTTTAAACTTGCTAAAAAAACAACAAAATCAAAAAGTAA
GTTTAGATTATAGCCAAAAGAGGGACAGCTCTTCTGGAACGGAAAAAACCTTTAATAGTGAATAATTAACAAAACAGCTT
TTAACCATTGTAGGCCTAAAAGCAGCCACCAATAAAGAAAGCGTTCAAGCTCAACATAAAATTTCAATTAATTCCATAAT
TTACACCAACTTCCTAAACTTAAAATTGGGTTAATCTATAACTTTATAGATGCAACACTGTTAGTATGAGTAACAAGAAT
TCCAATTCTCCAGGCATACGCGTATAACAACTCGGATAACCATTGTTAGTTAATCAGACTATAGGCAATAATCACACTAT
AAATAATCCACCTATAACTTCTCTGTTAACCCAACACCGGAATGCCTAAAGGAAAGATCCAAAAAGATAAAAGGAACTCG
GCAAACAAGAACCCCGCCTGTTTACCAAAAACATCACCTCTAGCATTACAAGTATTAGAGGCACTGCCTGCCCAGTGACT
AAAGTTTAACGGCCGCGGTATCCTGACCGTGCAAAGGTAGCATAATCACTTGTTCCTTAATTAGGGACTAGCATGAACGG
CTAAACGAGGGTCCAACTGTCTCTTATCTTTAATCAGTGAAATTGACCTTTCAGTGAAGAGGCTGAAATATAATAATAAG
ACGAGAAGACCCTATGGAGCTTAAATTATATAACTTATCTATTTAATTTATTAAACCTAATGGCCCAAAAACTATAGTAT
AAGTTTGAAATTTCGGTTGGGGTGACCTCGGAGAATAAAAAATCCTCCGAATGATTATAACCTAGACTTACAAGTCAAAG
TAAAATCAACATATCTTATTGACCCAGATATATTTTGATCAACGGACCAAGTTACCCTAGGGATAACAGCGCAATCCTAT
TTAAGAGTTCATATCGACAATTAGGGTTTACGACCTCGATGTTGGATCAGGACATCCCAATGGTGTAGAAGCTATTAATG
GTTCGTTTGTTCAACGATTAAAGTCCTACGTGATCTGAGTTCAGACCGGAGCAATCCAGGTCGGTTTCTATCTATTTACG
ATTTCTCCCAGTACGAAAGGACAAGAGAAATAGAGCCACCTTACAAATAAGCGCTCAAAGATGCAGGGGTAAAAGCTAAC
CGCATCTTTACCGACAAGGCATCCGGCAGTTCAACAGATCGGGAAGGGCTGGATTTGCTGAGGATGAAGGTGGAGGAAGG
TGATGTCATTCTGGTGAAGAAGCTCGACCGTCTTGGCCGCGACACCGCCGACATGATCCAACTGATAAAAGAGTTTGATG
CTCAGGGTGTAGCGGTTCGGTTTATTGACGACGGGATCAGTACCGACGGTGATATGGGGCAAATGGTGGTCACCATCCTG
TCGGCTGTGGCACAGGCTGAACGCCGGAGGATCCTAGAGCGCACGAATGAGGGCCGACAGGAAGCAAAGCTGAAAGGAAT
CAAATTTGGCCGCAGGCGTACCGTGGACAGGAACGTCGTGCTGACGCTTCATCAGAAGGGCACTGGTGCAACGGAAATTG
CTCATCAGCTCAGTATTGCCCGCTCCACGGTTTATAAAATTCTTGAAGACGAAAGGGCCTCGTGATACGCCTATTTTTAT
AGGTTAATGTCATGATAATAATGGTTTCTTAGACGTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGT
TTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATTGAAAAAG
GAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACC
CAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGC
GGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGT
ATTATCCCGT

Appendix 3. Exact commands for the mtDNA-seq data analysis steps.

```
############################## mtDNA-seq data analysis workflow ##############################
#### TRIMMING ####
# Read trimming for minimum length, quality and TruSeq adapters
flexbar -n 16 -r /path/to/reads1.fastq.gz -t output_prefix -f i1.8 -j -z GZ -q
28 -m 50 -ao 10 -at 1 -ae ANY > file.log.txt
#########################################################################################

#### ALIGNMENT to the NORMAL REFERENCE GENOME ####
# Create index for the reference genome (required only once)
bwa index -p reference reference.fa
# bwa mem alignment
bwa mem -t 15 -P -T 19 -B 3 -L 5,4 /path/to/reference/reference
/path/to/reads1_flexbar.fastq.gz > reads1.sam
# Sorting and indexing
samtools view -Sbu reads1.sam | samtools sort - -T reads1.sorted -o
reads1.sorted.bam ; samtools index reads1.sorted.bam
# Quality filter only mapped reads for further processing
samtools view -bq 1 reads1.sorted.bam > reads1.accepted.bam
samtools index reads1.accepted.bam
#########################################################################################

#### ALIGNMENT to the SPLIT REFERENCE GENOME ####
## Create the split reference genome (required only once) ##
# Create variables
split_genome=reference_split.fa
single_line=reference_singleline.fa
split_data=reference_split.data
# Transform the reference fasta file to contain the sequence as a single line
header=$(cat /path/to/reference.fa | grep '>')
cat /path/to/reference.fa | grep -v '>' | tr -d '\n' > $single_line
# Calculate the length of the input reference genome sequence to determine the cutting position
half_split=$(cat $single_line | awk 'BEGIN {junc=0} junc=int(length($0)/2)
{print junc}')
half_split1=$(echo $half_split | awk '{print $0+1}')
full_len=$(cat $single_line | awk 'BEGIN {len=0} len=length($0) {print len}')
paste <(echo $half_split) <(echo $half_split1) <(echo $full_len) > $split_data
# Take the genome halves according to the calculated positions
gen_start=$(cat $single_line | cut -c1-$half_split)
gen_end=$(cat $single_line | cut -c${half_split1}-$full_len)
# Combine the halves in correct order and restore the fasta format
gen_split=$(paste <(echo $gen_end) <(echo $gen_start) | tr -d '\t')
paste <(echo $header) <(echo $gen_split) | tr '\t' '\n' > $split_genome
####
# Create index for the reference genome (required only once)
bwa index -p reference_split reference_split.fa
# bwa mem alignment
bwa mem -t 15 -P -T 19 -B 3 -L 5,4 /path/to/reference/reference_split
/path/to/reads1_flexbar.fastq.gz > reads1_junction.sam
# Sorting and indexing
samtools view -Sbu reads1_junction.sam | samtools sort - -T
reads1_junction.sorted -o reads1_junction.sorted.bam
```

```
samtools index reads1_junction.sorted.bam
##################################################################################################
#### FILTERING the ALIGNED READS ### #
# Quality filter only uniquely aligned reads for further processing (bwa mem mapping quality 0 indicates multimapping reads)
samtools view -bq 1 reads1.sorted.bam > reads1.accepted.bam
samtools index reads1.accepted.bam

samtools view -bq 1 reads1_junction.sorted.bam > reads1_junction.accepted.bam
samtools index reads1_junction.accepted.bam
##################################################################################################

#### CALCULATE the COVERAGE ####
# For the normal reference alignment
bedtools genomecov -d -ibam reads1.accepted.bam -g reference.fa > coverage.txt
# For the split reference alignment
bedtools genomecov -d -ibam reads1_junction.accepted.bam -g reference_split.fa > coverage_junction.txt
## Combine the coverages to represent the entire mtDNA genome ##
# Take the middle and end points of the split junction genome for extracting correct lines
mid_point=$(cat reference_split.data | awk '{print $2}')
end_point=$(cat reference_split.data | awk '{print $3}')
# Middle part of the normal coverage file
cat coverage.txt | awk -v endpoint="$end_point" '$2 > 200 && $2 < endpoint - 200 {print}' > coverage.middle
# Re-coordinate the junction region and take only -200 and +200
cat coverage_junction.txt | awk -v midpoint="$mid_point" -v endpoint="$end_point" 'BEGIN {OFS = "\t"; pos = 0; test = 0; res = 0} {pos = $2; test = pos - midpoint; if(test <= 0) res = endpoint + test; else res = test; $2 = res; print}' | sort -nk2 | awk -v endpoint="$end_point" '$2 <=200 || $2 >= endpoint - 200 {print}' > coverage.junction_replacement
# Merge the middle and junction regions into a final result file
cat $normal_middle $junction_replacement | sort -nk2 > coverage_final.txt
##################################################################################################

####LoFreq* VARIANT CALLING ####
# First set the indel qualities for using --call-indels
lofreq indelqual --dindel --ref reference.fa --out reads1.indelqual.bam reads1.accepted.bam
lofreq indelqual --dindel --ref reference_split.fa --out reads1_junction.indelqual.bam reads1_junction.accepted.bam
samtools index reads1.indelqual.bam
samtools index reads1_junction.indelqual.bam
# Variant calling including indels
lofreq call-parallel --pp-threads 20 -f reference.fa -o reads1.nofilter.vcf -N -B -q 30 -Q 30 --call-indels --no-default-filter reads1.indelqual.bam
lofreq call-parallel --pp-threads 20 -f reference_split.fa -o reads1_junction.nofilter.vcf -N -B -q 30 -Q 30 --call-indels --no-default-filter reads1_junction.indelqual.bam
# Filtering the results if >85% of variant reads are on single strand
lofreq filter --no-defaults --snvqual-thresh 70 --indelqual-thresh 70 --sb-incl-indels -B 60 -i reads1.nofilter.vcf -o reads1.sbfiltered.vcf
lofreq filter --no-defaults --snvqual-thresh 70 --indelqual-thresh 70 --sb-incl-indels -B 60 -i reads1_junction.nofilter.vcf -o reads1_junction.sbfiltered.vcf
```

Combine the original and junction region results for the junction region
Read in the mid and end points of the split genome junction region
```
mid_point=$(cat reference_split.data | awk '{print $2}')
end_point=$(cat reference_split.data | awk '{print $3}')
```
Re-coordinate the variants and take only -200 and +200 region
```
cat reads1_junction.sbfiltered.vcf | awk '/#CHROM/ {flag = 1; next} flag
{print}' | awk -v midpoint="$mid_point" -v endpoint="$end_point" 'BEGIN {OFS =
"\t"; pos = 0; test = 0; res = 0} {pos = $2; test = pos - midpoint; if(test <=
0) res = endpoint + test; else res = test; $2 = res; print}' | sort -nk2 | awk
-v endpoint="$end_point" '$2 <= 200 || $2 >= endpoint - 200 {print}' >
junction_replacement.vars
```
Intermediate vcf header
```
cat reads1.sbfiltered.vcf | awk '/#/ {print}' >
reads1.sbfiltered_junction_combined.vcf
```
Take middle part of the original alignment variant calls
```
cat reads1.sbfiltered.vcf | awk '/#CHROM/ {flag = 1; next} flag {print}' | awk
-v endpoint="$end_point" '$2 > 200 && $2 < endpoint - 200 {print}' >
reads1.middle.vars
```
Combine junction replacement and original middle, sort and append to the original vcf header
```
cat reads1.middle.vars junction_replacement.vars | sort -nk2 >>
reads1.sbfiltered_junction_combined.vcf
```
Final junction fixed file for snv and indel separation
```
# Separate snvs only and indels only to different files for downstream analysis
lofreq filter --no-defaults --only-snvs -i
reads1.sbfiltered_junction_combined.vcf -o reads1.snvs.vcf
lofreq filter --no-defaults --only-indels -i
reads1.sbfiltered_junction_combined.vcf -o reads1.indels.vcf
```
##

snpEff FINAL FILTERING of the VARIANT RESULTS
Make sure the snpEff config file uses mitochondrial codons for annotations (required only once)
```
cd ~
mkdir snpEff_data
```
Modify snpEff.config file
data.dir = ~/snpEff_data/
Fix mitochondrial codon table usage in the .config file to the wanted genome version:
Original rows in snpEff.config
GRCm38.82.genome : Mus_musculus
GRCm38.82.reference : ftp://ftp.ensembl.org/pub/release-82/gtf/
Add MT.codonTable in between:
GRCm38.82.genome : Mus_musculus
GRCm38.82.MT.codonTable: Vertebrate_Mitochondrial
GRCm38.82.reference : ftp://ftp.ensembl.org/pub/release-82/gtf/
###
Filter variants for minimum number of supporting reads in total and on both strands
```
cat reads1.snvs.vcf | java -jar /software/snpEff/4.2/SnpSift.jar filter
"( (DP*AF >= 15) & (DP4[2] >= 3) & (DP4[3] >= 3) )" > reads1.filtered.vcf
```
Filter variants for mouse strain specific and highly strand biased variants
```
cat reads1.filtered.vcf | java -jar /software/snpEff/4.2/SnpSift.jar filter
"( (SB < 1000) & (POS != 4891) & (POS != 9461) & (POS != 9027) )" >
reads1.pos_filtered.vcf
```
Annotate the file
```
java -jar /software/snpEff/4.2/snpEff.jar -config ~/snpEff.config -no-downstream
-no-upstream -noStats -v GRCm38.82 reads1.pos_filtered.vcf > reads1.ann.vcf
```

-classic produces reference and alternative codon triplets and different amino acid change format than newer ANN (default) annotation

```
java -jar /software/snpEff/4.2/snpEff.jar eff -config ~/snpEff.config -classic
-no-downstream -no-upstream -noStats -v GRCm38.82 reads1.pos_filtered.vcf >
reads1.eff.vcf
```

Separate genome regions

tRNAs (contain also protein_coding rows from overlapping region)

```
cat reads1.ann.vcf | java -jar /software/snpEff/4.2/SnpSift.jar filter
"ANN[*].BIOTYPE has 'Mt_tRNA'" |
/software/snpEff/4.2/scripts/vcfEffOnePerLine.pl > reads1.ann.trna_tmp.vcf
```

Re-filter to have only tRNA rows and modify the final output columns

```
cat reads1.ann.trna_tmp.vcf | java -jar /software/snpEff/4.2/SnpSift.jar filter
"ANN[*].BIOTYPE has 'Mt_tRNA'" | java -jar /software/snpEff/4.2/SnpSift.jar
extractFields - -s "\t" -e "NA" CHROM "ANN[*].BIOTYPE" "ANN[*].GENE" POS REF ALT
QUAL DP DP4[0] DP4[1] DP4[2] DP4[3] AF SB "ANN[*].EFFECT" "ANN[*].IMPACT"
"ANN[*].HGVS_C" "ANN[*].HGVS_P" "ANN[*].CDS_POS" "ANN[*].CDS_LEN"
"ANN[*].AA_POS" "ANN[*].AA_LEN" "ANN[*].GENEID" "ANN[*].ERRORS" | sed '1d' | sed
'1iCHROM\tANN_BIOTYPE\tANN_GENE\tPOS\tREF\tALT\tQUAL\tDP\tDP4_REF_fw\tDP4_REF_rv
\tDP4_ALT_fw\tDP4_ALT_rv\tAF\tSB\tANN_EFFECT\tANN_IMPACT\tANN_HGVSc\tANN_HGVSp\t
ANN_CDSPOS\tANN_CDSLEN\tANN_AAPOS\tANN_AALEN\tANN_GENEID\tANN_ERRORS' >
reads1.ann.trna.vcf
```

rRNAs

```
cat reads1.ann.vcf | java -jar /software/snpEff/4.2/SnpSift.jar filter
"ANN[*].BIOTYPE has 'Mt_rRNA'" | java -jar /software/snpEff/4.2/SnpSift.jar
extractFields - -s "\t" -e "NA" CHROM ANN[0].BIOTYPE ANN[0].GENE POS REF ALT
QUAL DP DP4[0] DP4[1] DP4[2] DP4[3] AF SB ANN[0].EFFECT ANN[0].IMPACT
ANN[0].HGVS_C ANN[0].HGVS_P ANN[0].CDS_POS ANN[0].CDS_LEN ANN[0].AA_POS
ANN[0].AA_LEN ANN[0].GENEID ANN[0].ERRORS | sed '1d' | sed
'1iCHROM\tANN_BIOTYPE\tANN_GENE\tPOS\tREF\tALT\tQUAL\tDP\tDP4_REF_fw\tDP4_REF_rv
\tDP4_ALT_fw\tDP4_ALT_rv\tAF\tSB\tANN_EFFECT\tANN_IMPACT\tANN_HGVSc\tANN_HGVSp\t
ANN_CDSPOS\tANN_CDSLEN\tANN_AAPOS\tANN_AALEN\tANN_GENEID\tANN_ERRORS' >
reads1.ann.rrna.vcf
```

Control region

```
cat reads1.ann.vcf | java -jar /software/snpEff/4.2/SnpSift.jar intervals
~/scripts_master/ctrl_region_bases.txt | java -jar
/software/snpEff/4.2/SnpSift.jar extractFields - -s "\t" -e "NA" CHROM
ANN[0].BIOTYPE ANN[0].GENE POS REF ALT QUAL DP DP4[0] DP4[1] DP4[2] DP4[3] AF SB
ANN[0].EFFECT ANN[0].IMPACT ANN[0].HGVS_C ANN[0].HGVS_P ANN[0].CDS_POS
ANN[0].CDS_LEN ANN[0].AA_POS ANN[0].AA_LEN ANN[0].GENEID ANN[0].ERRORS | sed
'1d' | sed
'1iCHROM\tANN_BIOTYPE\tANN_GENE\tPOS\tREF\tALT\tQUAL\tDP\tDP4_REF_fw\tDP4_REF_rv
\tDP4_ALT_fw\tDP4_ALT_rv\tAF\tSB\tANN_EFFECT\tANN_IMPACT\tANN_HGVSc\tANN_HGVSp\t
ANN_CDSPOS\tANN_CDSLEN\tANN_AAPOS\tANN_AALEN\tANN_GENEID\tANN_ERRORS' >
reads1.ann.ctrl.vcf
```

OriL region

```
cat reads1.ann.vcf | java -jar /software/snpEff/4.2/SnpSift.jar intervals
~/scripts_master/OriL_region_bases.txt | java -jar
/software/snpEff/4.2/SnpSift.jar extractFields - -s "\t" -e "NA" CHROM
ANN[0].BIOTYPE ANN[0].GENE POS REF ALT QUAL DP DP4[0] DP4[1] DP4[2] DP4[3] AF SB
ANN[0].EFFECT ANN[0].IMPACT ANN[0].HGVS_C ANN[0].HGVS_P ANN[0].CDS_POS
ANN[0].CDS_LEN ANN[0].AA_POS ANN[0].AA_LEN ANN[0].GENEID ANN[0].ERRORS | sed
'1d' | sed
'1iCHROM\tANN_BIOTYPE\tANN_GENE\tPOS\tREF\tALT\tQUAL\tDP\tDP4_REF_fw\tDP4_REF_rv
```

```
\tDP4_ALT_fw\tDP4_ALT_rv\tAF\tSB\tANN_EFFECT\tANN_IMPACT\tANN_HGVSc\tANN_HGVSp\t
ANN_CDSPOS\tANN_CDSLEN\tANN_AAPOS\tANN_AALEN\tANN_GENEID\tANN_ERRORS' >
reads1.ann.oril.vcf
# Protein_coding
cat reads1.ann.vcf | java -jar /software/snpEff/4.2/SnpSift.jar filter
"ANN[*].BIOTYPE has 'protein_coding'" | java -jar
/software/snpEff/4.2/SnpSift.jar extractFields - -s "\t" -e "NA" CHROM
ANN[0].BIOTYPE ANN[0].GENE POS REF ALT QUAL DP DP4[0] DP4[1] DP4[2] DP4[3] AF SB
ANN[0].EFFECT ANN[0].IMPACT ANN[0].HGVS_C ANN[0].HGVS_P ANN[0].CDS_POS
ANN[0].CDS_LEN ANN[0].AA_POS ANN[0].AA_LEN ANN[0].GENEID ANN[0].ERRORS | sed
'1d' | sed
'1iCHROM\tANN_BIOTYPE\tANN_GENE\tPOS\tREF\tALT\tQUAL\tDP\tDP4_REF_fw\tDP4_REF_rv
\tDP4_ALT_fw\tDP4_ALT_rv\tAF\tSB\tANN_EFFECT\tANN_IMPACT\tANN_HGVSc\tANN_HGVSp\t
ANN_CDSPOS\tANN_CDSLEN\tANN_AAPOS\tANN_AALEN\tANN_GENEID\tANN_ERRORS' >
reads1.ann.protein_tmp.vcf

# Add codon position as a number to protein variants
# Calculate whether the variant is on 1st, 2nd or 3rd codon position for coding genes based on ANN_CDSpos column
sed '1d' reads1.ann.protein_tmp.vcf | cut -f 19 | while read i; do
 if [ -z $i ]; then
  echo "NA"
 elif [ $(($i % 3)) == 1 ]; then
  echo "1"
 elif [ $(($i % 3)) == 2 ]; then
  echo "2"
 elif [ $(($i % 3)) == 0 ]; then
  echo "3"
 fi
done | sed '1iCODON_POS' > codon_position.tmp
# Codons and amino acid change from EFF annotation for protein coding
cat reads1.eff.vcf | java -jar /software/snpEff/4.2/SnpSift.jar filter
"ANN[*].BIOTYPE has 'protein_coding'" | java -jar
/software/snpEff/4.2/SnpSift.jar extractFields - -s "\t" -e "NA" "EFF[0].CODON"
"EFF[0].AA" | sed '/^#/d;s/\//\t/g' | sed
'1iEFF_CODONREF\tEFF_CODONALT\tEFF_HGVSp' > protein.eff.tmp
# Combine everything to full protein coding tab-delimited file
paste reads1.ann.protein_tmp.vcf codon_position.tmp protein.eff.tmp >
reads1.ann.protein.vcf

# Combine all per element files into a full genome tab-delimited file to be used for mutation load calculations
sed -e '2,${/^CHROM/d' -e '}' reads1.ann.*.txt | sed '/^CHROM/ s/
$/\tCODON_POS\tEFF_CODONREF\tEFF_CODONALT\tEFF_HGVSp/' > reads1.ann.MT.vcf
##########################################################################################
```

ACKNOWLEDGEMENTS

This thesis work was carried out at Max Planck Institute for Biology of Ageing (Cologne, Germany) and as a part of Graduate School for Biological Sciences, University of Cologne (Germany). Max Planck Society (Germany) and Emil Aaltonen Foundation (Finland) are acknowledged for the financial support.

I am deeply thankful for my supervisor PhD Jim Stewart for having the faith in me and for offering me the PhD position although I had no prior experience about mitochondria or sequencing. During these years, I have learnt from you a great deal about the fascinating world of science. And most importantly, you provided the opportunity to focus on bioinformatics and sequencing data analysis, which have truly changed the course of my career and life – four years ago I could have not imagined that one day my professional title will be 'Bioinformatician'.

Dr. Dario Valenzano is sincerely acknowledged for kindly agreeing to be my official supervisor and a committee member as well as for reviewing my thesis. I also want to thank Prof. Dr. Andreas Beyer for participating to the committee and for reviewing my thesis. Prof. Dr. Jan Riemer is thanked for being the chair of the committee.

Learning the bioinformatics would not have been possible without the amazing people at the Bioinformatics Core Facility of the MPI-AGE. My warmest thanks to Dr. Anže Lošdorfer Božič, Dr. Jorge Bouças, Sven E. Templer, Franziska Metge and Dr. Tobias Jakobi – without your endless support, incredible kindness and patience (and all the other positive adjectives) as well as occasional dosages of beer or Kraken, I would have been lost and lonely with my work.

I am most grateful for Johanna and Timo Kauppila for their unconditional friendship over these years. Your contributions to both, personal and scientific matters have been priceless and I would have not been able to make it through the PhD or life in Germany without you two. Thanks to all previous and current Stewart Lab members, especially to Marie-Lune and Sara (for the crucial Hans im Glück group support sessions ;). I also want to thank everyone in Prof. Dr. Nils-Göran Larsson's lab for all the help and guidance with the lab work, especially thanks to Maria del Pilar Miranda for always being so kind to me and for Jakob Busch for making me laugh even at the most annoying drawbacks.

Finally, a huge special thanks to the friends outside academia or Germany. Jan, Karen, Steffi, Yoli, Leandro, Teppo, Krista, Riina, Elsa, Elina, Kaisa and Hanna – thank you for the countless hours of ”kitchen therapy” and for your invaluable support in general in life, beer/wine drinking, programming, science or simply fooling around. Thanks to Simo for caring and being there for me always when needed. Thanks also for contributing to my work by teaching me crucial lessons about programming practices or usage of graphics software.

Thank you all for guiding me through the most difficult times
as well as sharing all the happy moments with me –
I have learnt a lot from all of you!

ERKLÄRUNG

Ich versichere, dass ich die von mir vorgelegte Dissertation selbstständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde.

Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Dr. Dario Valenzano betreut worden.

Helsinki, October 2017

_____

Marita Isokallio

LEBENSLAUF

Marita Annika Isokallio

PERSÖNLICHE DATEN

Geburtsdatum:	11.02.1985
Geburtsort:	Huittinen, Finnland
Nationalität:	Finnisch
Adresse:	Kauppalantie 16 A 9, 00320 Helsinki, Finnland
Email:	marita.isokallio@gmail.com

AUSBILDUNG

Jan. 2014 – Dez. 2017
Doktorandin am Max Planck Insitute für Biologie des Alterns, Graduate School for Biological Sciences (GSfBS), Institut für Genetik, Department für Biologie, Universität zu Köln, Deutschland, in der Arbeitsgruppe von Dr. James B. Stewart
Forsuchungsgegenstand: High-throughput sequencing of mitochondrial DNA

Aug. 2004 – Aug. 2012
Grundstudium und Diplomarbeit an der Technische Universität Tampere, Finnland, Department für Chemie- und Bioingenieurwissenschaften Diplomarbeit an der Universität Helsinki, Finnland, Tierärztliche Fakultät, in der Arbeitsgruppe von Prof. Hannu Korkeala

Aug. 2002 – Mai 2004
Kaleva Gymnasium mit Abschluss Abitur, Tampere, Finnland

WEITERBILDUNG

Mai 2016	Designing and Presenting a Poster, BioScript, GSfBS, Köln
März 2016	Advanced Scientific writing, BioScript, GsfBS, Köln
Mai 2015	FELASA B (mouse), Universität zu Köln
März 2015	Statistical Literacy, Science Craft, GsfBS, Köln
Nov. 2014	Data Visualization with R, Science Craft, GsfBS, Köln
Nov. 2014	RNA-Seq Bioinformatics, ecSeq Bioinformatics, GsfBS, Köln
Feb. 2014	X Summer course in Bioinformatics, Universität von São Paulo, Brasilien

FINANZIERUNG

Oct 2016 – Jul 2017
Stipendium an Junge Forscher(innen), Emil Aaltonen Stiftung, Finnland
Sep 2016
Stipendium für Kongressreise, Deutscher Akademischer Austauschdienst
Jan 2014 – Sep 2016
Max Planck Stipendium für Promotionsarbeit

PUBLIKATIONEN

Dahlsten E.*, Isokallio M.*, Lindström M., Somervuo P., Korkeala H. (2014) Transcriptomic analysis of (Group I) Clostridium botulinum ATCC 3502 cold shock response. PLoS ONE 9(2):e8995. *equal contribution

Derman Y., Isokallio M., Lindström M., Korkeala H. (2013) The two-component system CBO2306/CBO2307 is important for cold adaptation of Clostridium botulinum ATCC 3502. International Journal of Food Microbiology, 167(1):87–91.