A SEQUENCING-BASED ANALYSIS OF EPIGENETIC MODIFICATIONS OF THE MITOCHONDRIAL GENOME

Iacopo Bicci
PhD candidate, Magdalene College

Department of Clinical Neurosciences
University of Cambridge

September 2021

This dissertation is submitted to the Board of Graduate Studies in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Abstract

A sequencing-based analysis of epigenetic modifications of the mitochondrial genome

Iacopo Bicci

Methylation on CpG residues is one of the most important epigenetic modifications of nuclear DNA, where it regulates gene expression. Methylation of mitochondrial DNA (mtDNA) has been studied using whole genome bisulfite sequencing (WGBS), but recent evidence has uncovered major technical issues, which introduce a potential bias during methylation quantification. In this study, we first validate the technical concerns with WGBS using publicly available datasets. Then we develop and assess the accuracy of a protocol for variant-specific methylation identification using long-read based technology Oxford Nanopore Sequencing. Our approach circumvents mtDNA-specific confounders, while enriching for native full-length molecules over nuclear DNA. Variant calling analysis against Illumina deep re-sequencing showed that all expected mtDNA variants can be reliably identified. By using simulated data sets, we were able to determine that the mtDNA methylation levels identified were likely false positives introduced by the technique. This observation was consistent across the multiple human primary and cancer cell lines and human tissues analysed in this study.

We therefore conclude that CpG methylation is not an epigenetic modification occurring in human mtDNA, thus resolving previous controversies. Additionally, we developed a reliable protocol to study epigenetic modifications of mtDNA at single-molecule and single-base resolution, with potential applications beyond CpG methylation

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except when specifically indicated in the text. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification. In accordance with the Clinical Medicine and Clinical Veterinary Medicine Degree Committee guidelines, this thesis does not exceed 60'000 words.

Signed: [Signature redacted]

Date: 21/07/2022

Iacopo Bicci, MSc

PhD Student, Department of Clinical Neurosciences, Magdalene College, University of Cambridge

Research Impact Statement form

Student name	Iacopo Bicci
USN	303552664
Department	Clinical Neurosciences
Supervisor	Prof. Patrick Chinnery
Thesis title	A sequencing-based analysis of epigenetic modifications of the mitochondrial genome
Research Impact Statement: Because of the Covid-19 pandemic we had to delay for months the execution and analysis of the experiments outlined in **Chapter 6**, particularly the sequencing of human tissues (which came available only in early 2021). Because of this delay we had to leave out of this dissertation the data we had collected on the study of the presence of adenine methylation (6mA) in mitochondrial DNA (mtDNA). Our original plan was to first generate positive (PC) and negative (NC) controls for this modification in a similar way to CpG methylation PC and NC (see **Paragraph 3.5**). With these, we planned to first train a Nanopolish model to detect this modification on mtDNA. We would have then on one side re-analysed all the raw Nanopore data used to check for CpG methylation, to assess 6mA methylation presence instead. In parallel, we would have studied mitochondrial 6mA presence in a new model, developed in collaboration with Dr Michele Frison. In this *in vitro* model we differentiated towards the neuronal lineage iPSCs derived from either patients with a mitochondrial disease and individuals without any known mutations, in order to check the presence of 6mA in mtDNA at various stages of differentiation. Because of the delay in completing the experiment and the analysis for the CpG methylation project (outlined in this dissertation), we were able to only partially complete the experiments required for this new part, and we therefore decided to leave those out of this dissertation.	
Student's signature *[by typing your name, you are providing your electronic signature]*	Iacopo Bicci
Supervisor signature *[by typing your name, you are providing your electronic signature]*	Patrick Chinnery

Publications arising from this thesis

Oxford Nanopore sequencing-based protocol to detect CpG methylation in human mitochondrial DNA. Iacopo Bicci, Claudia Calabrese, Zoe J Golder, Aurora Gomez-Duran, Patrick Chinnery. *Bioarxive*. 2021. https://doi.org/10.1101/2021.02.20.432086

Single-molecule mitochondrial DNA sequencing shows no evidence of CpG methylation in human cells and tissues. Bicci I, Calabrese C, Gomez-Duran A, Golder Z, Chinnery PF. *Nucleic Acids Research*. doi: 10.1093/nar/gkab1179

Acknowledgements

I would like to thank Professor Patrick Chinnery for giving me the opportunity and the resources to undertake this research in his laboratory. I would also like to thank the people who lend me their help and assistance, primarily Dr Claudia Calabrese, without whom this research would not have seen the light of the day and Dr Aurora Gomez-Duran for her invaluable assistance in the lab. A special thank goes also to the Human Research Tissue Bank (supported by the NIHR Cambridge Biomedical Research Centre) for providing access to the human tissues used in this study. I would also like to thank the rest of the members of the Chinnery lab and of the Mitochondrial Biology Unit, for making me look forward to going to work every day.

I would also like to thank my funding sponsors: the MRC DTP scholarship program, the Cambridge Trust, the Beverly-Sackler fund and Magdalene College Cambridge, for making it possible for me to undertake this PhD research project. Finally I would like to thank my parents for their unwavering support over all these years, I swear I'll find a real job soon!

Per experienzia e per sottile indagazione

Table of Contents

List of abbreviations

5hmC → 5-hydroxymethylcytosine

5mC → 5-methylcytosine

6mA → 6-methyladenine

ADP → Adenosine diphosphate

ATP → Adenosine triphosphate

AUC → Area under the curve

BG → Bias group

BS → Bisulfite

CpG → CG dinucleotide

CVD → Cardiovascular disease

dC → Cytosine nucleoside

ddPCR → Droplet digital PCR

DIN → DNA integrity number

DM → Differential methylation

DM-CpG → Differentially methylated CpG

DMEM → Dulbecco's modified eagle medium

DMRs → Differentially methylated regions

DNMT1 → DNA methyltransferase 1

DNMT3a and *3b* → DNA methyltransferase 3a and 3b

DPBS → Dulbecco's phosphate buffer saline

ES cells → Embryonic stem cells

ETC → Electron transport chain

EtOH → Ethanol

$FADH_2$ → Reduced flavin adenine dinucleotide

FBS → Foetal bovine serum

FP → False positive

FPR → False positive rate

gDNA → Genomic DNA

GFP → Green fluorescent protein

H3 → Histone H3

H3K4 → Lysine 4 of Histone H3

HMM → Hidden Markov model

HSP1 and HSP2 → H-strand promoter 1 and 2

IAP → Intracisternal A particle

IMM → Inner mitochondrial membrane

IMS → Inter-membrane space

LBG → Low-bias group

LC/MS → Liquid chromatography/mass spectrometry

LHON → Leber hereditary optic neuropathy

LLR → Log-likelihood ration

LR-PCR → Long-range PCR

LSD1 → Lysine-specific demethylase 1

MBD → Methyl-CpG-binding domain

Me-DIP seq → Methylated DNA immunoprecipitation sequencing

MELAS → Mitochondrial myopathy, encephalopathy, lactic acidosis, stroke-Like episodes syndrome

MERRF → Myoclonic epilepsy with ragged red fibers syndrome

MICOS → Mitochondrial contact site and cristae organising system

mtDNA → Mitochondrial DNA

MTS → Mitochondrial targeting sequence

NADH → Nicotinamide adenine dinucleotide

nDNA → Nuclear DNA

NC → Negative control

NCR → Non-coding region

NGS → Next-generation sequencing

NuMTs → Nuclear mitochondrial sequences

O_L and O_H → Origin of replication of the Light and Heavy strand

OMM → Outer mitochondrial membrane

ONS → Oxford Nanopore sequencing

ORF → Open reading frame

OXPHOS → Oxidative phosphorylation

PC → Positive control

PCR → Polymerase chain reaction

PEG → Poly-ethylene glycol

Pi → Inorganic phosphate

POLG → Polymerase gamma
qPCR → Quantitative PCR
QS → Quality score
RC → Respiratory chain
rCRS → Revised Cambridge reference sequence
rRNA → Ribosomal RNA
ROC → Receiving operating characteristic (curve)
ROS → Reactive oxygen species
RRBS → Reduced-representation bisulfite sequencing
RT → Room temperature
siRNA → Small interfering RNA
SMRT → Single-molecule real-time
SNV → Single nucleotide variant
SRA → Sequence read archive
SSBP1 → Single stranded binding protein 1
STED → Simulated emission depletion super resolution microscopy
TCA → Tricarboxylic acid cycle
TET → Ten-eleven translocation
TFAM → Mitochondrial transcription factor A
TIM → Inner membrane transporter
TN → True negative
TOM → Outer membrane transporter
TP → True positive
TRD → Transcriptional repression domain
tRNA → Transfer RNA
UHRF1 → Ubiquitin Like With PHD And Ring Finger Domains 1
VDAC → Voltage-dependent anion-selective channel
WB → Western blot
WGBS → Whole genome bisulfite sequencing
WGS → Whole genome sequencing

Chapter 1. Introduction

1.1 Mitochondrial Biology

1.1.1 Mitochondria origin

Mitochondria are intracellular organelles present in the majority of the eukaryotic cells, specialised in energy production in the form of adenine triphosphate (ATP) through oxidative phosphorylation (OXPHOS)[1]. Through this latter process, the coupling of adenine diphosphate (ADP) and inorganic phosphate (Pi) through a redox reaction is linked to a transfer of electrons from reduced cofactors (NADH and FAD_2) to O_2, via the electron transport chain (ETC) protein complexes[1].

To postulate the origin of these organelles within eukaryotic cells, the most accredited explanation is the *endosymbiotic* theory, which hypothesises that mitochondria descend from free-living α-proteobacteria, a type of organism which requires the insertion into a host cell as a step required for its replication[2]. These original endosymbiotic cells are thought to have helped the evolution of eukaryotes by complementing their nuclear genes functions[3]. A more recent study has somewhat challenged this view, claiming that mitochondria in fact evolved from a prokaryote group that branched before α-proteobacteria[4]. Regardless, horizontal gene transfer from the mitochondrial to the nuclear genome of the host cell has been suggested to be the crucial step for the transition from endosymbionts to organelles[3]. This relocation of the mitochondrial genes during evolution of eukaryotes has been suggested by the observation that the majority of the proteins essential to the integrity, replication and expression of the mitochondrion structure and genome are indeed encoded by the nucleus[2]. This evidence were further supported by comparative analyses performed with the latest DNA sequencing technologies, which additionally unveiled the presence of non-functional nuclear mitochondrial sequences (NuMTs) within the genome of both plants and animals, including humans[5].

1.1.2 Mitochondrion structure

Mitochondria possess a double phospholipid bilayer membrane, dividing the mitochondria into two internal aqueous compartments, the inter-membrane space (IMS) and the matrix[6].

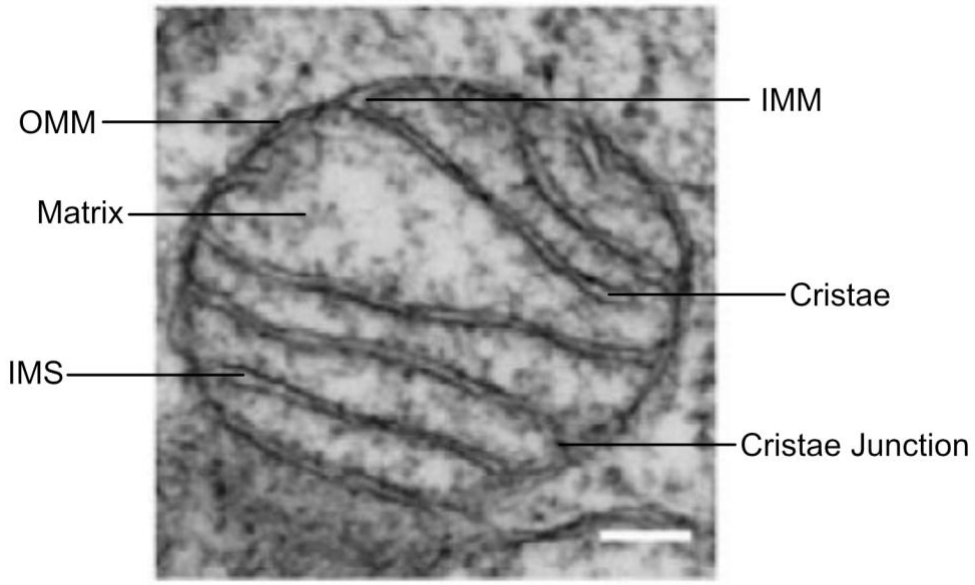


Figure 1.1: Mitochondrial structure. **Transmission electron microscopy image of a round shaped mitochondrion. OMM: Outer mitochondrial membrane; IMM: Inner Mitochondrial Membrane, IMS: Inter Membrane Space. (scale bar = 100 nm). Adapted from Prudent J. et al, 2015[7]**

The latter is the innermost mitochondrial compartment, it has a high pH (7.9-8) and an elevated protein density, and it hosts key metabolic processes such as the tricarboxylic acid cycle (TCA, also known as Kreb's cycle or citric acid cycle)[8]. The TCA is a series of chemical processes that participate to the breakdown of energy molecules (such as glucose) in order to produce NADH and $FADH_2$ feeding into the OXPHOS[1] reaction.
In the matrix are also present multiple copies of mitochondrial DNA (mtDNA), and as such it is the site where mtDNA replication, transcription and translation happens.
Enveloping the matrix, the inner mitochondrial membrane (IMM) is organised into invaginations called cristae, connected to the inter-membrane space (IMS, the second aqueous compartment the mitochondrion is divided in) at the cristae junctions[9]. The part of matrix contained within the cristae invaginations is called cristae lumen, and it is the site where the ATP production mainly takes place. This is due to the peculiar pH of the cristae lumen (7.2), forming the transmembrane electrochemical gradient driving the ATP synthesis, and the local concentration of the proteins of the ETC[10].
The cristae lumen is connected to the IMS via tubular openings located at the cristae junctions forming a complex structure called mitochondrial contact site and cristae organising system (MICOS)[11]. In this space, the proteins apt to the shuttling of metabolites, ions, ADP and ATP, along with the protein import complexes such as the

translocase inner membrane (*TIM*)[11] are also contained. This protein forms a super complex with the other protein importer complex, the translocase outer membrane (*TOM*), located on the outer mitochondrial membrane (OMM). This latter membrane is more porous than the IMM, with ions and small uncharged protein freely traversing through porins such as the voltage-dependent anion channel (*VDAC*)[12]. The OMM also possess proteins functioning as tethers anchoring the mitochondrion to other cellular organelles, and that it plays a central role in shaping mitochondrial morphology[13].

1.1.3 Mitochondria function

Mitochondria also play additional functions to energy production, such as regulators of apoptosis[14,15], calcium homeostasis[16–19], antigen presentation[20] and various immune response functions[21,22]. Mitochondria are also the source of a variety of important metabolites used in many cellular functions, such as the by-products of the TCA cycle[23], of the biosynthesis of cellular pyrimidines[24], steroids, heme, and the β-oxidation of fatty acids[25]. They also produce metabolites called reactive oxygen species (ROS), which have an important cellular signalling function, and mitochondria maintain the cellular redox balance through OXPHOS[26].

However, the principal activity of mitochondria is their capability to produce energy via OXPHOS and the ETC[27]. The latter is comprised of 4 complexes (I to IV), responsible of keeping the proton gradient across the IMM (except complex II), and two electron carriers (coenzyme Q10 and cytochrome C). All of these proteins are required for the correct functioning of the ATP synthase, or complex V[28]. Complex I (also called NADH-ubiquinone oxidoreductase) is comprised of 45 subunits, of which only 7 are encoded by the mtDNA (the rest are nuclear), while complex II (or succinate dehydrogenase) contains 4 completely nuclear-encoded subunits[29,30]. Complex III (or ubiquinol-cytochrome c oxidoreductase) has 11 subunits, only one of which is encoded by mtDNA (cytochrome b), whereas Complex IV (or cytochrome c oxidase) possesses 14 subunits (3 mitochondrial and the rest nuclear)[31,32]. Finally, Complex V (the ATP synthase) has 19 subunits and only *MT-ATP6* and *MT-ATP8* encoded by mtDNA[33].

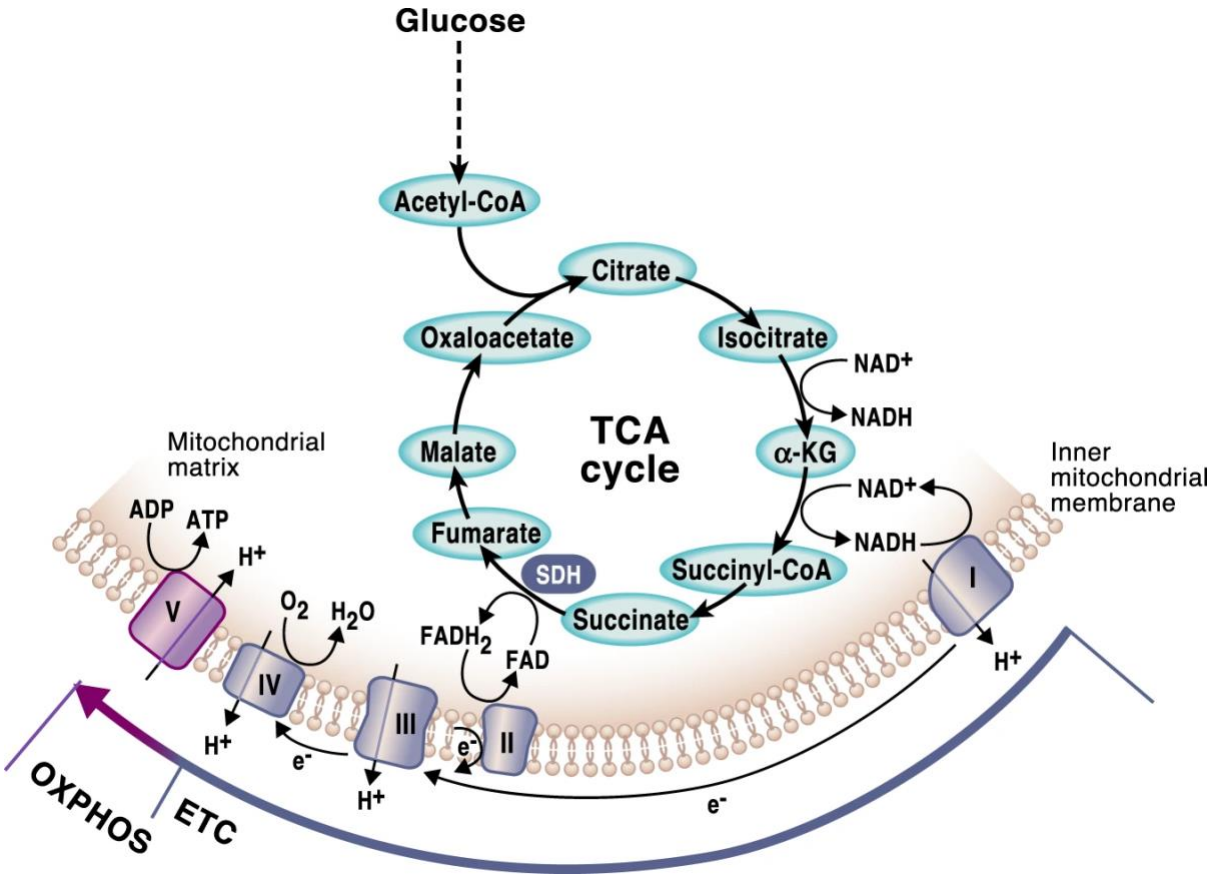


Figure 1.2: Overview of the respiratory complexes machinery and the TCA cycle.

The complexes forming the respiratory chain (in purple) are shown embedded in the mitochondrial inner membrane, and the electron transfer is shown. The steps of the TCA cycles are shown inside the mitochondrial matrix, and the formation of NADH and $FADH_2$ is shown. Complex V (ATP synthetase is represented in purple at the end of the respiratory chain. Adapted from Martinez-Reyes et al. 2020[34]

NADH and $FADH_2$ produced by glycolysis, the TCA cycle and/or the β-oxidation of fatty acids supply the electrons to Complex I and II, respectively, where NADH is oxidised to NAD^+ and $FADH_2$ to FAD. The electrons are then transferred to Coenzyme Q10 thanks to reduction of ubiquinone to ubiquinol. This process facilitates the transfer of $4H^+$ ions through the IMM, in order to generate the electrochemical gradient[35] Ubiquinol is then oxidised by the Q cycle of Complex III, where electrons pass to cytochrome c (via its reduction), facilitating even more transfer of H^+ ions through the IMM[36]. Finally, electrons flow through cytochrome c to oxygen (the final electron acceptor), through the action of Complex IV, which again facilitates the passing of H^+ ions through the IMM.

The electron transfer process generates reactive oxygen species (ROS), which can on one side participate in mitochondrial signalling[37] and on the other potentially damage cellular components[38]. The proton gradient across the IMM generated by the ETC complexes creates the so-called proton-motive force necessary for the ATP synthase to generate ATP from ADP and Pi, through a rotary catalysis process[39].

1.2 Mitochondrial genomics

1.2.1 The mitochondrial genome organisation

In addition to chloroplast DNA in plants, mtDNA is the only source of critical cellular proteins outside the nucleus of eukaryotic cells, and it is usually organised as a circular double-stranded DNA molecule.

The length in humans is generally consistent around 16569 base pairs (bp), but its length may vary greatly in other species (up to thousands of Kbp in angiosperms[40])[41]. The two strands have a different nucleotide composition, which makes them possible to separate physically by density centrifugation in alkaline thiophosgene (CsCl2) gradient[42]. The lighter strand (L-strand) is rich in cytosine residues, while the heavy strand (H-strand) has a higher guanine content.

Moreover, mtDNA is present in multiple copies inside of the mitochondria (between 100 and 10'000 copies), varying according to the cellular energy demands[43]. The mitochondrial genome contains 37 genes, the majority (28) on the H-strand and only 9 on the L-strand. Of these, 13 genes encode essential subunits of the Complexes I, III, IV and V of the mitochondrial respiratory chain, essential for energy production through OXPHOS. Additionally, 24 genes encode for transport RNA (tRNA) molecules and 2 for the essential components of the mitochondrial ribosome: 16s rRNA and 12s rRNA[43]. Unlike nDNA, ~93% of the mtDNA is comprised of coding regions, lacking introns, and two sets of genes (*MT-ATP6*/*MT-ATP8* and *MT-ND4/MT-ND4L*) share an overlapping region of the H-strand during transcription (**Figure 1.3**).

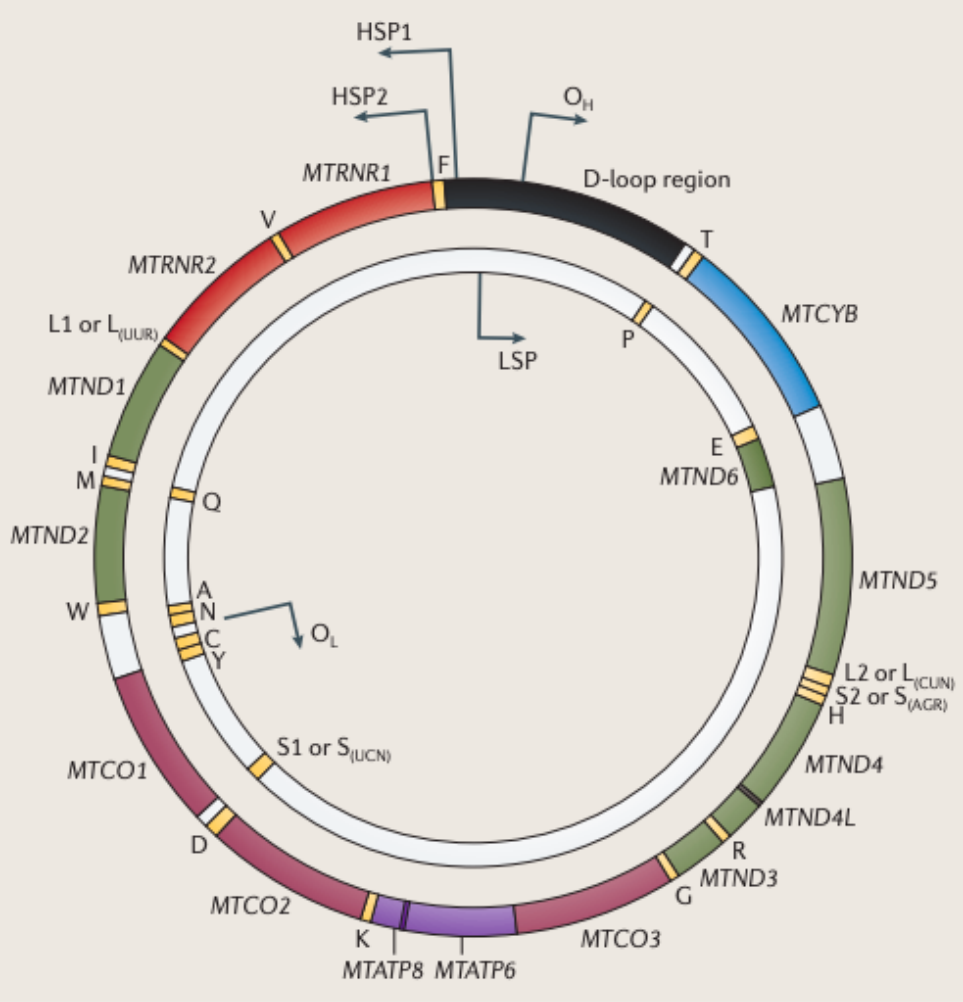


Figure 1.3: Mitochondrial genome structure. **The two strands are shown, with the H-strand on the outer and the L-strand in the inner side. Non-coding D-loop is show in black, the two rRNA genes in red and tRNA genes in yellow. Other genes forming subunit of the respiratory chain are shown. Adapted from Stewart and Chinnery, 2015**[44]

The rest of the genes are mostly contiguous, with coding genes usually separated by a tRNA and/or a few non-coding base pairs (**Figure 1.3**).

Only one significant non-coding region (NCR) is present in mtDNA, called the displacement loop (D-loop), because it has incorporated a small single strand DNA (called 7s DNA), complementary to the L-strand. The exact function of this structure has been debated, and it has been suggested that it probably plays a role in the regulation of mtDNA topology, replication or even association to the matrix membrane[45]. This NCR contains the sites for the initiation of mtDNA replication (called origin of the heavy strand synthesis, or O_H), two promoters for the transcription of the H-strand (HSP1 and HSP2) and one promoter for the transcription of the L-strand

(LSP), located on the opposite strand. The origin of the replication of the L-strand (O_L) is located 11 kbp downstream of O_H.

The mitochondrial genetic code is also different from the nDNA, as not all codons encode for the same amino acids as in nDNA, and two non-conventional codons ("AGA" and "AGG") induce a frameshift in human mitochondrial ribosomes[46].

During the formation of the mammalian zygote, sperm mtDNA is removed by ubiquitination during the transport through the male reproductive tract[47]. For this reason, mtDNA is maternally inherited, and the mtDNA content of the zygote is only determined by the contents of the unfertilised egg[43].

1.2.2 Nucleoids structure

In a parallel fashion to nDNA, mtDNA molecules are also packed around proteins, in order to form structures called nucleoids[48]. Nucleoids are mostly localised to the cristae junction of the IMM, although they could be identified in other parts of a mitochondrial network[49,50]. The protein responsible of nucleoids formation is the Mitochondrial Transcription Factor A (*TFAM*), and it has been described that up to over 1000 copies of this protein may coat a single mtDNA molecule[51–53]. Previous estimates gave a range from 2-10 molecules for every nucleoid, but recent evidence on stimulated emission depletion super resolution microscopy (STED) challenged those results and confirmed the presence of a single mtDNA molecule per nucleoid[53]. *TFAM* binding to mtDNA and to itself generate negative supercoiling, which bends the mtDNA molecule compacting and reducing its size[54]. Other proteins have been identified associated with mtDNA and *TFAM* in nucleoids structures, namely: the transcription factors *B1M* and *B2M* and the single stranded binding protein (*SSBP1*)[55].

1.2.3 mtDNA copy number regulation

As mentioned in **paragraph 1.2.1**, tissues with high energetic demands such as heart, muscles and the brain possess a higher number of mitochondrial copies comparing to others (such as the spleen, a far less energy-demanding tissue)[56,57].

The mitochondrial copy number is mainly maintained at an optimal level by nuclear-encoded proteins which regulate mtDNA replication[58], such as *TWINKLE*[59], *SSBP1*[60], DNA polymerase γ (*POLG*)[61] and *TFAM*, which levels are critical for the maintenance of mtDNA copy number[62,63].

However, the regulation of mtDNA copy number was shown to be also dependent on a number of additional factors, including: metabolic and transport enzymes[64], changes in mitochondria dynamics[65], cytoskeletal proteins[66], factors influencing mitochondrial biogenesis[67], protein chaperones[68] and various exonucleases and proteases[69,70].

1.2.4 Human mtDNA variation and mitochondrial diseases

As mentioned previously, mtDNA is polyploid, with multiple copies present inside of the mitochondria (between 100 and 10'000 copies), varying according to the cellular energy demands[43]. When there is a condition of genetic homogeneity of the mtDNA molecules of a cell/organism we define this condition as homoplasmy. On the contrary, whenever there is a situation of co-existence between a wild type and mutant DNA, this condition is referred to as heteroplasmy.

Because of the special circumstances of mtDNA inheritance (**paragraph 1.2.1**), there has been negligible intermolecular recombination of mtDNA[43], although some studies have confirmed that it does in fact happen in human mtDNA as well[71]. On the other hand, the evolutionary rate of mtDNA is higher than the average nuclear gene[72], resulting in a sequence variation that evolved as a sequential accumulation of maternally inherited mutations[73]. Because of this characteristic, it has been possible to retrace the whole human lineage back to a common matrilinear ancestor living approximately 200'000 years ago in Africa[74–76]. Subgroups of mtDNA haplotypes found in the present human genome pool which descend from a common ancestral mitochondrial genome are defined as haplogroups[72]. Thanks to analysis of human mtDNA haplogroups it has been possible to trace the migrations of humans out of Africa[72]. The common mtDNA variants that define an haplogroup are usually fixed (homoplasmic) in a population, and they generally do not possess any detrimental phenotypic effect[43], although it has been suggested that some regional variation could be explained by favourable effects that these variants have on the adaptation to the environment. For example, in colder climates common variants have been associated to a looser coupling between oxidation and phosphorylation, which in turn may generate additional heat production at the expenses of ATP production, favouring the survival of human populations in those areas[77].

Some rare mtDNA mutations are the primary cause of mitochondrial diseases in humans, such as Leber hereditary optic neuropathy (LHON), mitochondrial

encephalomyopathy, lactic acidosis and stroke-like episodes (MELAS) and myoclonic epilepsy with ragged red fibers (MERRF) syndromes. These rare mutations have an estimated incidence of 1 in 5000[78] and they primarily occur on either tRNAs or protein coding genes. This in turn results in a reduced energy production, either through impairment of mitochondrial protein synthesis or through reduced activity of the RC enzymes[79]. Differently from haplogroups-defining variants, these rare mutations are often heteroplasmic, with notable exception such as the *MT-ND1* primary LHON mutations (found in homoplasmy in > 90% of the cases)[80]. In the majority of the cases however, an heteroplasmic mutation has to reach a specific threshold to generate a pathological effect[81,82], and this threshold is very specific according to both the mutation and the context in which it occurs[81]. In fact, heteroplasmy level of a pathogenic mutation can not only vary from one cell to another in the same organ, but also between organs of the same person, and people of a same family[44,83].

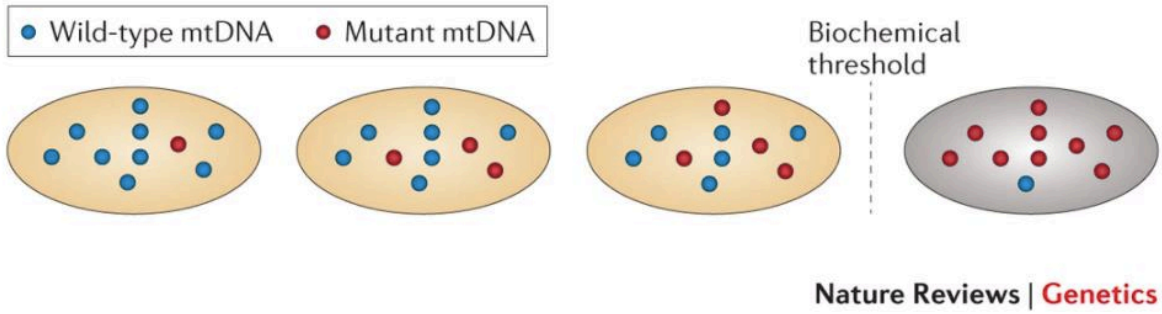


Figure 1.4: **mtDNA mutations threshold effect. Mitochondria represented on the left have different level of heteroplasmy of their mutated mtDNA molecules. This does not result in a phenotypic defect. Only when the heteroplasmy of the mutation reaches a biochemical threshold we can have a deleterious effect, as represented on the mitochondria on the right. Adapted from Stewart and Chinnery, 2015**[44]

1.3 Epigenetics and DNA methylation

By definition, epigenetics is the field that studies phenotypic changes that are inheritable but not related to alterations of the DNA sequence. Unlike genetic modifications such as mutations or DNA rearrangements, epigenetic changes are usually reversible, and do not involve changes in the DNA sequence, but rather they affect how the genetic information is interpreted[84]. Epigenetic modifications include direct modifications of DNA (such as methylation of the 5th cytosine carbon) or

modifications of the histones tails, which influence chromatin accessibility[85,86]. The epigenome is responsible for maintaining the phenotype specification of individual cells and tissues, by maintaining the cell- or tissue-specific gene expression states[87–89]. In mammals, it was thought that cytosine methylation in CpG context, a very stable epigenetic modification, was the only DNA modification present. However, recent evidence have shown presence of non-CpG cytosine methylation (in stem cells in particular[90,91]) or adenine methylation[92] (although the presence of this latter modification in the mammalian genome is still somewhat debated[93]).

Between 60-80% of the ~29 millions CpG residues of the human genomes are methylated[94], and they are classified according to their density. CpG islands are defined as high density methylation-resistant areas, with 7% of the CpG sites located in these regions[95]. Around ~70% of the annotated gene promoters are associated to CpG islands[96].

1.3.1 *De novo* DNA methylation patterns

Because of binding of transcription factors, exclusion by nucleosomes rearrangement or the action of histone modifiers (such as *H3K4* methyltransferase *SETD1A*[97]), most of the CpG islands-associated promoters are protected from methylation[97]. Despite this, some promoters become methylated during development, repressing in turn the gene expression (*de novo* methylation)[97]. This process is executed by the action of the DNA methyltransferases *DNMT3a* and *DNMT3b*[98], in complex with the protein *DNMT3L*, a related homolog lacking catalytic activity[99,100]. This protein is responsible for the recognition of unmethylated *H3K4* and the recruitment of the *de novo* methyltransferases to the site[99].

The targeting of *DNMT3a* and *DNMT3b* to gene promoter is usually in conjunction with other repressors of the gene expression, such as histone deacetylases and *H3K9* methyltransferases[101], and it is started by the binding of transcription factors[97]. The observed crosstalk between histone modification and DNA methylation suggest the former initiate the formation of heterochromatin and DNA methylation serve as a stable silencing of the target promoter[97].

1.3.2 Maintenance of DNA methylation modifications

In humans, DNA methylation patterns are transmitted during DNA replication through the action of the high-fidelity DNA methyltransferase *DNMT1*[102,103], which shows a strong preference for hemi methylated DNA[103]. In fact, this protein gets recruited directly into the DNA replication fork through interactions with the protein *PCNA* and *UHRF1*, which recognises hemi methylated sites through its SRA domain[104].

The activity of *DNMT1* is regulated through post-translational modifications[97]. The most important protein modulating *DNMT1* stability through regulation of its methylation status is lysine-specific demethylase 1 (*LSD1*), which is then in turn essential for the maintenance of global DNA methylation patterns[105].

Another factor regulating *DNMT1* stability during the S phase is the methylation of the lysine 9 of the histone *H3*, the binding site of *UHRF1*[106]. These kind of interactions with other heterochromatin-associated proteins make sure that *DNMT1* is active and stabilised only during the DNA replication phase, to provide fidelity to global DNA methylation replication[97].

1.3.3 DNA methylation removal

Two possible mechanisms of DNA methylation are known, either passive or active. Methylation is gradually removed from the DNA pool when it is not maintained during the successive replication rounds. In the active demethylation process, specific enzymes use energy to remove methyl groups bound to 5mC[107]. Such enzymes have been identified in the ten-eleven translocation (*TET*) enzyme family, *TET1*, *TET2* and *TET3*[108]. To remove the methylation residue the first step is to oxidise 5mC to 5hydroxymethylcytosine (5hmC), then to the other oxidated states 5-formylcytosine and 5-carboxylcytosine[109]. These modified bases have attracted much attention lately as they are being increasingly recognised as having independent roles as modifiers of gene expression rather than mere oxidated intermediates of 5mC[107]. DNA methylation is finally completed by either diluting the 5mC oxidation derivatives during replication, or by base excision repair[110,111].

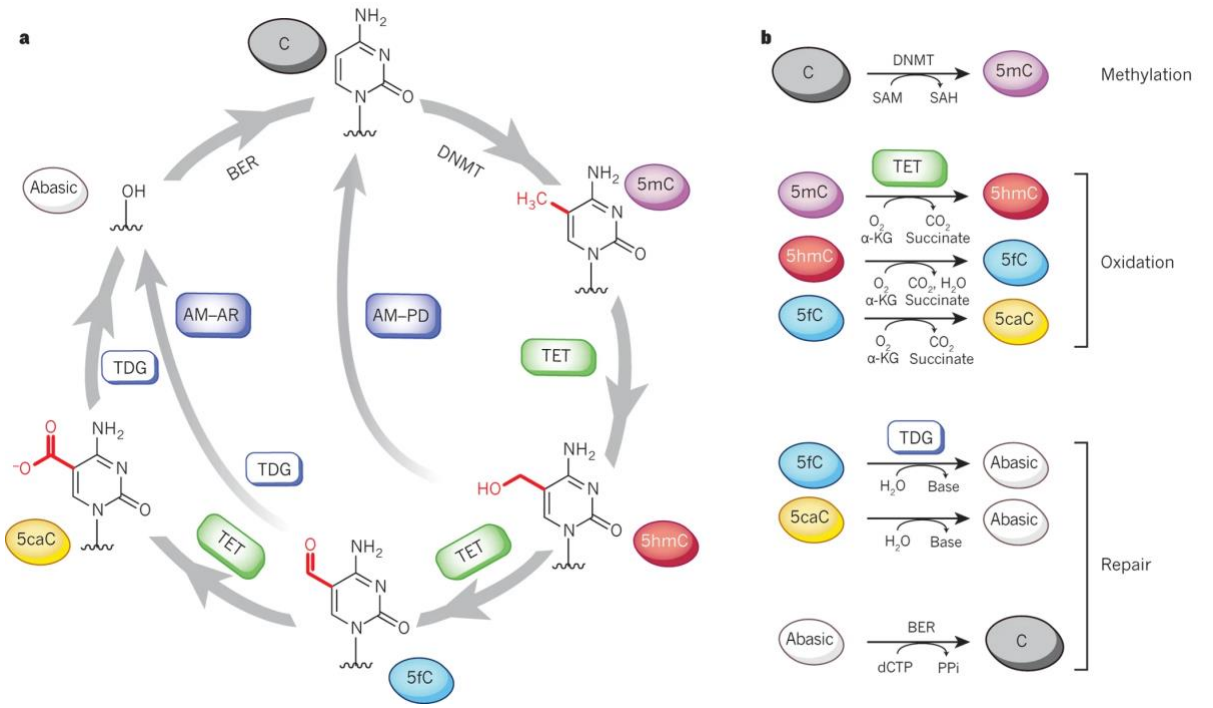


Figure 1.5: **Diagram showing the cytosine oxidation states. a) Diagram representing the cytosine oxidation cycle. b) Individual reactions to generate the oxidised residues shown in diagram in a). DNMT: DNA methyltransferase; TET: ten eleven translocation; thymine DNA glycosylase; a-KG: Alpha-ketoglutarate; 5mC: 5-methylcytosine; 5hmC: 5-hydroxymethylcytosine; 5fC: 5-formylcytosine; 5caC: 5-carboxylcytosine. Adapted from Kohli and Zhang, 2013**[107]

1.4 DNA methylation and gene expression

As mentioned above, not only the distribution of CpG residues in the genome is non-random, but their methylation status is also tightly regulated by the *DNMT* and *TET* enzymes. DNA methylation is essential for silencing retroviral elements, regulating tissue-specific gene expression, genomic imprinting, and X chromosome inactivation, and it may exert different influences on gene activities based on the underlying genetic sequence.

1.4.1 Silencing of retroviral elements

Transposable viral elements are potentially harmful DNA sequences that have the potential of being replicated and inserted in the human genome, causing gene disruption and DNA mutations[112–116]. It is estimated that 45% of the mammalian genome consists of transposable and viral elements that are silenced by bulk

methylation[117]. However, the majority of these elements are inactivated by DNA methylation or by mutations acquired over time as the result of the deamination of 5mC[118].

As it is demonstrated in the case of the intracisternal A particle (IAP), one of the most aggressive retrovirus in the human genome[118], one of the main roles of DNA methylation in intergenic region is to repress the expression of these potentially harmful elements. In fact, not only IAP is heavily methylated throughout life in gametogenesis, development, and adulthood[118,119], but even within the embryo, when the rest of the genome is relatively hypomethylated, *Dnmt1* maintains the repression of IAP elements[119]. On the contrary, when *Dnmt1* is depleted by genetic mutations, leading to extensive hypomethylation, IAP elements are expressed[118,120].

1.4.2 How CpG island methylation controls gene expression

As mentioned in **paragraph 1.3**, stretches of at least 200bp containing a high CpG density are defined as CpG islands, and are often found unmethylated[87]. Around 70% of gene promoters are found within these DNA regions[96], especially those for housekeeping genes[121], and for this reason they are usually found highly conserved between the species[122].

CpG islands have evolved to promote gene expression by the regulation of transcription factor binding and of the chromatin structure. In fact, a common feature of CpG islands is their low nucleosomes content compared to other DNA stretches[123–125], which are also often associated with histone modification involved in enhancing gene expression[123,126]. Despite more than half of the CpG islands are known to contain transcription start sites, they are usually devoid of other common promoter elements such as TATA boxes[127]. Despite this, CpG islands still enhance the accessibility of DNA and promote transcription factor binding, because many such binding sites are rich in GC residues.

On the contrary, methylation of CpG islands results in stable silencing of gene expression[128]. As mentioned in the previous paragraphs, methylation patterns are established during both gametogenesis and early embryonic development, when CpG islands go through a process of differential methylation[129–132], for example during the

establishment of imprinted genes[125,129–131] (where the expression of a gene is determined by the parent of inheritance).

Moreover, CpG island methylation regulates gene expression during development and stem cells differentiation[128,133–136]. Contrary to what could be expected, only CpG islands in intragenic and gene body regions (and not those associated with transcription start sites) have tissue-specific methylation patterns[122,136–139]. These regions are called CpG islands "shores", they can be located as far as 2 kbp from CpG islands and their methylation correlates with a reduction in gene expression[140].

1.4.3 Methylation of CpG sites in gene bodies

"Gene body" is defined as the region of a gene past the first exon, as it was shown that methylation of the first exon leads to gene silencing (similarly to promoter methylation)[141]. Experimental evidence show that methylation of the gene body is associated with elevated gene expression in dividing cells[142–144], while this become not significant in slowly dividing or nondividing cells such as neurons[144–146]. Despite these evidence, it is still unclear how DNA methylation of the gene body specifically contributes to gene expression regulation.

1.5 Reading DNA methylation patterns

Although DNA methylation can impair gene expression directly by steric hindrance of transcription factor binding, a specific class of proteins achieves the same purpose by binding with high affinity to 5mC. In particular, DNA methylation is recognized by three separate families of proteins: *MBD*, *UHRF* and the zinc-finger proteins.

MBD (methyl-CpG-binding domain) contain a conserved MBD domain that confers the proteins a high affinity for single methylated CpG sites[147]. This protein family includes: *MeCP2* (the first identified methyl-binding protein), *MBD1*, *MBD2*, *MBD3*, and *MBD4*[148–150]. They are found highly expressed in the brain (more than any other tissues), as many *MBD*s are important for neural development and function[151]. A particular characteristic of most *MBD* proteins is the presence of a transcriptional repression domain (TRD), that allows *MBD* proteins to bind to a variety of repressor complexes[147,152,153]. Also, *MeCP2* does not only act as a transcriptional repressor, but it also appears to have a unique role in the maintenance of DNA methylation. *MeCP2*

binds to *DNMT1* via its TRD and can recruit *DNMT1* to hemi-methylated DNA to perform maintenance methylation[154].

Another class of methyl-binding proteins are the ubiquitin-like, containing PHD and RING finger domain (*UHRF*) proteins, which includes *UHRF1* and *UHRF2*. As their name implies, these proteins bind methylated cytosines via a SET- and RING-associated DNA-binding domain[155]. However, contrary to *MBD* proteins, the primary function of *UHRF* proteins is not to bind to DNA and repress transcription. Instead, proteins of the *UHRF* family first bind to *DNMT1*, then they target it to hemi methylated DNA in order to maintain DNA methylation during DNA replication[156–158].

Finally, the third family of methylated DNA-binding proteins is composed of the protein *Kaiso*, *ZBTB4* and *ZBTB38*, which all share the presence of a zinc-finger domain in their molecular structure[159,160], and are found highly expressed in the brain. Similarly to proteins of the *MBD* family, zinc-finger domain proteins repress transcription in a DNA methylation-dependent manner[159–162].

These families of methyl-binding proteins also serve as a link between DNA methylation and modifications on the histone tails, as both the *MBDs* and the *UHRF* proteins interact with methylated DNA and histones to facilitate gene repression[152,153,163–165]. For example, not only *MeCP2* recruits histone deacetylases to remove active histone modifications, repressing gene transcription[163,166,167], but it also enhances the repressive chromatin state by recruiting histone methyl-transferases that add repressive H3K9 methylation[167].

1.6 Technologies used to study DNA methylation and mtDNA-related problems

1.6.1 Mass spectrometry

The most sensitive technology that can be used to assess the methylation level of a DNA sample is based on mass spectrometry[168]. In the most recent method developed, DNA is analysed with liquid chromatography coupled with mass spectrometry (LC/MS) after digestion of DNA to individual nucleosides[169]. In the context of mtDNA research, this implies that mitochondria have to be isolated from the target cells/tissues before sequencing, to avoid contamination with nDNA. Therefore, steps that assure the purity of the mitochondrial preparation (WB, fractionation, sucrose gradient isolation, etc) first, and of the mtDNA elution later (such as RNAse treatment) are required[169]. The

earliest studies that identified methylation of the mitochondrial genome were based on this technology[170,171] and the methylation that was measured was ~5%. These results were refined further by a 2018 study by Matsuda and colleagues[169], which found even lower mtDNA methylation levels, at around 0.3%. Apart from the laboursome preparation protocol to analyse mtDNA methylation with this method, the principal disadvantage of mass spectrometry is that this technology does not provide any information on the DNA sequence, because each genome is digested to individual nucleoside level.

1.6.2 Bisulfite sequencing

Most of the technology currently available to measure DNA methylation, including the current gold standard whole genome bisulfite sequencing (WGBS) are based on the chemical treatment of DNA with sodium bisulfite. This reaction facilitates the conversion of unmethylated cytosines to uracils (which then become thymines upon PCR amplification), while leaving methylated ones unaffected. Based on this simple principle a variety of technologies have been developed, all with specific advantages and disadvantages, that are briefly described in the next paragraph. Some downfalls of using bisulfite treatment on DNA are worth for consideration. Firstly, the bisulfite treatment is very harsh, leading to DNA degradation and problems in PCR amplification, therefore large amounts of input DNA are often required. This problem is particularly important in mtDNA methylation research, as recent reports have in fact highlighted that bisulfite preferentially degrades unmethylated C-rich regions, such as the mitochondrial L-strand. This could potentially introduce biases in mtDNA methylation calculation, and this hypothesis will be explored in **chapter 4** of this dissertation.

The analysis of bisulfite-converted data require dedicated bioinformatic tools that are more sophisticated than those required for unconverted DNA. Usually the bisulfite-converted reads needs to be aligned to a bisulfite-converted reference genome, in order to infer the methylation calls.

As WGBS is based on Illumina sequencing, it suffers from the same issues common to other short-read sequencing technologies, such as mapping issues in repeated or low complexity regions (including heavily GC rich regions and repetitive DNA). These

problems are further exacerbated by the loss of sequence diversity following bisulfite conversion (see **Figure 3.1**)[172].

1.6.3 Other bisulfite-based sequencing technologies

When only a small number of genomic loci are to be investigated (≤ 20), the most effective solution is amplicon sequencing, where DNA is treated with bisulfite first, then amplified with specific primers, barcoded and sequenced[173].

For a larger number of regions, capture-sequencing is usually preferred, as it avoid labour-intensive primer pairs design although it does require the synthesis of a probe panel. Capture by hybridisation can then be performed either before[174] or after[175–177] bisulfite conversion. In this latter case, there is a risk of introducing biases in the methylation quantification based on the preferential binding of the probes to certain methylation states. This solution is usually cost-effective for large cohort-based studies, where commercially available probes panels are used[178].

Another cost-efficient solution to analyse mammalian genomes is reduced representation bisulfite sequencing (RRBS)[136,179]. This technology is based on the digestion of the target genomes by the restriction enzyme *MspI*, which cuts CCGG motives irrespective of their methylation state[180]. This enables the enrichment of regions of high CpG density such as CpG islands, enhancer and promoters[95]. RRBS has been found to be informative for 85% of CpG islands, representing < 3% of the genome and therefore greatly reducing sequencing costs[180].

However, it is obvious that the main drawback of RRBS is its reliance on the presence of MspI restriction sites. Also, *MspI* digestion creates an intrinsic lack of diversity at the beginning of all sequenced reads, which could possibly interfere with calibration and cluster detection on the latest Illumina sequencers[181].

1.6.4 Methylated DNA immunoprecipitation sequencing (Me-DIP seq)

Me-DIP sequencing is a well-established approach for identifying CpG-rich genomic sequences and to identify differentially methylated regions. As the name implies, Me-DIP sequencing is based on the immunocapture of methylated DNA residues using anti 5mC monoclonal antibodies, coupled with next generation sequencing of the isolated, fragmented products[182]. The high specificity of monoclonal antibodies make

this technique ideal for the specific detection of 5mC (either in CpG or non-CpG context), but also 5mGC or 5hmC as well[183]. DNA input requirements are also relatively low, and the downstream analysis is more straightforward compared to bisulfite-based methods, making Me-DIP seq an attractive alternative[184].

The principal downside of Me-DIP seq is that methylation levels cannot be resolved at single-base resolution level, nor hemi-methylation status can be determined. Me-DIP seq data is shown as differentially methylated regions (DMRs), which are represented as peaks of methylation enrichment across the genome. Since the maximum resolution size of the peaks is around 100-150 bases, this is usually sufficient for most studies, since methylation bordering CpGs has been shown to correlate significantly with regions distant up to 1 kbp[138].

If methylated regions are inferred by the read enrichment, unmethylated ones must be inferred by lack of reads. Therefore, predictions of unmethylated regions rely critically upon achieving high sequencing depths and more importantly by the use of appropriate controls (such as input DNA without enrichment or methylated and unmethylated DNA controls).

Me-DIP seq has been used to detect mtDNA methylation in post-mortem blood and brain specimens, and brain region-specific patterns of methylation could be identified[185]. Also, when testing valproic acid toxicity on primary hepatocytes, Wolters and colleagues revealed using Me-DIP seq that some methylation modifications they identified on mtDNA were reversible after a 3-day washout period[186].

1.6.5 Single-molecule long-read sequencing

Among the most recent developments in sequencing technologies there is the sequencing of very long fragments as single-molecule DNA (single-molecule long-reads sequencing). These technologies have been developed by PacBio technologies, under the form of PacBio SMRT sequencing and by Oxford Nanopore Sequencing (ONS)[187]. The main advantage of these two methods is that they are able to sequence native, DNA (i.e.: not treated with any chemical reagent, including bisulfite, nor amplified with PCR). Therefore, they are able to avoid the biases introduced by bisulfite-induced DNA degradation and of the subsequent PCR amplification, while still collecting information at the individual cytosine level. Recent advancements in software development, in particular neural network technology, now

also allow the identification of modified bases from the raw signal coming from PacBio and ONS data[188,189]. These modifications are usually 5mC or 6-adenosine methylation (6mA), for which models are already provided by the software developers, although it is common that new models can be trained on *ad-hoc* controls to recognise other kinds of modified bases.

A major advantage of long-read sequencing is the potential to phase epigenetic and genetic information, providing allele-specific 5mC patterns that allow insight into the effect of mutations, structural variants, or parental origin on gene regulation[190,191].

A downside of this approach is that since PCR amplification needs to be avoided to collect information on modified bases, a large amount of input DNA is required to obtain enough data (PCR can however still be performed in simple WGS approaches).

1.6.6 PacBio SMRT sequencing

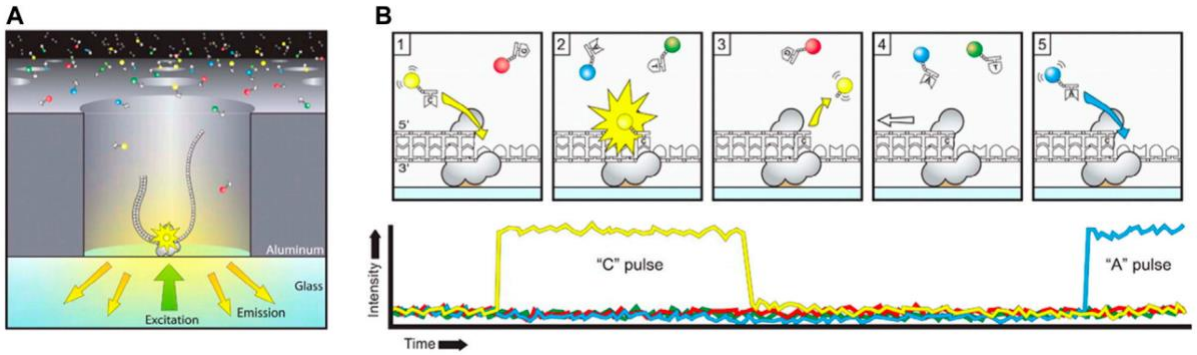


Figure 1.6: PacBio SMRT sequencing principle. a) A schematic diagram of the SMRT sequencing. The adaptor binds to a polymerase immobilized at the bottom of a zero-mode waveguide, where the light excitation and emission occurs. b) (top) Sequencing-by-synthesis procedure; (bottom) graph showing PacBio raw signal intensities over time. Adapted from Rhoads and Au, 2015[192]

The key technology at the hearth of SMRT sequencing are the so-called *zero-mode waveguides*, small nanowells that are able to contain a single DNA polymerase fixed at the bottom. In each of the nanowells a circularised DNA strand is then inserted, and the original DNA sequence is reconstructed by registering fluorescent pulses over time, emitted when a different fluorescently labelled nucleotide is incorporated by the polymerase (sequencing-by-synthesis)[187]. Therefore, with SMRT sequencing modified bases do not affect the basecalling of the sequence, but instead they influence the kinetics by which the polymerase incorporates the complementary

base[188]. By analysing inter-pulse durations, base modifications can be inferred by comparison to an *in silico* model or to an unmodified template[188]. The various patterns of perturbations depend on the genomic context in which each modification is inserted, and the magnitude of the perturbations of the polymerase kinetics depend on the modification itself[188].

Because of the way PacBio sequencing detects base modifications, the signal/noise ratio is usually low, and it varies greatly according to the modification analysed. Therefore, high read depths are recommended, for example 25x for 6mA or even as high as 250x for 5mC and 5hmC (unless they are enriched or modified beforehand to produce an even greater polymerase incorporation shift)[193]. Because of this issue SMRT sequencing typically only achieves single-molecule resolution for certain modifications, and using relatively short fragments (≤ 2 kb), which have to be read a large number of times by the polymerase[190].

1.6.7 Oxford Nanopore Sequencing (ONS)

1.6.7.1 Overview of the technology and nucleotide calling

In ONS, a single strand of a target DNA sequence is unwound through a protein pore by the action of a helicase placed at the edges of the DNA sequence by *ad-hoc* sequencing adaptors. The flow of the DNA strand through the pore interrupts an ion current which is maintained across the synthetic membrane in which the pore is embedded, and variations in the resulting electric signal are registered over time. Neural network-based algorithms are then capable to reconstruct the original DNA sequence by analysis of the electrical current variations, in a process called basecalling. Recent advances in basecalling technologies have enabled the identification of modified bases by their electric signature. In a typical bioinformatic workflow using ONS there are usually 3 steps: (i) basecalling with canonical bases only; (ii) anchoring the raw electric signal to a genomic reference and (iii) comparing the raw signal to a model to assess the probability of a specific base in that particular genomic context to be modified or not.

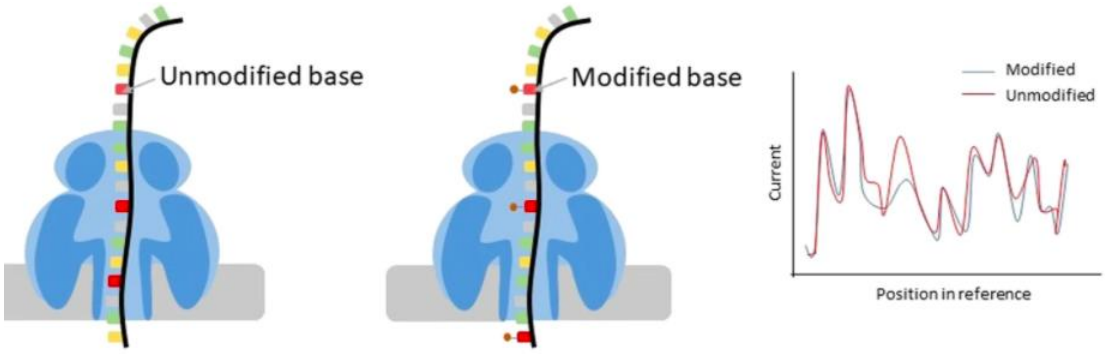


Figure 1.7: Principle of Oxford Nanopore Sequencing-based modified bases detection. On the left a diagram of a Nanopore is shown. When a DNA strand with unmodified bases passes through the nanopore, this produces a current registered over time, represented by the blue line in the diagram on the right. When a modified bases is sequenced, this can be seen as different in the electric current over time, represented as the red line in the diagram on the right. Adapted from Xu and Seki, 2019[189]

1.6.7.2 ONS accuracy and error rates

Accuracy of basecalling can be assessed at the read level or at the consensus sequence level. The former measures the identity of individual basecalled reads to a known reference, while the latter measures identity of a consensus sequence constructed overlapping multiple reads coming from the same genomic location, and increases with the read depth (i.e.: accuracy of a consensus built using 10 reads is lower than one built with 100 reads). There may be a correlation between read and consensus accuracy, although more accurate reads do not necessarily produce more accurate consensuses. In fact, since random errors occur in the minority of their reads at their locus, it is unlikely that they will appear in the consensus. Therefore, Low accuracy reads are able to produce perfect consensus sequences, assuming their errors are random and the read depth large enough. Vice versa, high accuracy reads can produce an imperfect consensus if they contain systematic errors, irrespectively of the read depth.

One of the main drawbacks of ONS is the elevated error rate compared to all of the other sequencing technologies. The most common errors found in ONS are insertions and deletions[194]. These are caused by the fact that, on one side the helicase that unwinds DNA through the nanopore does not do so always at a constant rate, and this results in more deletions when translocases speeds are high[194]. On the other hand, a major issue lies in homopolymeric repeats, which generate constant current alterations over time, and are challenging to interpret by the basecalling software[195] .

Moreover, as with ONS it is possible to use native DNA or RNA, the accuracy can be further altered by the presence of secondary modifications like CpG methylation.

In ONS, read and consensus accuracies depend on, and are improved with the advancement of both the sequencing chemistry (i.e.: better translocases, shorter pores, etc.) and the basecalling software. A characteristic of ONS since the company was formed in 2015 is the very fast rate at which these two aspects have improved[196]. Because of this and all the variables on which the accuracy determination depend on, to get an update estimate of the ONS accuracy, multiple aspects have to be considered, including pore and sequencing kit versions, basecalling software version and sample type (DNA/RNA, native/PCR).

Currently (February 2022), in latest technology update by ONS, the company claims to achieve a raw read accuracy of 99.3% (https://nanoporetech.com/accuracy), which is higher than the maximum theoretical accuracy that could be achieved with the kit, pores and basecalling software version that were used in this study (which was performed between 2017-19). However, this accuracy is still very far from the levels that can be currently achieved with Illumina sequencing (>99.99%)[197].

Nevertheless, because of the constant rate of ONS technology improvements, it is advisable to consider re-analysing old sequencing data using the latest release of the basecalling software.

1.6.7.3 Methylation calling with ONS

Nanopolish[198] is one of the most widely used software for modified bases detection using ONS. It shows a good correlation with WGBS regarding the calling of 5mC in mouse and human genomes[199]. Nanopolish already possess a model to assess 5mC, so usually the comparison with a non-methylated and methylated control is not needed. The output of Nanopolish is a log-likelihood ratio quantifying the probability for a base to be modified[198]. Other software for 5mC detection have been developed by independent laboratories, such as signalAlign[200] mCaller[201], DeepSignal[202], DeepMod[203], as well as software developed by ONS technologies, such as Megalodon (https://github.com/nanoporetech/megalodon) or Tombo[204]. 6mA identification is usually less accurate than 5mC, and depends greatly on the development of accurate positive and negative controls. This holds true also for other modifications in general, since for software (including Nanopolish) which model a previous knowledge on the expected modification signal, the use of *ad-hoc* controls is fundamental[201]. These

controls are typically PCR amplicons either left untreated (negative control) or either synthesised chemically or treated *in vitro* with a modifying enzyme (such as a methyltransferase)[198,201].

Very recent advances of the basecalling software (after Guppy v3.2.1) now allow the direct basecalling of the modified bases on the raw electric signal. Despite still being restricted to 5mC in the CG and CC(A/T)GG contexts and 6mA in the GATC context, this technology is very promising as it avoids complicated downstream bioinformatic analyses.

Despite this, the detection of DNA modifications with ONS is still in development, as we do not yet know the full extent of the modifications that can be observed, the sensitivity limits for each detection or ways to detect more than a modification at the same time. Moreover, ONS technologies frequently updates both the pore chemistry and the basecalling algorithms, so it is virtually necessary to re-train the algorithms at every upgrade.

1.7 Mitochondrial DNA methylation

As outlined in **section 1.3** and **1.4**, DNA methylation has been studied in detail in nDNA. In parallel to these studies, short after the discovery of mtDNA many groups have researched the presence and the possible role of 5mC methylation in mtDNA as well (see **paragraph 1.7.1**). However, contrary to what has been established for nDNA, there is still not agreement over not only the role but also the existence of relevant methylation on mtDNA.

1.7.1 Early evidence

In 1971, the group by Vanyushin and colleagues reported activity of *DNMT1* in mitochondria of loach embryos[205]. This was the very first evidence that suggested that mtDNA could be methylated. The first study that measured 5mC levels in mtDNA was performed in 1973 by the same group that discovered mtDNA[170]. Using mass spectrometry, they were able to demonstrate that 5mC was the only DNA modification present in mtDNA of various cell lines, but they reported levels well below those present in nDNA, at around 2-5% methylation[170]. Also, because of the technology used it could not be possible to determine sequence information, nor they could identify which methyltransferase was responsible for the maintenance of the observed low

methylation levels. These results were also challenged by a report published shortly after, dismissing the observed methylation as an artefact[206]. These initial reports contributed to picture mtDNA as essentially deprived of significant methylation, a view that would endure for decades. Despite this, other groups were reporting *DNMT1* activity in mitochondria isolated from beef heart[207], and later it was also reported the 5mC mitochondrial methylation profile in tissues of various species[171]. However, the main view remained that mtDNA was not methylated, thus the few studies that followed these early reports confirmed the reported low levels of methylation, although without reporting sequence information[208,209].

1.7.2 Recent studies

In 2001, the first major breakthrough in mtDNA methylation research came from a study performed by Shock and colleagues[210]. When analysing the *DNMT1* sequence, they noted the presence of additional ORFs, one of which producing a peptide with possible mitochondrial targeting. After establishing that the peptide could produce a mitochondrially-targeted *DNMT1* isoform, they arguably show presence of *DNMT1* in mitochondria by fractionation and WB. Then, they established that the mitochondrially-targeted *DNMT1* was associated to mtDNA, and further identifying some levels of 5mC associated with mitochondrial genes by immunoprecipitation. They also reported presence of 5hmC in mtDNA for the first time[209]. Recently, another study showed that the mitochondrially-targeted *DNMT1* corresponds to the isoform 3 of *DNMT1*[211]. This report paved the way for a new wave of research focussed on identifying mtDNA methylation and its possible role.

1.7.2.1 Environment effect on mtDNA methylation

Several studies have been conducted on a Belgian cohort of mothers and new-borns (ENVIR*ON*AGE cohort[212]) over the years, which explored the effect on babies and mothers of environmental stressors. For example, they explored the effect of pregnancy smoking or exposure to pollutants during pregnancy on different phenotypic variables on the newborns, including mtDNA methylation. Within this framework, the researchers identified over the years differences in the methylation of mitochondrial genes *MT-RNR1*, *MT-TF* and the D-loop between patients and controls[213–219]. Another study measured whether mtDNA methylation was affected in

the newborn by intrauterine growth restriction and preeclampsia, and found that mtDNA levels were increased in all pathologic groups compared to control, while D-loop methylation was further decreased in the most severe cases and associated to umbilical vein pO_2. *MT-CO1* methylation levels were inversely correlated to mtDNA content[220].

1.7.2.2 mtDNA methylation research in cancer

Another field where mitochondrial DNA methylation was extensively investigated is cancer research. Although an initial pre-2011 study failed to identify any relevant methylation in cancer cell lines[221], shortly after the report by Shock and colleagues[210], mtDNA methylation was found associated with the L1 region of the papillomavirus HPV16 in samples from infected patients, prompting the suggestion that it could be used as a marker for precancerous and cancerous cervix disease[222]. Studies conducted on colon cancer samples revealed instead that demethylation of the D-loop region in patients possibly modulated mtDNA copy number *MT-ND2* expression, facilitating cancer growth[223–225]. In a study that took both nDNA and mtDNA sequence differences into account, various CpG residues distributed across the whole mitochondrial molecule were differentially methylated in cell lines derived from glioblastoma and osteosarcoma patients[226].

1.7.2.3 mtDNA methylation and ageing

Detailed analysis of a large number of human blood samples part of a cohort comprising all possible age groups revealed methylation of the *MT-RNR1* gene and the co-presence of both unmethylated and methylated cytosines in most samples. High methylation levels (>10%) were more frequent in old women with respect to younger controls[227]. A 9-year-long follow-up survey showed that subjects with high methylation levels exhibit a mortality risk significantly higher than subjects with low levels[227]. A similar analysis on a smaller and different sample group detected low and variable levels of mtDNA methylation at 54 of 133 CpG sites interrogated, with 12S ribosomal RNA gene showing an inverse correlation with subject age[228].

1.7.2.4 Role of mtDNA methylation in neurological diseases

The presence of mtDNA methylation was linked to motor neuron cell death, through *DNMT3a* upregulation[229]. *DNMT3a* (but not *DNMT1*) was found in mitochondria of

skeletal muscle and CNS of transgenic mouse models of amyotrophic lateral sclerosis, together with a peculiar 5mC pattern on mtDNA[230]. Post-mortem brains of patients with late-onset Alzheimer's disease showed decreased methylation levels compared to controls[231], while they were found increased in another study[232]. Loss of methylation in the D-loop was observed in substantia nigra of Parkinson's patients compared to controls[232]. A study based on the analysis of post-mortem brain tissues revealed region-specific patterns of mitochondrial DNA methylation[185].

1.7.2.5 Role of mtDNA methylation in stem cells research

The first study that examined mtDNA methylation presence in human stem cells was part of a bigger project aiming at mapping 5hmC distribution at a single-base resolution. Coincidentally, the authors found that the highest presence of non-CpG 5hmC presence was in mtDNA, although no explanation on the mechanism was investigated[233]. Another group showed that after inactivation of *Dnmt1*, *Dnmt3a*, and *Dnmt3b* in mouse embryonic stem (ES) cells, a reduction of the CpG methylation in the D-loop was observed, while the non-CpG methylation was apparently not affected[234]. This suggested that D-loop epigenetic modification is probably only partially established by those enzymes[234]. A recent report studying mtDNA methylation patterns during development reported methylation presence as early as soon after implantation, with DNMT1 as the main enzyme responsible for establishing and maintaining such modification [235].

1.7.2.6 mtDNA methylation and its identification in other diseases

A 2015 study showed both in an *in vitro* model and in human retinal microvasculature from donors with diabetic retinopathy that the retinal mtDNA is hypermethylated in diabetes, and compared to other regions of mtDNA, the D-loop showed higher degree of methylation[236]. Also, *Dnmt1* appears to play an active role in mtDNA methylation, as its expression is increased in the mitochondria, and inhibition of *Dnmt1* by its siRNA ameliorated hyperglycaemia-induced decrease in mtDNA transcription and increase in apoptosis, suggesting a critical role of D-loop methylation in the development of diabetic retinopathy[236].

Clusters of methylated cytosines were described in the D-loop of senescent endothelial cells, where based on their position it was hypothesised that that could play a role in mtDNA replication rather than gene expression[237].

Another study tested the effect of the antiepileptic drug valproic acid *in vitro* on primary human hepatocytes and found that 7 mtDNA regions are transiently hypomethylated when cells are temporarily exposed to the drug[186].

In a cohort of overweight and obese patients platelet mtDNA was assessed for methylation to test whether it could be used as a predictor for cardiovascular disease (CVD). This study found that methylation of the *MT-CO1*, *MT-CO3*, and *MT-TL1* genes are strong predictors of future CVD incidence[238].

1.7.3 Evidence against mtDNA methylation presence

In parallel to the emerging evidence that supported the presence of mtDNA methylation, a number of publications emerged with a more critical approach to this research, questioning not only the role but the very existence of 5mC presence in mtDNA.

In 2013, a seminal study from Hong and colleagues was the first after the discovery of the mitochondrially-targeted *DNMT1* to challenge the view of a methylated mtDNA[239]. In their study, using sodium bisulfite DNA conversion they examined the same mitochondrial regions that had been identified by the Shock group in their original study, using the same cell line strain[210]. They failed to identify any methylation higher than 0.18% in any of the regions they analysed, even after enrichment of the CpG sequences using RRBS. They repeated their analysis in primary human cell lines and in publicly-available WGBS experiments, and once again they failed to identify any methylation above 1% across the whole mitochondrial genome[239].

Following on Hong and colleagues work, in 2017 a study by Mechta and colleagues expanded further the amount of evidence against mtDNA methylation presence[240]. Using WGBS to analyse mouse tissue and cell lines that were reported either containing or not methylation on mtDNA, they reported for the first time a strict correlation between the amount of unconverted (i.e.: possibly methylated) cytosines and their relative read depth. After having excluded the possibility of NuMTs contamination by careful analysis of the aligned reads, they hypothesised that this phenomenon could be explained by the mtDNA secondary conformation. Because of its circular nature, DNA can in fact form supercoiled structures[241]. This could in turn cause some specific mtDNA areas to be more affected to the sonication process at the beginning of the WGBS procedure, thus ending up being overrepresented

compared to the others more tightly wound to the supercoiled structure (and therefore less likely to ligate to the WGBS adaptors). To test this, authors digested mtDNA with BamHI (a restriction enzyme that cut mtDNA once) in order to linearise it before sonication. When analysing different mitochondrial regions, they found statistically significant differences between the methylation levels calculated on the same regions in digested Vs undigested samples. For example, in the D-loop (positions 6–298) of Lonza cells, the range of methylation identified was 0-4.8% in undigested samples, dropping to 0-0.6% after BamHI digestion[240].

Shortly after the publication of Mechta and colleagues, Olova and colleagues performed an extensive comparative analysis of the available WGBS protocols, to assess how the sequence coverage and methylation outputs are affected by: 1) BS-induced DNA degradation, 2) PCR amplification, 3) DNA modifications, and 4) incomplete BS conversion[242]. Their main result was that the bisulfite conversion step is mainly responsible for introducing sequencing biases, due to a selective and context-specific DNA degradation[243] and incomplete conversion efficiency, with subsequent PCR amplification only expanding biases already introduced previously. A major discovery was done on mtDNA as well: while assessing the effect of various bisulfite treatments on sequences with uneven C distribution across the strands, they found that mtDNA was affected in a major way, with more than 60% aligning only from the C-poor H-strand, as part of the reads from the L-strand are lost due to degradation. Based on these results, in **chapter 4** we will explore the effect that this bias could have on the mtDNA methylation detection.

Lastly, the work of Matsuda and colleagues extensively examined mtDNA using 3 different technologies, to look for traces of 5mC[169]. Initially, they performed WGBS on rat liver and brain mtDNA isolated from mitochondria preparations and linearised with BglIII. They failed to identify any significant methylation signal in all of the samples tested[169]. The same observation was confirmed when the analysis was done using a 5mC-specific restriction enzyme, McrBC, which failed to digest any of the mtDNA samples they analysed[169]. Finally, they used LC/MS on mtDNA isolated from mitochondria preparations and digested to single nucleosides. They then compared the obtained levels of methylation to standard curves to obtain an absolute quantification of the 5mC present in their samples. The measured levels of 5mC over dC were ranging from 0.3%-06%, demonstrating once again that the levels of methylation in mtDNA in their samples is extremely low.

Overall, this evidence seem to point towards a vision of mtDNA as generally unmethylated. However, what is still unknown is whether even the low methylation observed in these latter studies may have a functional relevance (for example, if it is concentrated on a few highly methylated molecules or if it is more diluted but present always in the same pattern). Therefore, there is a need of developing an accurate method for specifically analyse mtDNA methylation at single-base resolution.

1.7.4 Introductory final remarks

The idea of an epigenetic control of mtDNA has been fascinating to researchers since early after the discovery of the molecule itself. As we described in the previous paragraphs, this topic still remains a matter of debate, particularly because of the contradicting findings in the works that have been published until now regarding this topic.

However, its relevance could have far-reaching ramifications, ranging from control of mitochondrial gene expression in a variety of diseases to effects on mtDNA replication. None of these aspects have been studied in deep so far, with the majority of the published works exploring simple associations of methylated mtDNA patterns to diseased states or other conditions (**paragraph 1.7.2**).

Therefore, there is a need to investigate further whether these association evidence are indeed linked with a molecular mechanism that regulates mitochondrial gene expression or replication. The presence of DNMT enzymes (DMNT1 in particular), albeit disputed, seems to suggest that the former might be true.

However, before assessing the validity of these hypotheses, an essential step is the establishment of a reliable method to assess mtDNA methylation presence, since this is the key readout that is used to explore any role of DNMT1 or other possible molecular mechanisms behind the epigenetic regulation of mtDNA. This topic will be thoroughly explored in this dissertation, where we propose an innovative method based on long-read sequencing to assess CpG methylation on mtDNA.

Chapter 2: Aim of the work

Studying the presence of CpG methylation in the mitochondrial genome has potentially very important applications. Primarily, this epigenetic modification could play a role in regulating gene expression, mirroring its function on the nuclear genome. To understand the mechanisms by which this is regulated could be of central importance in expanding our knowledge on mitochondrial gene expression. Additionally, mtDNA methylation has already been suggested to be potentially used as a biomarker in various pathological contexts[244]. Therefore, it is crucial to possess the most up-to-date tool to study this epigenetic modification in a context such as mtDNA genomics which already possess intrinsic difficulties. This is an aspect which is still debated in the field of mitochondria epigenetics. While on one side presence of CpG methylation on mtDNA is being observed, others questioned such results based on technical problems in the technology used to analyse this modification. The aim of this work can be summarised in the following objectives:

(i) By analysing publicly available WGBS studies, we aim at highlighting any problems intrinsic with this technology when specifically investigating mtDNA methylation.

(ii) Using Oxford Nanopore Sequencing we aim to develop both a novel library preparation method targeting mitochondrial sequences and to analyse in detail any advantages/pitfalls intrinsic to this technology. The aim of this objective is to develop a tool that circumvents the technical issues we identified in WGBS for single-base methylation assessment.

(iii) Finally, by investigating the presence of mtDNA CpG methylation in a variety of human samples using our new method, the overarching aim of this project is to determine whether there is significant methylation on human mtDNA at single-nucleotide resolution.

Chapter 3. Materials and Methods

3.1 Cell culture

Primary and immortalised cell lines used in this study are listed in **Appendix 1**. Cells were maintained in DMEM high glucose (Gibco) with 10% foetal bovine serum (Gibco) and no antibiotics at 37°C in a humidified 5% CO2 atmosphere. Cells were grown until ~80% confluence in 10mm dishes (Corning). When ready, cells were placed under sterile conditions in a class II cabinet. Cells were washed with sterile DPBS (Gibco), then incubated with 0.05% trypsin (Gibco) for 5 minutes at 37°C. Cells were then collected in a 15 ml plastic tube, then centrifuged at 1500 g for 5 minutes. Old media was removed by aspiration. At this point, for routine cell passaging, pellets were resuspended in 1 ml of DMEM 10% FBS medium then split into 10 mm dishes, in a ratio varying from 1:3 to 1:10. For DNA extraction, pellets were washed once with PBS, then placed on ice. Resuspended pellets were then centrifuged at 10000g for 10 minutes at 4°C, then, after PBS removal, snap-frozen in liquid nitrogen and kept at -20°C until further use.

3.1.1 Cell counting

Human cell lines and primary fibroblasts were counted using a Countess II FL Automated Cell Counter (Thermo Fisher Scientific). The cell suspension was diluted with a 1:1 ratio of 0.4% Trypan Blue (Thermo Fisher Scientific) in 20 µl final volume. 10 µl of this solution was loaded into a Countess Cell Counting Chamber Slide and placed into the Countess II FL Automated Cell Counter to measure the number of cells/ml. Only the concentration of alive cells was used to calculate the number of cells for seeding.

3.2 DNA extraction using Qiagen kits

3.2.1 DNA extraction from cell pellets

All DNA from immortalised or primary cell lines, was extracted from snap-frozen pellets using the QIAmp DNeasy blood and tissue kit (QIAGEN) following the manufacturer's instructions on how to extract DNA from cultured cells. Briefly, pellets were thawed at RT, then resuspended in 200µl of PBS. This was followed by the addition of 20µl of Proteinase K and 200µl of AL lysis buffer. Lysed samples were then incubated at 56°C

for 10 minutes, then 100% EtOH was added to stop the reaction. The mixture was then added to a DNeasy Spin Column and centrifuged at 6000g for 1 minute. Eluate was discarded and 500µl of Buffer AW1 was added to the column. Samples were centrifuged at 6000g for 1 minute, then eluate was discarded. 500µl of Buffer AW2 was then added to the columns. Samples were centrifuged at 20000g for 3 minutes, then eluate was discarded and columns were transferred to a 1.5 Eppendorf tube. 200µl of PCR-grade H_2O was then added to the columns and column membranes were soaked for 1 minute. Finally, samples were centrifuged at 6000g for 1 minute, and this last step was repeated for a total of 2 times to increase DNA recovery yield.

3.2.2 DNA extraction from human tissues

DNA from human tissues was extracted using the QIAmp Fast DNA Tissue Kit (QIAGEN). A lysis buffer mastermix was prepared by adding the following reagents in the amounts described in **Table 3.1** below per each sample.

Reagent	Amount
AVE Buffer	200 µl
VXL Buffer	40 µl
DX Reagent	1 µl
Proteinase K	20 µl
RNase A (100mg/ml)	4 µl

Table 3.1: mastermix preparation for DNA extraction using QIAmp Fast DNA Tissue Kit

Lysis buffer mastermix was added to supplied Disruption Tubes (265 µl per sample), then under a class II biological safety cabinet ~25 mg of human tissues (Appendix 1) were added to each individual tube. Tissues were then homogenised by mechanical disruption by shaking on a vortex for 5 minutes. This was followed by incubation on a Thermomixer (Thermo Scientific) at 56°C for 10 minutes, shaking at 1000 rpm. If required, the last two steps were repeated once to increase tissue disruption. After homogenisation, 165 µl of MVL media were added to each sample, then mixed by vortexing. Samples were then added to individual QIAmp Mini spin columns, then centrifuged at 20'000 rpm for 1 minute. Eluate was discarded, then 500 µl of Buffer AW1 were added to the columns. Samples were centrifuged at 20'000 rpm, then eluate

was discarded. This was followed by addition of 500 µl of Buffer AW2 and further centrifugation at 20'000 rpm for 1 minute. Finally, columns were transferred to 1.5 ml Eppendorf tubes, and 75 µl of water was added to the membrane columns. After 1 minute, tubes were centrifuged at 20'000 rpm for 1 minute to recover DNA. This step was repeated for a total of 3 times to increase DNA recovery yield.

3.3 DNA quantification using Qbit dsDNA kits

All DNA was quantified using the Qubit dsDNA kit (Invitrogen), either Broad Range (BR) or High Sensitivity (HS) following the manufacturer's instructions (identical for the two kits). Working solution (WS) was prepared by adding the Qbit Reagent in Qbit buffer at a concentration of 1:1000. Volumes of WS were calculated by preparing 200 µl of WS per sample (considering an excess of 1 or 2 samples) and 200 µl per calibrating control. Controls were prepared by adding 10 µl of either Control #1 or Control #2 to 190 µl of WS into Qbit tubes (Invitrogen). Samples were prepared by adding 2 µl of sample to 198 µl of WS. Calibration was performed every time new samples were analysed. Calibration and sample reading were performed on a Qbit 2.0 instrument following the guidelines on the instrument for either Qbit HS or BR reagents. The choice between the two kits depended on whether the measured DNA concentration fell in the detection range of either kit.

3.4 Long-range polymerase reactions

Long-range polymerase reactions (LR-PCR) amplification reaction was performed using PrimeSTAR GXL DNA Polymerase kit (Takara) according to manufacturer's instructions. The primers used are detailed in **Appendix 2**. The resulting PCR product is an amplicon of 15412 bp length, covering the positions 1157-16569 of the human mitochondrial genome. Amplification reactions were performed using the following cycling conditions: 94°C for 1 minute, followed by 30 cycles of 98°C for 10 seconds, 55°C for 15 seconds and 68°C for 10 minutes.

3.5 Generation of negative and positive controls

Untreated LR-PCR amplicons were used as negative controls for methylation. To generate positive controls, the same amplicons were treated in vitro with the

recombinant CpG methyltransferase *M.SssI* (NEB). To find the optimal reaction conditions, various parameters were tested, as outlined in **chapter 5**. The final protocol is detailed here: 1 µg of amplicon DNA per 50µl reaction were treated for 4 hours at 37°C with 50 units of *M.SssI* in the presence of 1x NEB buffer #2 and 160µM of S-adenosylmethionine (NEB). To test the efficiency of the M.SssI reaction, 10 units of methylation-sensitive restriction enzyme BstUI were added at the end of the M.SssI incubation. This was followed by a further incubation at 60°C for 1 hour.
Protection of the M.SssI-treated amplicons from *BstUI* digestion was assessed using the Genomic DNA ScreenTape System (Agilent) on an Agilent 2200 TapeStation platform following manufacturer's instructions described below (**Figure 5.2**).

3.5.1 TapeStation

The Agilent 220 TapeStation system is designed to perform fast and automated DNA electrophoresis. The platform utilises various formats of ScreenTape reagents, designed to separate different DNA sizes and at a variable sensitivity. In this study, only Genomic ScreenTape was used. Genomic ScreenTape is designed to analyse fragments ranging from 200 to > 60'000 bp. Also, Genomic ScreenTape provides an estimation of objective genomic DNA integrity by calculating a DNA Integrity Number (DIN), which ranges from 0 (highly fragmented DNA) to 10 (intact DNA). DIN was used to assess DNA integrity before each Nanopore sequencing experiment, and only DNA with a DIN ≥ 9 were processed for further analysis. An internal control ladder is provided and analyse together with experimental sample, to which sample peaks are compared to identify their size.
To analyse samples using Genomic ScreenTape, all reagents were taken out at RT and let equilibrate for 10 minutes. Then, Genomic ScreenTape Sample Buffer (Agilent) were added to 0.2 µl plastic PCR tubes, 10 µl per sample. The number of samples to analyse at one time needs to always be odd, since the Agilent 220 TapeStation system may only analyse couples of samples at the time, and the ladder must always be included with the samples. Then, 1 µl of Genomic DNA ladder was added to the first tube, followed by 1 µl of experimental sample per every other tube. Tubes were then briefly centrifuged, then vortexed at 2000 rpm for 1 minute. Meanwhile, Genomic ScreenTape was added to the Agilent 220 TapeStation system, together with the sample plate holder and tips holder. Details of the experiment were added using the

Agilent 220 TapeStation controller software, and tips were added in the tips holder following the pattern specified in the software. Once the vortexing was over, tubes were briefly centrifuged then added into the tubes holder, after carefully removing their lids (if present) using a scalpel. Agilent 220 TapeStation system was then closed and the experiment started using Agilent 220 TapeStation controller software.

3.6 Mitochondrial DNA enrichment for single-molecule sequencing

3.6.1 Exonuclease V-based protocol

Following the protocol developed by Jayaprakash and colleagues[245], 1µg or 4 µg of gDNA were digested with Exonuclease V (NEB) for 48 hours at 37°C. For each reaction 10 enzymatic units per µg of DNA were used, in addition to 0.3 mM of ATP and 1x NEB buffer 4 (provided with the enzyme). AMPure beads (Beckman Coulter) were then used to isolate high molecular weight digestion products and to remove the enzyme (see **section 3.4.2**).

To purify digested DNA, AMPure beads were used. The principle for the size selection used in AMPure beads purification is based on the concentration of the negatively charged DNA around the magnetic beads, in the presence of high salt and PEG concentrations (contained in the beads solution). Therefore, at low beads/DNA concentrations (corresponding to low salt and PEG levels) only the high molecular weight DNA fragments will be concentrated with the beads and eluted out.

To achieve purification, digested gDNA was transferred into 1.5 ml Lo-Bind microcentrifuge tubes (Eppendorf). A beads/DNA ratio of 0.5 was used to enrich for fragment of the desired molecular weight. Mix were incubated at RT for 10 minutes and beads were pelleted using a DynaMag-2 magnet (Invitrogen). Bead pellets were washed twice with 750 µl of 70% ethanol (vol/vol). Bead pellets were air-dried for 5 minutes and resuspended in 25 µl of DNase-free water (Ambion) and incubated at 37°C for 5 minutes. Tubes were placed on a DynaMag-2 magnet (Invitrogen) and supernatant was collected and transferred to new Lo-Bind 1.5 ml microcentrifuge tubes (Eppendorf).

3.6.2 BamHI-based protocol

The protocol for mtDNA enrichment Is based on the simultaneous linearisation of mtDNA and its isolation together with high molecular fragments. To achieve this, 2 µg of genomic DNA (nuclear + mitochondrial DNA) per 50 µl reactions were digested with 40 units of the recombinant restriction enzyme BamHI-HF (NEB) for 1 hour at 37°C in the presence of CutSmart buffer (NEB).

To achieve combined DNA purification and selection of high molecular weight fragments, DNA was first purified using a ratio of AMPure beads/DNA of 0.1x, 0.3x and 0.5x, following the protocol described above.

At a later stage, purification and selection was achieved using the Monarch® PCR & DNA Cleanup Kit (NEB), following manufacturer's instructions. Binding Buffer was added at a ratio of 2:1 buffer:sample (100 µl buffer per 50 µl reaction). The mixture was then added to the PCR Monarch® PCR & DNA Cleanup columns, followed by centrifugation at 16'000 g for 1 minute. Eluate was discarded then 200 µl of Wash Buffer was added to the columns, followed by centrifugation at 16'000 g for 1 minute and eluate removal. This step was repeated once, after which columns were transferred to 1.5 ml Eppendorf tubes. The last step was modified following the recommended protocol modification to enrich for long DNA fragments: 20µl of Elution Buffer was heated to 50°C, then added to the columns. Membranes were soaked for 1 minute, then columns were centrifuged at 16'000g for 1 minute and eluate was retained.

3.7 Quantification of mtDNA levels using ddPCR

Droplet Digital PCR (ddPCR) was used to quantify relative mtDNA enrichment following BamHI-HF (NEB) treatment of control DNA. This technique has the advantage to enable the quantification of the absolute mtDNA copy number without relying on a standard curve. On an average run, samples are fractionated to form ~10'000-20'000 droplets, where individual PCR reactions take place. To quantify relative mtDNA copy number[246], a mitochondrial and nuclear target (the genes *MT-ND1* and *RNASE P*, respectively) were amplified and fluorescent signal was generated using the primers and probes detailed in **Appendix 2**. ddPCR protocol was performed following manufacturer's instructions. PCR reaction master mix was prepared in 1x (final concentration) ddPCR Supermix for Probes (No dUTP, BioRad),

by adding 300nM of each primer and 200nM of each probe in 19µl final volume. Then, 1 ng of sample DNA was added to the mastermix. Droplets were generated using an Automated Droplet Generation instrument (BioRad) in a 96 well plate (Bio-Rad). After droplet generation, the plate was sealed with foil using a PX1 PCR Plate Sealer (Bio-Rad). Droplets were quickly subjected to PCR amplification, performed using the following cycling conditions: 95°C for 10 minutes, followed by 39 cycles of 94°C for 30 seconds and 58°C for 1 minute, followed by a final stabilisation step at 98°C for 10 minutes. Droplets were then loaded into a QX200 droplet reader (BioRad) and analysed using an absolute quantification protocol (ABS) to measure the absolute copy number of each probe. Droplet analysis was performed using the QuantaSoft analysis software (BioRad) that determines if droplets are positives or negatives. The separation threshold was adjusted manually if necessary. Results were represented as a ratio of *MT-ND1/RNASE P* copy numbers.

3.8 ONS library preparation and sequencing on the MinION instrument

Approximately 1 µg of native genomic DNA (from cell lines or human tissues, **Appendix 1**) or purified LR-PCR amplicons were prepared for ONS sequencing on R9.4.1 flow cells using the Ligation Sequencing Kit SQK-LSK109 (Nanoporetech), in combination with the Native Barcoding Expansion Kit EXP-NBD114 (Nanoporetech). Genomic DNA was fragmented either through BamHI digestion (see **paragraph 3.6.2**) or sheared to 10 kbp using g-tubes (Covaris), following manufacturers' instructions, while amplicons were left untreated. To shear DNA using g-tubes, 50µl of DNA were placed inside the g-tube column, then centrifuged at 6000 rpm for 1 minute on an Eppendorf 5424 centrifuge. The column was then flipped and centrifuged again at 6000 rpm for 1 minute on an Eppendorf 5424 centrifuge. Sheared DNA was collected from the g-tube cap.

Simultaneous DNA repairing, end-repairing, and dA-tailing was achieved using the NEBNext FFPE Repair Mix (NEB) and the Ultra II end-repair module (NEB). DNA repair is an optional step which increases DNA quality after sequencing. DNA end-repairing allows the addition of a few nucleotides at either strand on both ends of a DNA sequence to obtain a blunt sequence end. To this, dA-tailing adds an adenine to one of the DNA strands to allow the subsequent ligation of an additional

barcode/adapter, both of which will have a thymine at one of their strands ends for complementarity. Samples were then incubated at 65°C for 5 minutes then 20°C for 5 minutes in a thermocycler. Barcodes (part of the EXP-NBD114 kit) were ligated to individual samples using Blunt/TA Ligase Master Mix (NEB), by incubation at RT for 10 minutes. Barcodes are short DNA sequences with a known nucleotide pattern and they allow the labelling of a specific sample. This allows the pulling together of several samples in one sequencing experiment.

Samples were then combined and AMII adapters were ligated using NEBNext® Quick Ligation Module (NEB), by incubation at RT for 10 minutes. AMII are special adapters containing the motor proteins needed for sequencing using Nanopore technology. These motor proteins are helicases which can separate the two DNA strands while attached to the Nanopore protein pore.

AMPure XP beads (Beckman Coulter) at a concentration of 1x, 1x and 0.5x, respectively, were used to purify DNA between the library preparation steps. Final libraries were loaded onto R9.4.1 flow cells and samples were sequenced using a single MinION Mk 1B, together with sequencing buffer and sequencing beads (part of the SQK-LSK109 kit). To keep the sequencing throughput consistent, where possible a maximum of 6 biological samples were pooled and sequenced for 24 hours (**Appendix 5**) . LR-PCR amplicons were pooled and sequenced for 6 hours.

For all the experiments, live basecalling was turned off and only raw signal data was collected. Mux scans were performed every 6 hours. In a R9.4.1 MinION flow cell the sequencing pores are divided in 4 sequencing groups based on sequencing performance. Mux scans allow the periodic reset of these 4 groups based on the changing performance of the sequencing groups over the course of the experiment.

3.9 WGBS data analysis

3.9.1 Data download

Raw WGBS experiments part of the Roadmap Epigenome Project[247] were downloaded from the GEO Database. Downloaded files from single-ended WGBS sequencing experiments were converted from SRA (Sequence Read Archive) format to fastq files using `fastq-dump` (**Appendix 4**) with the following options: `--readids --skip-technical -W --read-filter pass --gzip`. Respectively, these options allow to: append the read id after the spot id as 'accession.spot.readid'; dump

only the biological reads; clip adapter sequences; optionally filter reads by value: “pass”; compress output files using gzip. SRA is a public repository of data which contains thousands of high-throughput-sequencing experiments, usually short-reads-based sequencing experiments (typically less than 100bp).

3.9.2 Quality control and trimming

Read quality of the converted fastq files was assessed with FastQC v0.11.5[248]. FastQC is a commonly used software which provides a quick way to assess the quality of the raw data files. Specifically, in this study it was used to assess the fastq files extracted from the downloaded SRA files. Relevant graphs are generated and can be visualised on an html file.

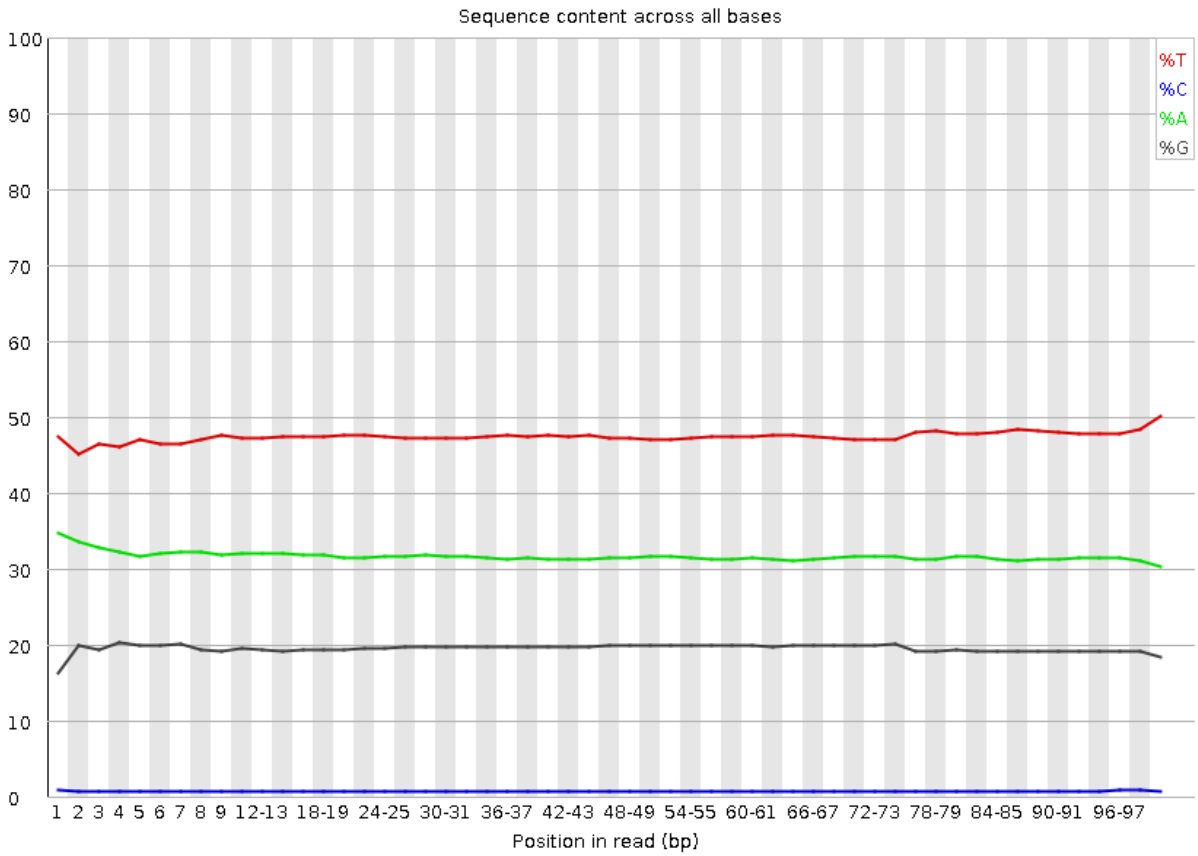


Figure 3.1: Example of a FastQC sequence content per position graph. Graph showing the average percentage (y axis) per position (x axis) of each of the 4 bases (different colours) across all the reads of a WGBS fastq file.

All the reports generated from FastQC were manually checked to determine whether a trimming of low-quality reads and/or adapters was needed. Where trimming was

deemed necessary, TrimGalore! v0.4.5[249] was used. This software automatically trims adapter sequences from the reads (if present) and retains those with an average Phred quality score ≤ 20 (before and/or after trimming). Reads shorter than 45 bp after trimming were discarded using the `--length` option.

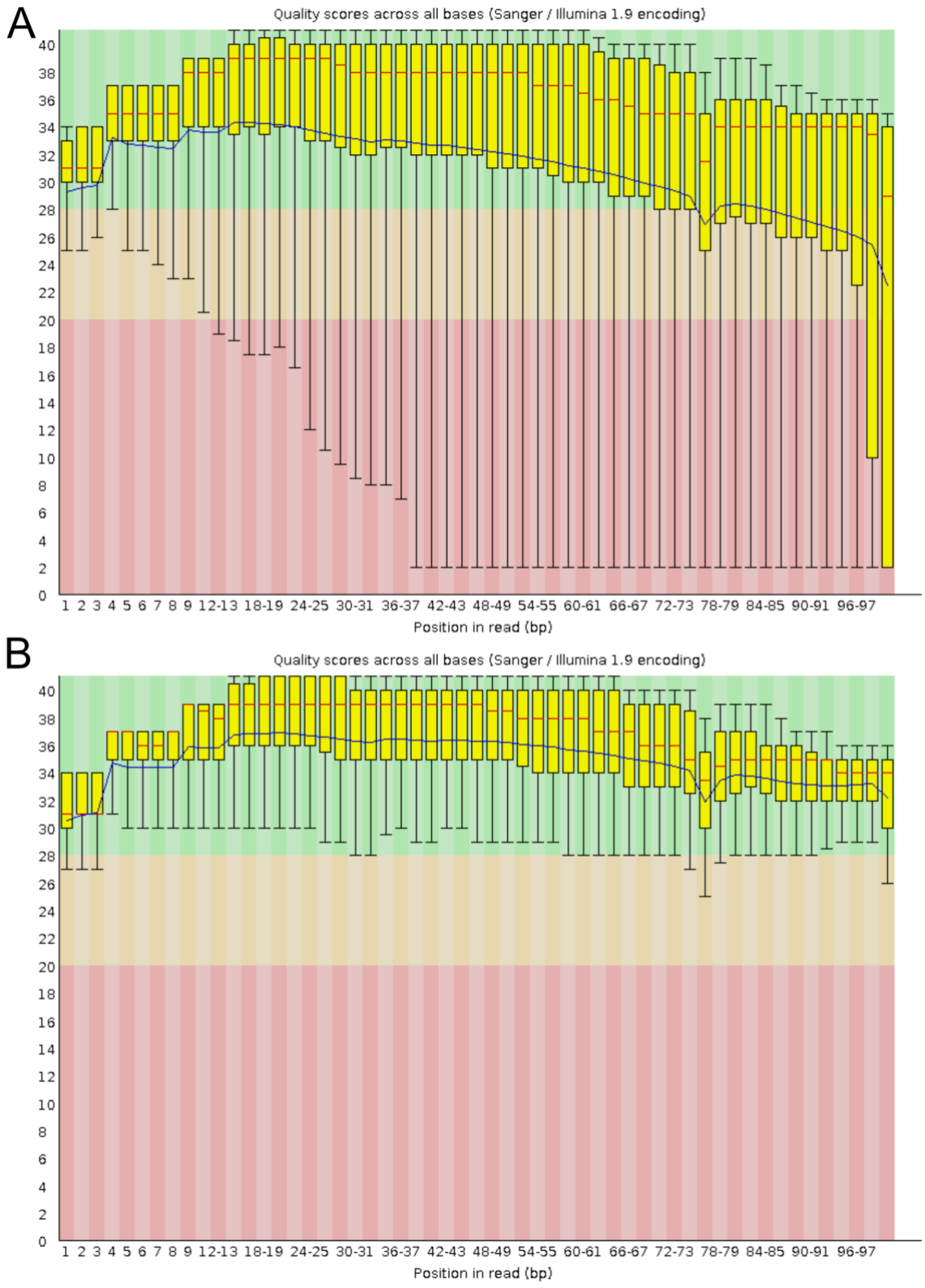


Figure 3.2: Comparison of pre- and post-trimming quality scores per position distributions in WGBS experiments. **The graphs show the distributions of average quality scores per position(s) in one WGBS sample, a) before and b) after the trimming by TrimGalore!.**

Y-axis represent quality scores expressed on the Phred scale; X-axis represent the position(s) on the read in basepairs.

3.9.3 Reads alignment

Alignment of the sequenced reads to a reference sequence is necessary to identify where on the genome a read should be assigned to. Upon quality check and trimming, both alignment of the WGBS fastq files to the reference human genome sequence (GRCh38) and extraction of the methylation information were carried out with bowtie2 v2.3.2[250] and Bismark v0.19.0[251], respectively. Coverage was calculated from BAM files using samtools depth[252]. This was defined as the percentage of mtDNA genome in each strand covered by at least 5 reads.

3.9.4 Calculation of the methylation levels

Extraction of the methylation information was performed by Bismark by comparing which cytosines were converted to thymines in the sequenced read after bisulfite treatment (because of their unmethylated status) and which instead remained as cytosines (because they were originally methylated). This information and the sequencing context where this modification is found (i.e.: CpG or non-CpG context) were reported in an additional column in the BAM file which is generated after alignment.

Methylation extraction was carried out from the BAM file using the `bismark_methylation_extractor` package with the following options: `--comprehensive --merge_non_CpG --gzip --bedGraph --CX_context`. These set of options reported information on cytosines in both CpG and non-CpG context, but for the purposes of this study only CpG residues were considered for further analyses. The generated files were further processed using custom scripts to extract information about the mitochondrial genome alignment bias, coverage, etc.

3.10 ONS data analysis

3.10.1 Base calling

Raw signal from ONS experiments is a collection of information on the variation of the current in each flow cell pore over time, stored in fast5 files. To convert these signals to sequence information, the software `Guppy` utilises a machine-learning algorithm to

calculate a probability for each variation to correspond to one of the 4 DNA bases (recent advances allow the identification of modified bases directly at this level, but this was not performed in this study)[198]. Base-calling of fast5 files containing raw electric current information was performed by the `guppy_basecaller` package of Guppy v3.2.2+9fe0a78 (Nanoporetech). Base-called, barcoded reads were de-multiplexed into individual samples using the `guppy_barcoder` package of Guppy v3.2.2+9fe0a78 (Nanoporetech).

Demultiplexing is performed by aligning the initial bases of each basecalled read (corresponding to the expected length of an ONS barcode) to a reference list of barcodes sequences. A read is assigned to a barcode (i.e.: to a different sample) when a complete alignment is found and the read is then added to fastq files in dedicated custom repositories. Unaligned or partially aligned reads are excluded and collected in fastq files in an "`unassigned`" repository.

3.10.2 Reads alignment and quality check

To simultaneously enrich for linear full-length mitochondrial sequences, exclude ligation artifacts and minimise the presence of NuMTs, we applied a stringent filter on read sequence length (for LR-PCR controls: min=14000 bp, max=17000 bp; for biological samples: min=4000 bp, max=17000 bp) and quality (Phred quality score ≥ 9) using the software NanoFilt v2.2.0[253] on the barcoded fastq files (**Figure 5.17**). The minimap2 v2.10-r761[254] software was used to perform the alignment of Nanopore reads onto the GRCh38 reference (which includes the mitochondrial rCRS reference sequence, NC_012920.1), specifying the `-x map-ont` option. Secondary alignments (when a read completely aligns both to two genome regions) were identified in the BAM files by specific flags (256, 272) and excluded. Also, because of the length of the ONS reads, it is possible that reads partially align to two or more genomic regions. This are defined as "supplementary alignments" and identified in the BAM file by "SA" flags. Supplementary alignments represent a risk of NuMTs contamination when one of the alignments is on the nuclear DNA and the other on the mtDNA. To avoid this risk, we excluded all these cases aligned both on nuclear and mtDNA and marked by the "SA" flag. In the case of fragmentation experiments, we also identified supplementary alignments aligning only to the mtDNA. This is an artifact due to the circularity of the mtDNA and it is typical of reads which span the D-loop (with 16569-0

bp boundaries), that minimap2 does not recognise as circular sequences. Using custom scripts we therefore retained all such supplementary alignments, but only if they aligned in the same orientation on the same strand (H- or L- strand).
Similarly, to avoid the same issue with reads spanning the BamHI cut site in the ND6 gene (14258-14259 bp of the mtDNA reference sequence), we created an alternative rCRS (or sample-specific, see **paragraph 6.2.2**) mitochondrial reference sequence with a modified starting site at base 14259 instead of base 1. All the experiments where the samples were digested using BamHI were aligned to this alternative sequence (gene annotations were adapted accordingly). Quality control plots and sequencing statistics of aligned reads were automatically generated using NanoPlot v1.13.0[253].

3.10.3 ROC curve generation

ROC curve generation was performed by Dr. Claudia Calabrese. We calculated a ROC curve to assess the accuracy of our CpG methylation calling, using a previously published procedure[198]. To do this, we randomly chose 50,000 mtDNA CpG sites from positive and negative controls and classified each CpG call as true positive (TP) or false positive (FP), depending on which of the two controls each site came from and on whether methylation fell above or below a log-likelihood methylation threshold. We repeated the TP and FP calculation by varying log-likelihood threshold values within a range of -20 to 20 (to build the ROC curve) and 0 to 10 (to calculate accuracy, intended as the proportion of true calls, either TP or true negatives (TN), with a step of 0.25, as explained by Simpson and colleagues[198]

3.10.4 Mitochondrial variant calling of ONS samples

Variant calling was performed by Dr Claudia Calabrese. Because Nanopore technology allows a simultaneous read of epigenetic modifications while sequencing the target DNA, we performed a mitochondrial variant calling on the fastq files filtered with NanoFilt v2.2.0[253]. For this we used a modified version of the MToolBox pipeline[255], a workflow developed to analyse mtDNA from high-throughput sequencing data, which was adapted to long-reads sequencing analysis and is available in the Github public code repository (https://github.com/mitoNGS/MToolBox/tree/MToolBox_Nanopore).

For reads mapping we used the GRCh38 human genome assembly. For variant calling, we set a minimum Phred quality score (QS) threshold to retain variants to 10 (using the `-q` option of the `assemblyMTgenome.py` script). Variants with a read depth per position ≥ 30x and heteroplasmy ≥ 10% were retained). Finally, we performed haplogroup predictions, automatically generated by the MToolBox pipeline using a consensus FASTA sequence with all major alleles found in each sample[255] compared against the human phylogeny annotated in the Phylotree build 17[256] (**Appendix 6**). Moreover, we used Haplogrep 2 v.2.1.1[257] as additional tool to confirm individual haplogroup predictions, by running it on MToolBox-generated VCF files including only homoplasmic variants (with heteroplasmy ≥ 90%).

3.10.5 CpG methylation detection

Detection of methylation in CpG context was carried out using Nanopolish v0.11.0 call-methylation package[198]. In a similar way to Guppy basecalling, Nanopolish utilises a trained Hidden Markov Model to detect modified bases by comparing raw electric signals of modified/unmodified cytosines with expected signal from a reference sequence. The methylation calling output is a log-likelihood ratio where a positive value indicates evidence supporting methylation. Nanopolish utilises as input fast5 files containing raw electric signal information, basecalled fastq files and BAM alignment files, to generate an index file used by the algorithm to determine methylation Log-likelihood ratios. Log-likelihood ratios were then converted to a binary methylated/unmethylated call for each read, then percentage of methylation was obtained by calculating the fraction of methylated reads, using the `calculate_methylation_frequency.py` script available with the Nanopolish package. After accuracy determination using positive/negative controls, the default calling threshold of ≥ 2.5 LLR was modified to a more stringent ≥ 5 LLR to increase the accuracy of the call by modifying the script. Since Nanopolish groups neighbouring CpG sites and calls them jointly, CpG sites in the same group were separated and assigned the same methylation frequency using the `-s` option.

3.10.6 CpG methylation analysis

We applied a series of stringent quality filters to remove possible artefacts of the CpG methylation calling and errors introduced by the Nanopolish algorithm. We first

removed in all samples those CpGs calls that had a methylation frequency greater than two standard deviations from the methylation frequency mean in negative controls (false positives, **Appendix 3**). We also removed: I) all calls supported by less than 60 reads (since after analysing the relationship between read depth and methylation level we decided that this was the minimum acceptable read depth threshold to obtain reliable methylation results); II) calls with methylation frequency similar to the background (i.e. with a methylation frequency ≤ 0.5%, which is the average methylation frequency observed in the negative controls) and III) calls neighbouring any heteroplasmic nucleotide variant (i.e.: with heteroplasmy < 0.9) in a ± 5 nucleotides window.

This last approach was deemed necessary after noticing that Nanopolish introduced a false methylation call every time a homoplasmic haplogroup-defining variant position fell within ± 5 nucleotides from a CpG. As 11 nucleotides is the kmer size that Nanopolish considers to calculate CpG LLR, we hypothesized that the introduction of a nucleotide variant within ± 5 nucleotides from the CpG altered the Nanopolish methylation determination, leading to an incorrect methylation call. To demonstrate this, we used MToolBox[255] to generate a consensus sequence from the Illumina data, carrying only the major alleles at each position, and we used this new sequence to perform another methylation calling on our ONS samples. As expected, no methylation was identified in the CpGs close to the haplogroup-defining variants this time (**Figure 6.3**).

Differential methylation analysis was performed on cell lines and primary fibroblasts using the R package DSS[258] following the protocol detailed by Gigante and colleagues[191] using the H haplogroup and control fibroblasts as baseline, respectively. Differentially methylated mtDNA positions and regions (defined by overlapping tiles of 50nt) were deemed significant if False Discovery Rate was below 1%.

3.10.7 Dataset simulation and background noise modelling

To elucidate the relationship between the methylation levels and the read depth in ONS data, we generated in silico multiple datasets of simulated sequencing experiments, subsampling the negative control BAM file. We used samtools `-s` (read fraction) `-b BAM > simulated.sam`. We selected 30 different read fractions matching the read depths achieved with both the fragmentation and BamHI-based

sequencing experiments on native DNA. Once the simulated SAM files were generated, we proceeded with the methylation calling using Nanopolish, following the same workflow used for cell lines, primary fibroblasts and tissues. Methylation levels calculated on the simulated data were therefore considered background noise introduced by either the ONS technique or the methylation calling procedure, as previously observed. The following analyses were performed by Dr. Claudia Calabrese. We chose a function describing an exponential decay (1) to model the background noise, given the inverse relationship we observed in simulated data (high methylation levels corresponding to low read depth and vice versa).

$$(1) \quad Y = m * e^{(-t*x)} + b$$

The goodness of fit test showed that the exponential function in (1) well explained the variation of the simulated data (R2 = 0.94), therefore we set out to use the estimated parameters (m, t and b) and the equation in (1) to calculate the background noise present in all downstream ONS sequencing experiments. The background noise model fitting was performed using the `optimize.curve_fit` function of the Scipy Python module. All analyses have been performed in Python 3.0 and code is available at https://github.com/ib361/scripts_paper

3.11 Illumina Miseq library preparation and sequencing

Sequencing of human samples on the Illumina Miseq platform was performed by Dr. Zoe Golder. MiSeq libraries were prepared from genomic DNA by amplification of the mitochondrial DNA in two overlapping fragments, using the primers outlined in (**Appendix 2**). Amplicons were individually purified, quantified, and then were pooled in equal amounts from each sample. Libraries were prepared using NEBNext Ultra library prep reagents (NEB) according to manufacturer's instructions and sequenced using a 2 × 250-cycle MiSeq Reagent kit v3.0 (Illumina, CA).

3.12 Miseq variant calling analysis

Variant calling was performed by Dr. Claudia Calabrese. Fastq files generated with Illumina Miseq were checked for quality using FastQC v0.11.5[248]. Illumina adapters

and read ends showing poor *per*-base quality were trimmed using TrimGalore! v0.4.5[249], setting a minimum per-base QS = 20 and minimum read length after trimming = 35 bp. Mitochondrial variant calling was then performed with the MToolBox pipeline v1.2[255], which mapped reads to the human reference genome (GRCh38) with the two-mapping step protocol integrated in the pipeline, to exclude possible NUMT. Single nucleotide variants with ≥ 5 reads of support (and at least 1 read of support on each strand) and minimum QS per base ≥25 were retained. Haplogroup predictions were performed using both MToolBox and Haplogrep 2 v.2.1.1[257] and based on the human phylogeny annotated in the Phylotree build 17[256]. Haplogrep2 predictions were based on homoplasmic mtSNVs only (with heteroplasmy ≥ 90%).

3.13 Statistical tests

Each data distribution was checked for normality by using the Shapiro-Wilk test. For pairwise comparisons, we chose to use the parametric Student's t-test or Anova one-way test when values were normally distributed. When not stated, distributions were non-normal and a Wilcoxon two-tailed test was used instead. Spearman's rank test has been used to calculate correlation between variables.

Chapter 4. CpG methylation analysis of mtDNA with WGBS

4.1 Introduction

Methylation of the 5th carbon of cytosines in CpG context is a well-described epigenetic modification of the DNA, which has a central role in regulating gene expression both during development[97] and throughout life[259]. Its presence and role in mtDNA has been researched by multiple groups soon after the discovery of mtDNA[170]. While early studies showed that mtDNA possesses very low or absent methylation levels[205–207], the recent discovery of a mitochondrial isoform of the methyltransferase *DNMT1*[210] rekindled the interest in mitochondria epigenetics. As reviewed more extensively in the introduction, a series of recent studies have tried to show that mitochondrial CpG (and sometimes non-CpG) methylation is present at various degrees in a variety of contexts including ageing[227], environmental exposure to pollutants[219], cancer[226,260] and different neurological diseases[231,232]. However, these studies fail to reach a consensus on both whether a specific pattern of methylation is shared in the different context analysed and on which of the three nuclear methyltransferases (*DNMT1*, *3a* or *3b*) is the one responsible to establish mtDNA methylation patterns. In parallel to the various studies demonstrating presence of mtDNA methylation, a few works were published on the opposite view, that mtDNA methylation in fact does not exists, and that the results of published studies could in fact be ascribed as artefacts [169,239,240,261]. Whole genome bisulfite sequencing (WGBS) is the gold standard technique used to study the presence of CpG modifications on the nuclear genome[186,262,263] and is also the one at the heart of most of the studies that show presence of mtDNA methylation. Some of the arguments against the presence of mtDNA methylation revolve around whether the harsh chemical treatment with bisulfite, the basis of WGBS, introduces a bias in DNA sequences that have an uneven distribution of unmethylated cytosines between the strands (such as the mtDNA L-strand)[242]. Work by Olova and colleagues has in fact shown that such sequences are especially sensitive to degradation following bisulfite treatment. This in turn may potentially lead to a bias being generated during methylation calling of the two strands. Such bias has even been recently identified by Dou and colleagues[264] but it has been interpreted by the authors as an intrinsic property of mtDNA.

To help elucidate this ongoing issue, in this chapter we sought independent evidence on whether WGBS has indeed limitations in establishing accurately the level of CpG

methylation on mtDNA or if biases are introduced and on what level. To do so, we applied a commonly used bioinformatic workflow to analyse publicly available WGBS experiments on human cell lines and tissue. Particularly, we focussed on quantifying alignment biases and their possible effects on mtDNA CpG methylation calling. The data has been sourced from the NIH Human Epigenome Roadmap Project[247] repository. The consortium was an international effort aimed at building a database of experiments which could give insights into the epigenetic landscape of the human genome.

The core of our bioinformatic workflow is based on a software called `bismark`[251], used to perform the alignment of WGBS data. During the WGBS library preparation process, unmethylated cytosines (which represent the majority in the human genome) are chemically converted to uracils by the deamination on the C4 carbon. Uracil is then converted to thymine the PCR amplification step of the library preparation. For this reason, WGBS reads present a reduced nucleotide complexity compared to standard NGS experiments (**Figure 3.1**), which precludes the use of standard alignment tools. To solve this problem, `bismark`[251] generates an *in silico*-converted reference genome where all cytosines are virtually converted to thymines in the reference provided. Read alignment is then performed using the converted genome and, once a match is found, the original unconverted reference genome is compared to the WGBS read to check which cytosines were converted (thus marked as unmethylated) and the ones that remained unconverted after the bisulfite treatment (thus considered as methylated).

4.2 Results

4.2.1 WGBS experiments quality control

We downloaded data from 67 human cell lines and tissues from the NIH Human Epigenome Roadmap Project[247] repository. Fifty-five passed quality control (Methods) and were aligned to the human genome build GRCh38 (**Appendix 4**). As reads were aligned to both the nuclear and mitochondrial genome (rCRS[265]), we were able to identify and exclude those with a double nuclear-mitochondrial alignment (i.e.: secondary alignments), which likely represented nuclear-mitochondrial sequences (NuMTs)[5].

From the BAM files we were able to extract the information on mitochondrial reads and their strand alignment. By looking at the distributions of the reads per mtDNA strand, we arbitrarily divided the samples in two groups, depending on the percentage of the reads mapped on the H strand ("Bias" and "Low Bias"). The Bias group included 58.2% (N= 32/55) samples, with a majority of reads mapped to the mitochondrial H-strand (≥ 55% reads; P = ≤ 0.0001, **Figure 1.1 A-B**), and a more pronounced per strand coverage bias (L-strand coverageBG = 6.2%-88.3%; H-strand coverageBG = 83.5%-91.7%, **Figure 1.1 C** left panel). The remaining samples (41.8%, N = 23/55, "Low Bias" group, LBG), showed a milder mapping bias on the H-strand (between 51%-55% reads; P = ≤ 0.0001, **Figure 1.1 A-B**), although present in all the samples analysed, but no coverage bias (**Figure 1.1 C** right panel).

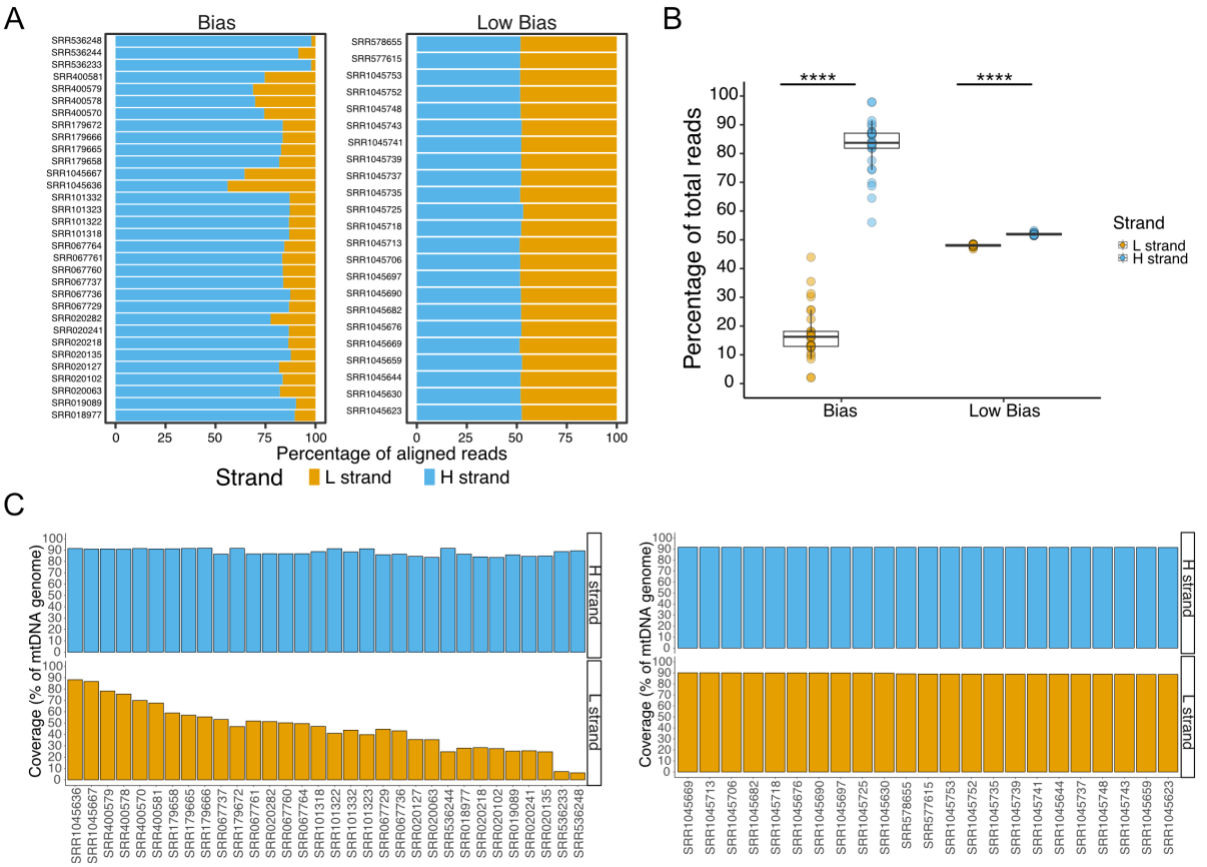


Figure 4.1: Quantification of alignment and coverage bias. a) Percentage of reads aligned to the mtDNA reference per sample, identifying samples with a marked (Bias, N = 32/55) or low (Low Bias, N = 23/55) per-strand-bias. b) Percentage of reads aligned to mtDNA, divided by bias group. Boxplot shows the percentage of reads aligned to the mtDNA reference. The lower and upper hinges correspond to the first and third quartile of the distribution, with median in the centre and whiskers span no further than 1.5*interquartile range. Stars indicate

significance (**: two-sided P ≤ 0.0001, Wilcoxon test). C) Percentage of mtDNA covered by at least 5 reads on the two mtDNA strands (H and L) in (top) Bias and (bottom) Low Bias sample groups.**

This bias is reflected in the differences observed in the average mitochondrial read depth per position calculated in the two groups (i.e.: the average number of reads aligned to the mitochondrial genome per position): 66.32 ± 28.84x in the Bias Group versus 148.77 ± 55.45x in the Low Bias Group (mean ± sd; Mann-Whitney test: P = ≤ 0.0001, **Figure 4.2**).

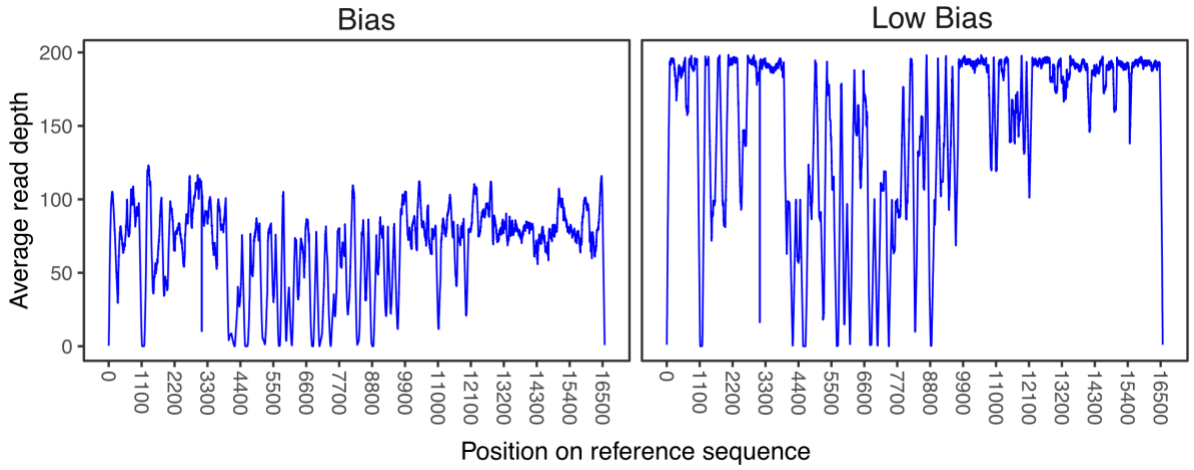


Figure 4.2: WGBS samples read depth distributions. Distribution of the average read depth per mtDNA position in the two per-strand-bias groups.

4.2.2 WGBS experiments methylation analysis results

Finally, methylation analysis in both groups revealed higher methylation levels in the L-strand compared to the H in all the samples analysed, despite this difference being more pronounced in the Biased Group samples compared to the Low Bias Group samples (L-strandBG= 4.97% ± 8.79 vs H-strandBG= 2.01% ± 1.92 (mean methylation ± sd); L-strandLBG= 1.43% ± 0.77 vs H-strandLBG= 1.39% ± 0.7 (mean methylation ± sd); P = ≤ 0.001; **Figure 4.3 A,B**).

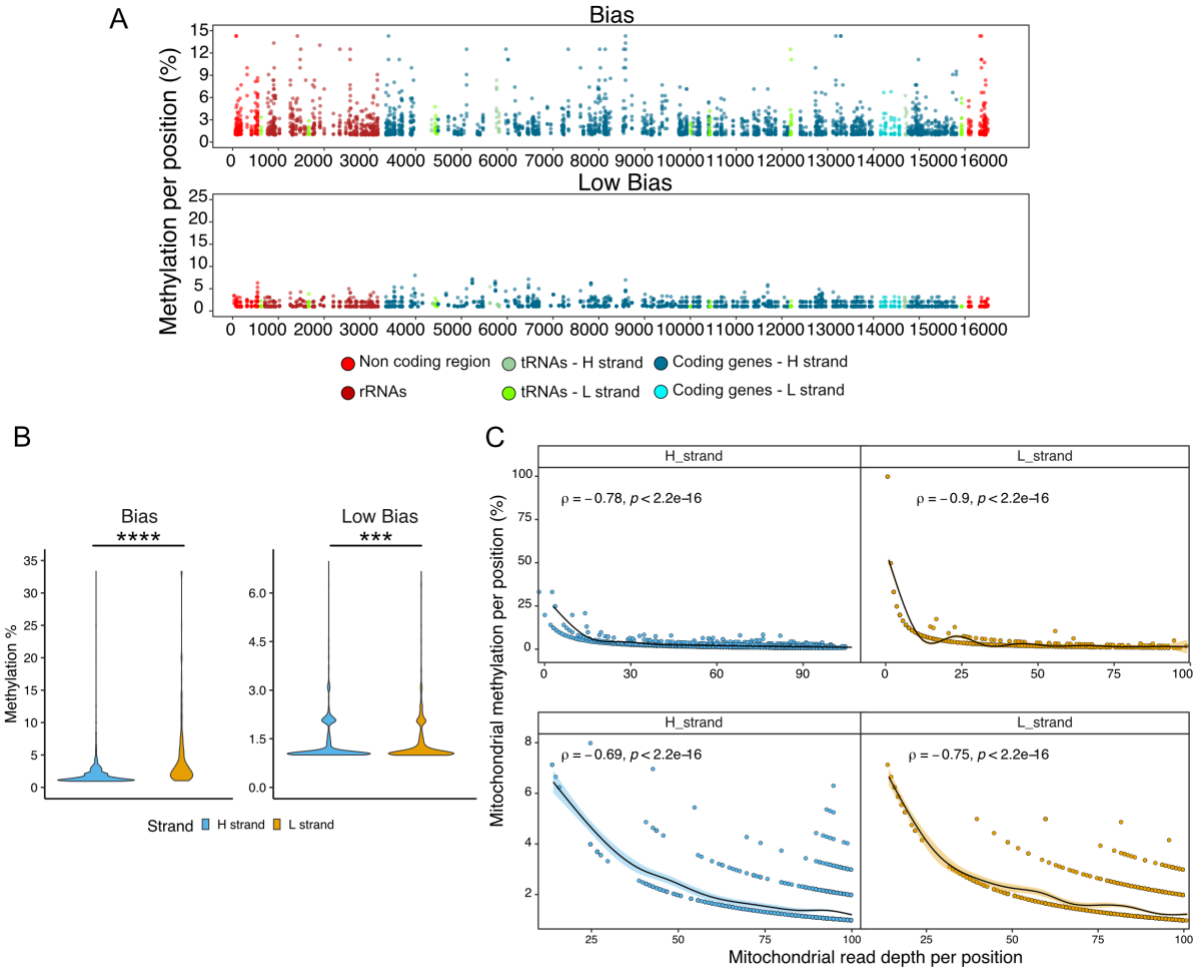


Figure 4.3: Results of the mtDNA methylation analysis on WGBS samples. a) Distribution of the methylation percentage per mtDNA position in CpG context, in (top) Bias (N = 32) and (bottom) Low Bias (N = 23) groups. Each dot represents every CpG in every sample. Methylation values are expressed in % of methylation. b) Quantification of the average CpG methylation per strand (H and L), divided by per-strand-bias group. The lower and upper hinges of the violinplot correspond to the first and third quartile of the distribution, with median in the centre. Stars indicate significance (*: two-sided P ≤ 0.001; two-sided ****: P ≤ 0.0001, Wilcoxon test). c) Correlations between average read depth and average methylation percentage for every cytosine in CpG context, in bias (upper graphs) and low-bias (lower graphs) groups and mtDNA strands (H and L). Spearman's rank test correlation coefficient and two-sided P-values are shown. For all the plots in b,c), Average methylation is intended as the mean methylation value across all the WGBS samples analysed.**

Since the methylation level per CpG is expressed as a ratio of the reads supporting a methylated CpG over the total of the reads covering that CpG position, we reasoned that the bias we observed per strand could be explained by strand-specific fluctuations of the read depth. Indeed, we found a significant inverse correlation between

methylation levels detected and the read depth (Spearman's rank test P < 2.2e-16; average rho coefficient = -0.78, **Figure 1.3 C**), leading to the appearance of higher methylation levels where read depth is low. This holds true also for the Low Bias Group samples (which possess a milder alignment bias), where local fluctuations in the read depth alter CpG methylation levels (**Figure 1.4 A,B**).

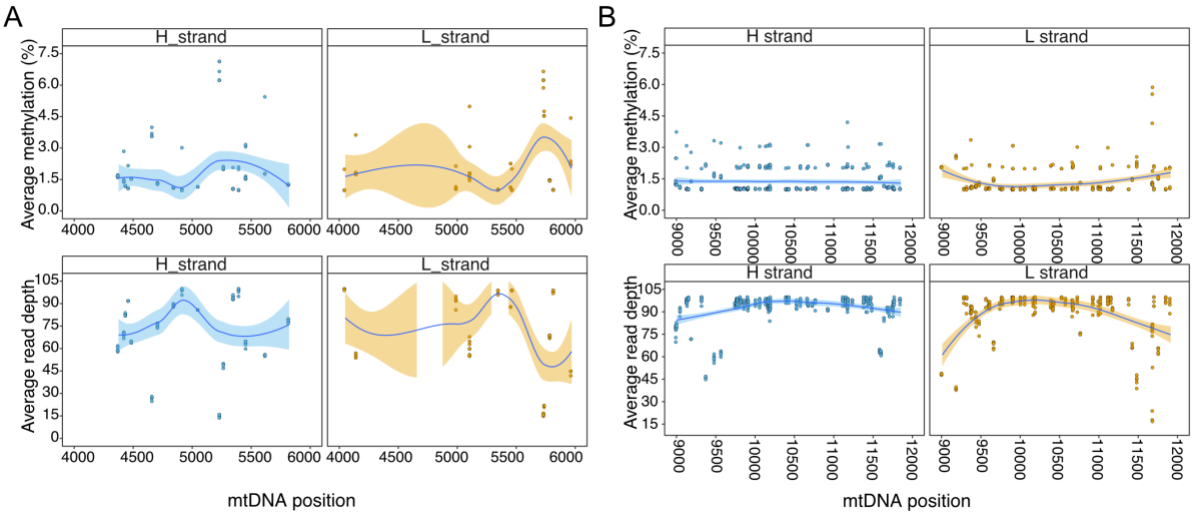


Figure 4.4: methylation patterns in WGBS Low Bias group samples. a,b) CpG methylation and read depth profiles of a 2kb (a) and 3kb (b) mtDNA genome region, per each position in the Low Bias sample group, divided by mtDNA strand (H and L). Each dot represents all the CpGs in the specific area in all the Low Bias samples. Methylation values are expressed in % of methylation. Blue lines indicate the mean over all the data points (calculated using the "loess" geom_smooth R function) and shaded surrounding regions represent 95% confidence interval.

4.3 Conclusions and Discussion

The results of our analysis on mtDNA alignment bias in human WGBS experiments are in line with what was described in the study by Olova and colleagues[242]. In their in-depth analysis of the available WGBS library preparation methods Olova and colleagues show that the principal source of biases in WGBS is the bisulfite-mediated degradation of DNA, followed by a PCR step amplifying any initial biases. Moreover, they show that the bias introduced by bisulfite is non-random, targeting specifically unmethylated C-rich regions of the DNA, such as repeated chromatin regions (i.e.: satellite DNA) or mtDNA. Therefore, we believe that this is a plausible explanation of the bias we also observed in our analysis: the bisulfite treatment degraded

preferentially the C-rich mtDNA L-strand, which in turn could not be fully sequenced (**Figure 4.1 C**).

The strict relationship between methylation and read depth supports the view of an unmethylated mtDNA: if mtDNA was indeed methylated, we would have expected that the amount of unconverted (i.e. methylated) reads would increase in line with the read depth, plateauing around the expected methylation level. The strict inverse relationship observed between these two measurements suggests instead that with the increase in read depth we draw closer to the true mtDNA (low/absent) methylation value, as more information is gathered to describe the methylation status of the individual cytosines. This is further supported by the established notion that circular mtDNA is more resistant to bisulfite conversion[240].

Higher mitochondrial read depths are unlikely to solve the problem either. First, by looking at the average read depths profile in both BG and LBG (**figure 4.2**), it is possible to observe that the read depths achieved are already far higher than what is usually considered good in WGBS (10x for nDNA). Secondly, as shown by the distribution profiles of the read depth in the LBG group (**figure 4.4**), having a uniform distribution of the read depth across the molecule is more important than achieving high read depths (which would anyway be easier to achieve on mtDNA because of the multiple copies present in each cell). However, as shown in the average read depths distribution profile (**figure 4.2**), not only this is not the case, but it seems that the read depth is probably dependent on the mtDNA sequence itself, as some sections are always lowly represented. This would be interesting to investigate in depth, as it would be in line with what was published by Olova and colleagues[242] regarding the preferential degradation of cytosine-rich sequences by the bisulfite treatment.

We could not find a clear explanation to the stark differences observed between the Biased and Low Bias groups in their respective alignment biases. In their study, Olova and colleagues[242] state that the choice of the right kit for WGBS library preparation and the polymerase to use for the subsequent PCR amplification that precedes sequencing are important factors to take into account as a strategy to avoid the introduction of sequencing biases. Even more important is the timing of the bisulfite treatment (before or after sequencing adapter ligation), in order to avoid excessive DNA degradation. A manual search of the library preparation method of all the samples analysed in this study revealed no differences in the kit used for preparing

the sequencing library between the two sample groups, nor in the polymerase used, which were therefore factors excluded as being the cause of the observed bias.

The presence of a methylation bias is something that has been previously described in the literature. At least 19 works published before the one by Olova and colleagues[242] have identified differences in the methylation patterns between the mitochondrial strands[185,186,227,229–232,234–236,266,213–215,218,220,222–224]. Having reviewed those studies we believe that it is possible that those results could indeed be ascribed to the presence of an underlying alignment bias. However, even after the publication of the guidelines for limiting WGBS-derived alignment bias[242], multiple groups have continued to investigate mtDNA methylation using bisulfite-based technologies, without accounting for biases presence[219,226,267–271]. One study in particular stands out: the work by Dou and colleagues[264] identifies a strikingly similar pattern to the one we identified in the Roadmap samples. They analyse both publicly available WGBS experiments and samples sequenced by the group, both in humans and other species. However, while the authors correctly report the presence of an alignment bias in all the samples they analyse, they claim that the differences observed in the methylation levels between the two mtDNA strands are a biologically relevant phenomenon, probably dependent on *DNMT3A* regulation.

Studies of this kind show that the question regarding the presence and the role of mtDNA methylation is far from being fully addressed. The common issue undermining most of the results published in this field is that the principal technology used to detect mtDNA methylation is vulnerable to the introduction of intrinsic biases that lead to contradicting interpretations of the results. This, combined with the added complexity of studying mtDNA (i.e.: presence of NuMTs; multiple mtDNA copies, etc.) has prompted us to look at alternative methodologies to WGBS to analyse mtDNA methylation.

Chapter 5. Experimental setup for CpG methylation detection on mtDNA using Oxford Nanopore Sequencing

5.1 Introduction

The work detailed in **chapter 4** discussed the intrinsic weakness common to all bisulfite-based technologies when it comes to draw conclusions on mtDNA methylation. In light of these considerations, we had to exclude the most used technologies to analyse DNA methylation at single-base resolution level (i.e.: WGBS, RRBS, etc)[180,263,272], as well as the rest of the non-bisulfite based methods (such as MeDIP sequencing, mass spectrometry, etc)[273,274], as none of these possess neither the sensitivity, nor they provide the same amount of information that can be obtained by single-base detection methods.

Long-reads-based sequencing technologies could instead be a valid alternative for mtDNA methylation detection. Currently, such technologies are represented by PacBio SMRT sequencing and by Oxford Nanopore Sequencing (ONS)[275], reviewed in the introduction. For practical and budget reasons (the ONS MinION sequencing platform is a portable device and the starting kit cost is 1k £), we chose to explore only the potential of ONS to measure methylation in mtDNA. Briefly, in ONS DNA is unwound through a protein pore embedded in a synthetic membrane, across which an electric flow of ions is maintained[276]. Fluctuations in the electric flow caused by the passing of the DNA strand through the pore are registered and interpreted (either in real-time or later) by a neural network-based algorithm which re-constructs the original DNA sequence bases identities (a process called "basecalling")[196]. Contrary to Illumina-based sequencing technologies, ONS can sequence native DNA (i.e.: nor treated nor PCR amplified). Therefore, it is possible to identify the signal coming from modified bases such as 5mC (and others), thanks to recent advances in the basecalling technologies[189]. At the time of writing, there are multiple specialised software which can identify such modifications[189], including the principal software used for basecalling, `guppy` [195]. However, all the methylation calculation in this study have been performed using the software Nanopolish[198], as at the time of the analysis that was the most advanced software available.

Briefly, Nanopolish utilises a Hidden Markov Model (HMM) to distinguish between raw electric signal originating from a modified or unmodified cytosine (in the case of 5mC

modification in CpG context). This model is trained to recognise these differences using positive (i.e.: methylated) and negative (i.e.: unmethylated) DNA control sequences[198]. The resulting output of Nanopolish methylation calling is a Log Likelihood Ratio (LLR) that describes the probability of a cytosine in CpG context to be methylated. The direct identification of DNA modifications without the need of PCR amplification is the main advantage of ONS compared to WGBS, potentially overcoming the intrinsic biases created by the bisulfite treatment and subsequent PCR amplifications. In the first part of this chapter we assessed the sensitivity and accuracy of Nanopolish mtDNA methylation calling, using ad-hoc positive and negative controls. To do so we apply a custom bioinformatic workflow combining common tools used in ONS quality control and analysis (**Figure 5.2**).

One of the characteristics of ONS is that virtually there is no limit to the length of the reads generated by the sequencing[276]. Studies have in fact shown that the principal factor limiting the length of an ONS-sequenced DNA read is the physical fragmentation occurring during the library preparation. Ultra-long reads have been used to generate *de novo* plant and human genome assembly that span regions which were usually hard to map using only short reads[275,277].

Therefore, in the second part of this chapter we devised a modification of the standard ONS library preparation (based on random fragmentation) which could allow sequencing of the full-length mtDNA molecule. The advantages of this alternative method would include: 1) identifying methylation not only at the single-base but also at single-molecule level; 2) combining the identification of nucleotide variations with epigenetic modifications using a single technology; 3) the possibility of phasing distant nucleotide variants (and possibly methylation) on the same molecule. Moreover, to save time and resources we devised our method to avoid the need to isolate mitochondria from the biological samples for subsequent purification of the mtDNA. Two mtDNA enrichment methods from gDNA were first assessed, then the most promising one was evaluated against the standard ONS library preparation method, using DNA from control cell lines.

5.2 Results

5.2.1 Negative and positive controls generation

To assess the accuracy of the Nanopolish methylation calling we generated positive and negative controls, replicating the protocol described by Simpson and colleagues[198]. Similarly, we used PCR amplicons either untreated (i.e.: negative control) or treated *in vitro* with a recombinant methyltransferase (i.e.: positive control). However, instead of short PCR sequences we decided to use long-range PCR (LR-PCR) amplicons, covering almost the entirety of the mitochondrial sequence. As the LR-PCR protocol requires a very long amplification cycle (>5 hours), the risk of NuMTs amplification was high. To assess this possibility, we used DNA from a Rho 0 cell line (deprived of mtDNA)[278] in parallel to control DNA used as template for LR-PCR. Agarose gel results revealed that indeed some aspecific bands lower than the expected ~16 kbp band appeared after amplification (**Figure 5.1**). However, as the ~16 kbp band did not appear in the Rho 0 DNA sample, and since the aspecific bands were lower in intensity compared to the ~16 kbp one (denoting lower concentration), we decided to proceed with the methyltransferase reaction and sequencing without further purification steps. A strict filtering on the sequencing reads was applied bioinformatically at a later stage, to avoid the alignment of sequenced reads lower than 14 kbp (likely aspecific).

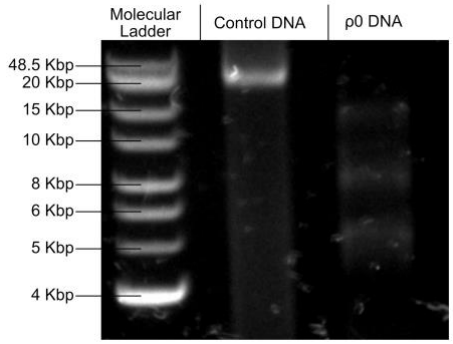


Figure 5.1: aspecific amplification test. Agarose gel results show presence of aspecific low molecular weight bands in the LR-PCR performed on DNA from Rho 0 cells. The apparent higher molecular weight of the mitochondrial amplicon is probably due to incomplete resolution during the gel run.

To generate positive controls, we incubated the LR-PCR products with the recombinant bacterial *M.SssI* methyltransferase. We tested whether the methylation rate is influenced by 3 parameters: I) reaction time II) enzyme concentration and III)

DNA concentration. As we required at least 1 µg of DNA for the ONS library preparation, we decided to use that amount of DNA in all reactions and to vary only reaction times and enzyme units amounts, to test which of these two variables had the biggest effect on the methylation levels. Therefore, while keeping DNA concentration (1 µg), temperature (37°C), reaction buffer and methyl donor s-adenosine-methionine (SAM) concentration (160 µM) steady, we set up a series of reactions in parallel. As readout, we used a restriction enzyme reaction by the enzyme BstUI, which cuts only unmethylated CpGs. Generation of fragments from the BstUI reactions were assessed on the TapeStation instrument, and results are illustrated in **Figure 5.2**. The reactions conditions which resulted in the highest protection from BstUI reaction were 4 hours of incubation with 50 U of M.SssI.

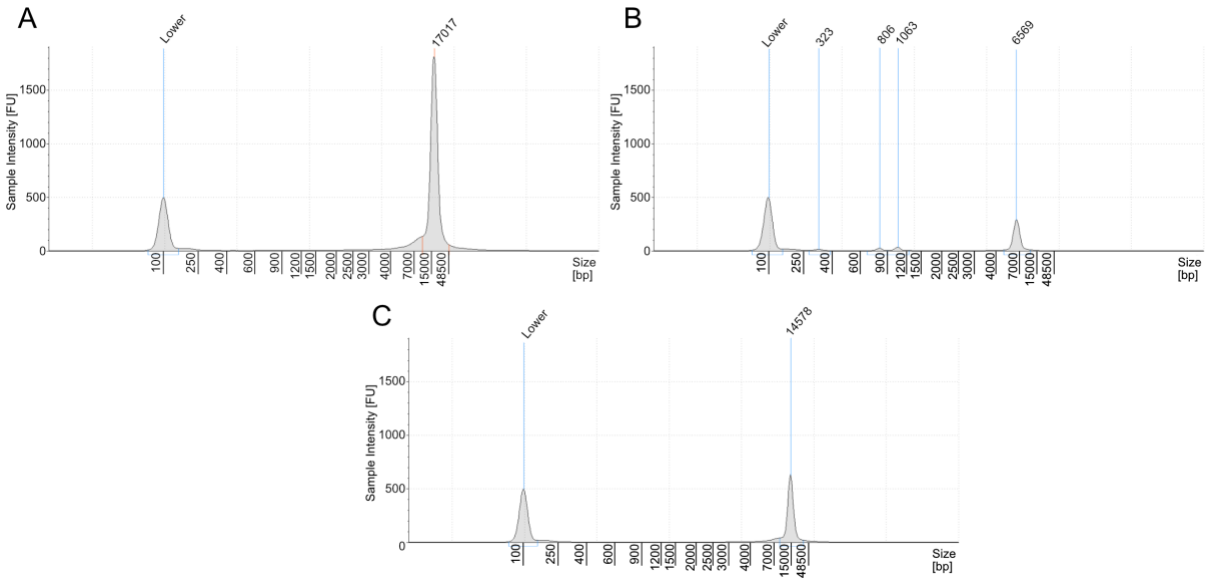


Figure 5.2: Positive control generation assessment. **Fragment peak analysis results using Genomic ScreenTape. a) Fragment peak profile of untreated mitochondrial LR-PCR amplicon. b) Fragment peak profile of mitochondrial LR-PCR amplicon treated for 1 hour with the restriction enzyme BstUI. c) Fragment peak profile of mitochondrial LR-PCR amplicon treated first with the recombinant methyltransferase M.SssI for 4 hours (at the reaction conditions stated in the methods section), then treated for 1 hour with the restriction enzyme BstUI. The different peak sizes shown in a) and c) are possibly due to the low resolution of the tape at high molecular weights.**

5.2.2 Bioinformatic workflow

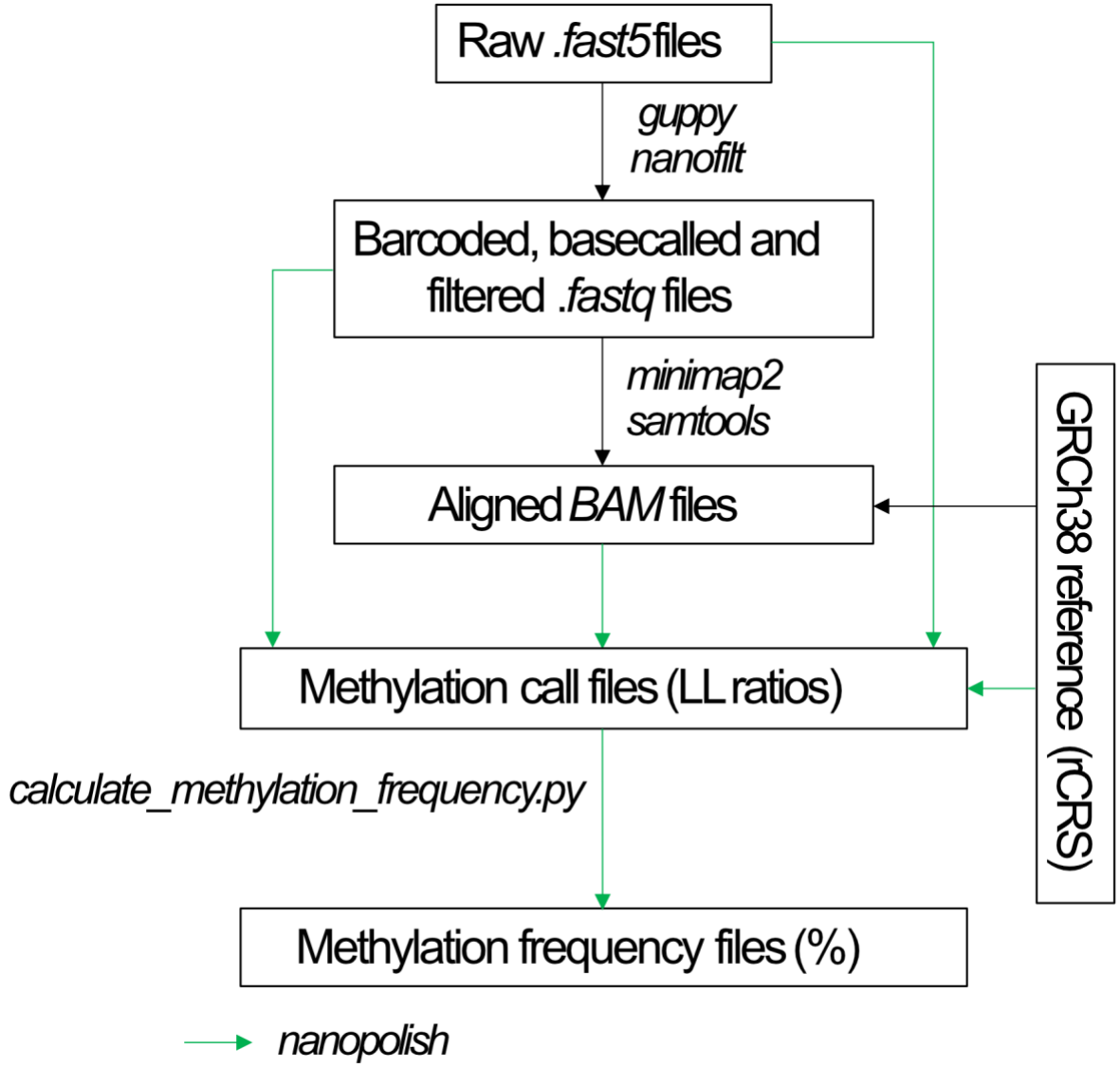


Figure 5.3: Bioinformatic pipeline overview.

The workflow used to analyse negative and positive controls data was overall identical to the one used for the rest of the samples that were analysed in this study and is outlined in **Figure 5.3** (any specific ad-hoc modifications introduced later are discussed in the following chapters). Briefly, raw fast5 files were basecalled and separated by barcodes using `guppy`. Specifically for LR-PCR controls, sequenced reads shorter than 14 kbp were filtered out before alignment to the hg38 human genome assembly. BAM files, basecalled fastq files and raw fast5 files were then used by Nanopolish[198] to generate methylation LLR for all the cytosines in CpG context found in the sequenced negative and positive controls (**Figure 5.4**).

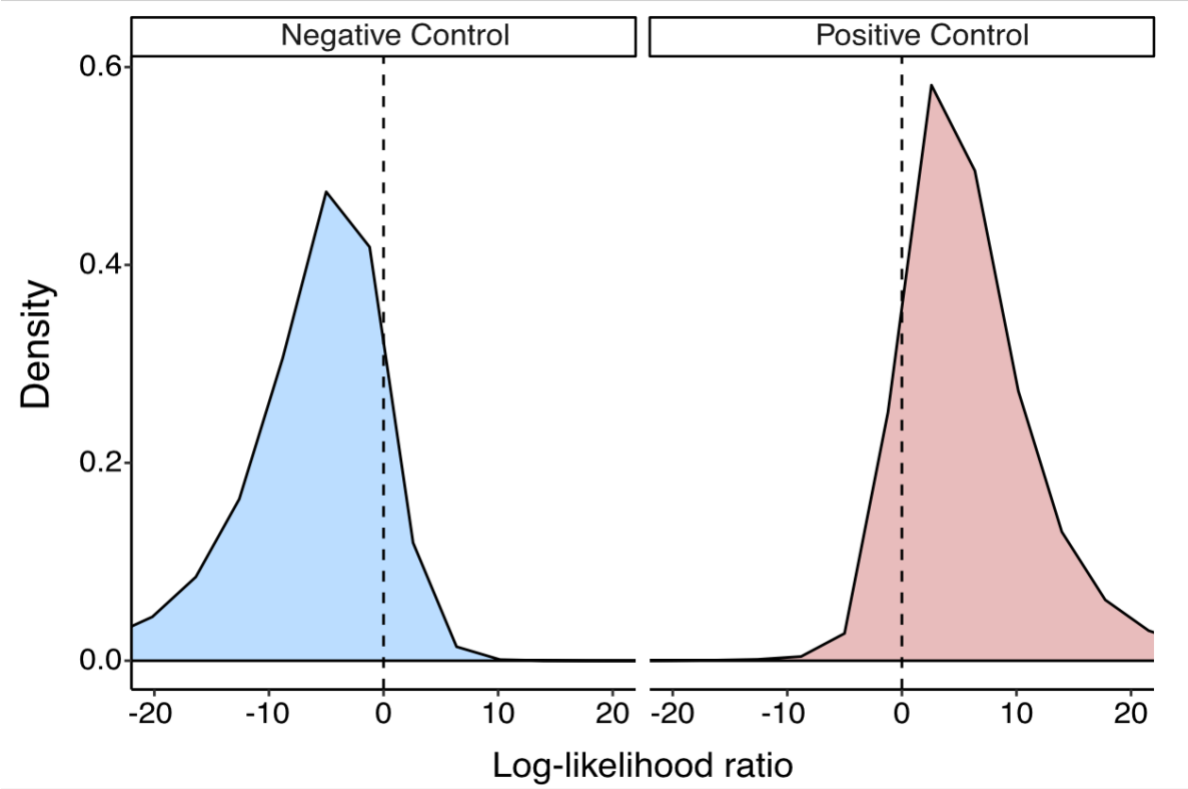


Figure 5.4: Distributions of LLR from negative and positive controls. Log-likelihood ratio values of methylation calculated by Nanopolish, using the positive and negative controls. The log-likelihood ranges between -20 and 20 (used to build the Receiver operating characteristic (ROC) curve) are shown

5.2.3 Accuracy assessment of Nanopolish methylation calling

To determine the methylation status of the individual CpGs Nanopolish applies a default threshold of 2.5 on the calculated LLRs, above which a site is considered methylated. It is worth noting that this threshold has been determined by the analysis of nuclear DNA data only. Therefore, to choose the most accurate methylation calling threshold for mtDNA, we used LLRs derived from our known controls to determine true and false positives ratios at varying threshold values (from -20 to +20, with 0.25 increments), following a procedure published previously by the Nanopolish developers[198] (Methods). We then calculated a receiving operating characteristic (ROC) and methylation calling accuracy (intended as proportion of true calls; **Figure 5.5 A,B**). The ability to distinguish between mtDNA unmethylated and methylated sites was measured by the area under the ROC curve (AUC), which was equal to 0.97

(**Figure 5.5 A**). Also, at the default Nanopolish threshold of ≥2.5 the accuracy was 97.7% (**Figure 5.5 B**).

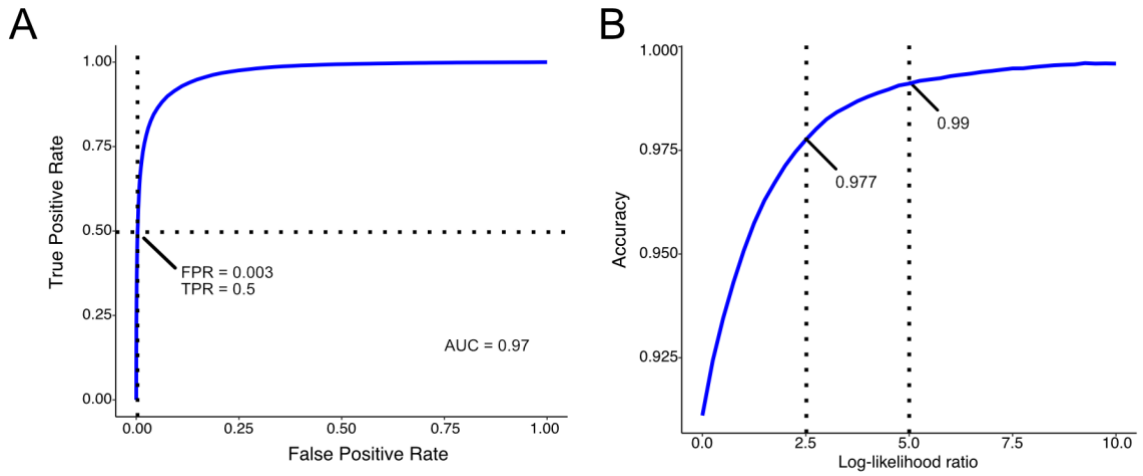


Figure 5.5: Assessment of Nanopolish methylation calling accuracy. a) ROC curve calculated by changing the methylation call log-likelihood ratio threshold from a value of -20 to 20, with a step of 0.25. The dash lines are drawn at FPR (False Positive Rate) and TPR (True Positive Rate) values obtained by setting the ratio equal to 5. AUC = area under the curve. b) Methylation call accuracy calculated at increasing values of log-likelihood ratio (ranging between 0 and 10). The dash lines indicate the accuracy achieved at the ratio equal to 2.5 (accuracy = 0.977) and 5 (accuracy = 0.99).

To increase the sensitivity in detecting mtDNA methylation, we decided to improve the methylation calling accuracy further. To do that, we chose a more stringent calling threshold of LLR ≥ 5. This increased the accuracy to 99% (**Figure 5.5 B**), dropping the false positive rate (FPR) from a 0.016 FPR (at LLR ≥ 2.5) to 0.0032 (at LLR ≥5). Additionally, by looking at the methylation profiles of the negative control, we identified 13 residues with a methylation percentage consistently higher than 2 standard deviations from the average negative control methylation value (**Appendix 3**, **Figure 5.6**). We checked the methylation level of these 13 positions in all of the samples sequenced in this study, and we found that they were consistently methylated at around the same level (data not shown). For this reason, these residues could represent false positives (possibly due to their sequence context, although this was not explored in this study) and were therefore excluded from all of our analyses.

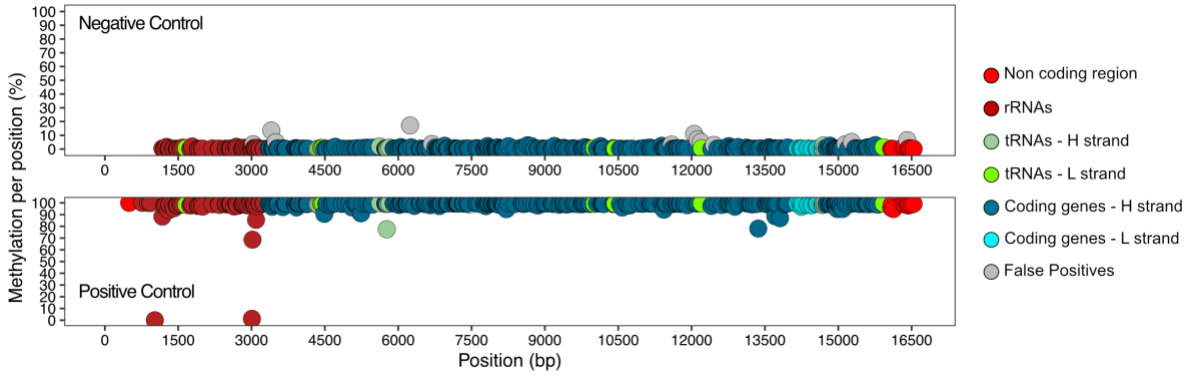


Figure 5.6: Results of the Nanopolish methylation calling on NC and PC. Distribution of the methylation percentage per mtDNA position in CpG context, in (top) negative and (bottom) positive controls. Methylation values are expressed in % of methylation. Grey values represents the 13 positions identified as likely false positives.

5.2.4 Improvement on ONS library preparation: advancement over the standard protocol

Many of the published studies sequenced mtDNA isolated from pure mitochondria preparations, which usually involves a time-consuming step requiring large amount of biological material[279]. If on one side this approach has the potential advantage of reducing/avoiding NuMTs contamination before the sequencing step, on the other side the purity of the mitochondria preparations may vary greatly between the available methods[279]. This approach is even more problematic for results obtained with mass spectrometry: not only nuclear DNA could end up mixing with mtDNA in the mitochondrial preparation, but bacterial DNA may act as an important confounder too, as it is methylated and indistinguishable from mtDNA[280]. Therefore, we set out to develop a protocol that avoids the mitochondria isolation step and still enrich for mtDNA sequences starting from gDNA.

To achieve this, we tested two different approaches. The first is a method published by Jayaprakash and colleagues[245], based on the reaction of the enzyme Exonuclease V with gDNA. This enzyme degrades linear sequences (nuclear DNA) while keeping circular ones (mtDNA) intact. Linearisation of mtDNA for sequencing with ONS is achieved by BamHI digestion. The other method is based directly on the restriction reaction of the enzyme BamHI, which cuts nuclear DNA multiple times (at the multiple BamHI restriction sites present on nDNA), while cutting mtDNA once. Subsequent

selection of the longer restriction fragments ultimately allows the enrichment of full-length linearised mtDNA molecules.

5.2.4.1 Exonuclease V approach

For the first approach (based on Exonuclease V) we digested control DNA and Rho 0 DNA used as positive and negative controls, respectively. After digestion, DNA was purified using AMPure beads at 0.5x ratio. Analysis of purified DNA by agarose gel electrophoresis revealed the presence in the control lanes of a low molecular weight band, maintained after purification. This band could correspond to supercoiled mtDNA molecules migrating faster into the agarose gel, as it was reported for plasmid vector purification[281] (**Figure 5.7**).

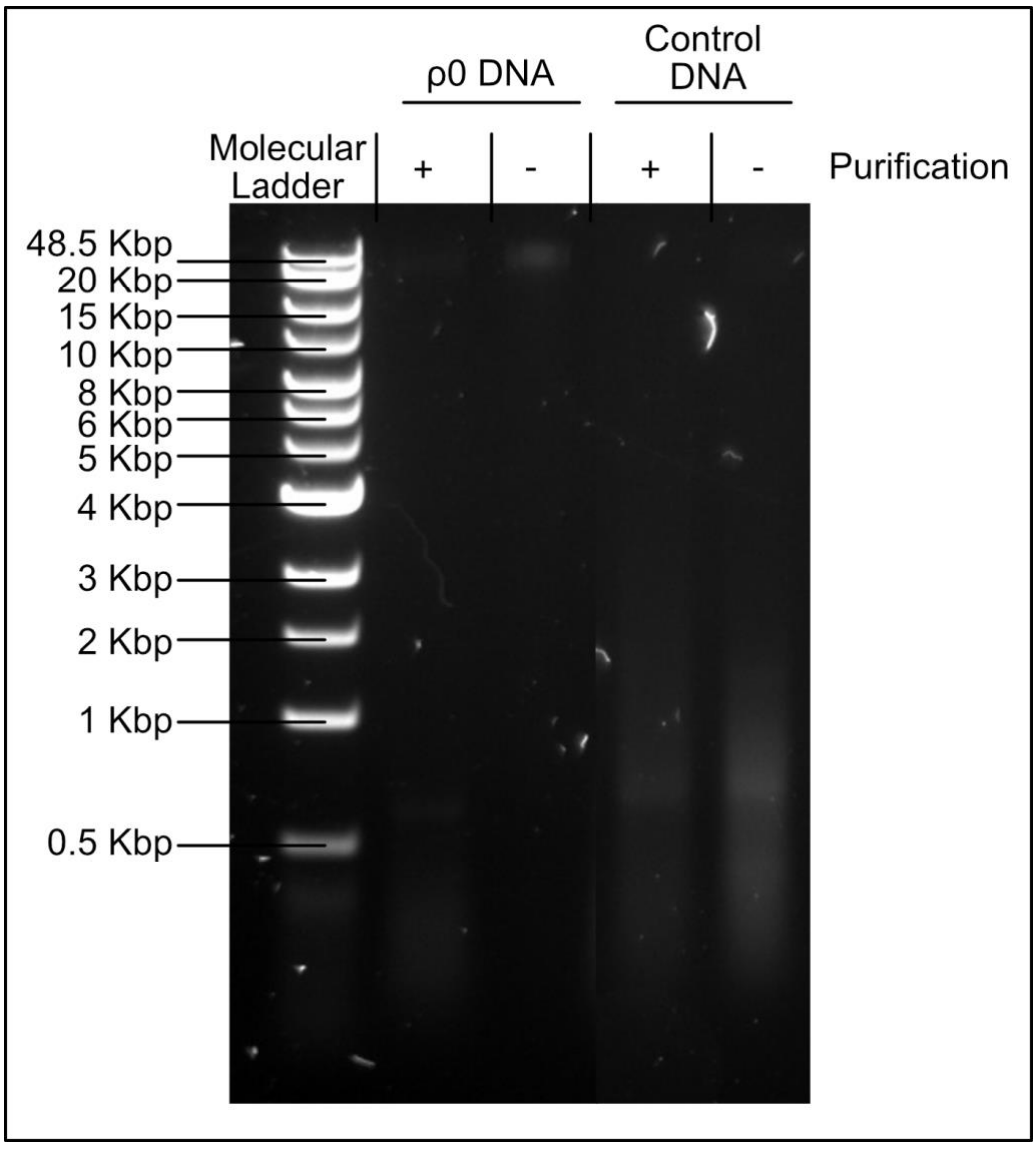


Figure 5.7: Exonuclease V protocol results. **Analysis of DNA digested with Exonuclease V of Rho 0 and control cell lines before or after AMPure beads purification (0.5x).**

To determine if full-length mtDNA was maintained after Exonuclease V digestion, we first performed a LR-PCR reaction on purified samples. We analysed amplification results by agarose gel electrophoresis, using as positive and negative controls LR-PCR performed on pure mtDNA isolated from mitochondria of Rho 0 or control cell lines. As expected, we did not observe any amplification in the digested Rho 0 sample. On the other hand, we observed one full length amplicons of the expected size (16.5 kbp) from digested Control sample (**Figure 5.8**).

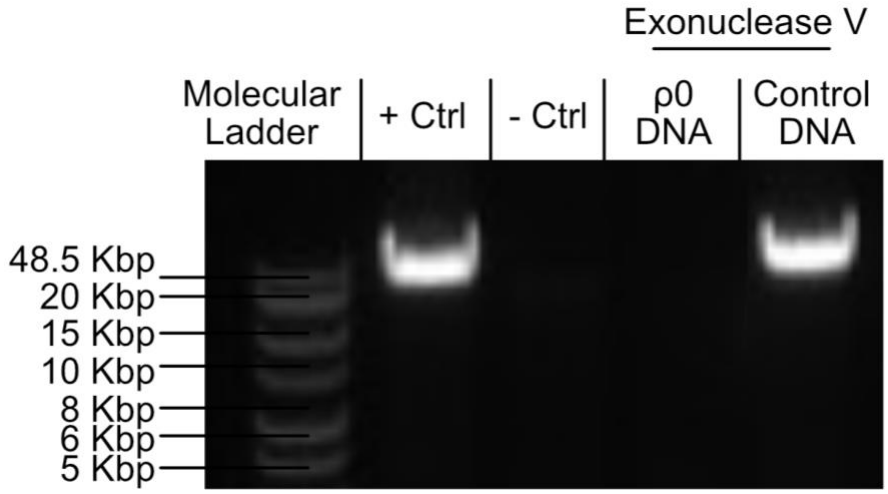
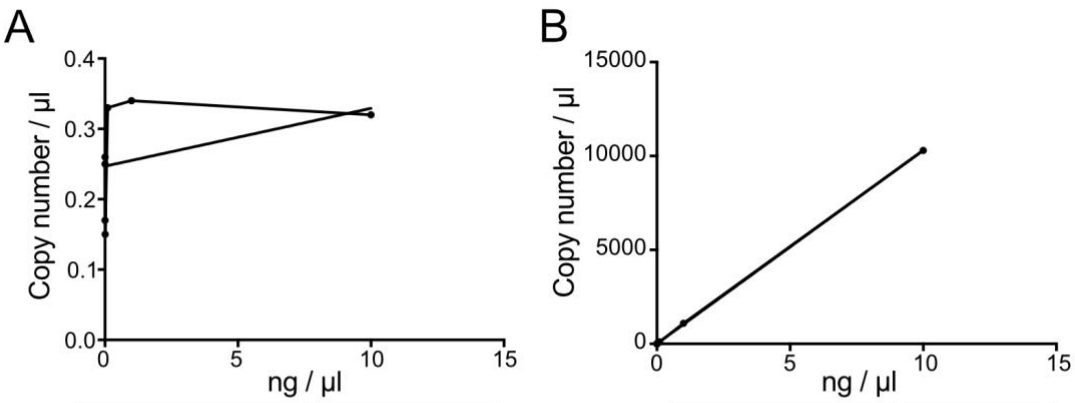


Figure 5.8: LR-PCR on Exonuclease V-digested samples. **Analysis of the presence of intact circular mtDNA after enzymatic digestion and AMPure beads purification. "+ Ctrl" = LR-PCR on untreated control DNA, "- Ctrl" = LR-PCR on untreated Rho0 cells DNA.**

Next, we performed ddPCR using a nuclear (*RNASE P*) and a mitochondrial probe (*MT-ND1*). As expected, we saw no amplification of either probe in the Rho 0 samples digested with ExoV and amplification of the mitochondrial probe only in the control samples (**Figure 5.9**).



Goodness of Fit	
R square	0.1603
Sy.x	0.07689
Is slope significantly non-zero?	
F	0.9545
DFn, DFd	1, 5
P value	0.3734
Deviation from zero?	Not Significant
Equation	Y = 0.008227*X + 0.2469

Goodness of Fit	
R square	1
Sy.x	26.25
Is slope significantly non-zero?	
F	128230
DFn, DFd	1, 5
P value	<0.0001
Deviation from zero?	Significant
Equation	Y = 1030*X + 10.31

Figure 5.9: ddPCR results of Exonuclease V-digested samples. **a) copy number/µl ratio of the nuclear RNASE P gene in samples treated with Exonuclease V b) copy number/µl ratio of the mitochondrial MT-ND1 gene in samples treated with Exonuclease V. Correlation statistics are shown in the boxes below. Pearson correlation coefficient is shown. N =1**

5.2.4.2 BamHI approach

For the second approach, we used the restriction enzyme BamHI to digest gDNA. We chose BamHI as it had been used previously in the literature[169,240]. Again, we digested control DNA (used as positive control) and Rho 0 DNA (used as a negative control). We purified fragmented products with AMPure beads at 3 different DNA/beads ratios (0.5, 0.3 and 0.1), to determine in which condition we could obtain less low molecular weight products. On agarose gel electrophoresis, we observed DNA migrating at the same molecular weight as the positive control on the 0.5x AMPure-purified control and Rho 0 DNA lanes (**Figure 5.10**).

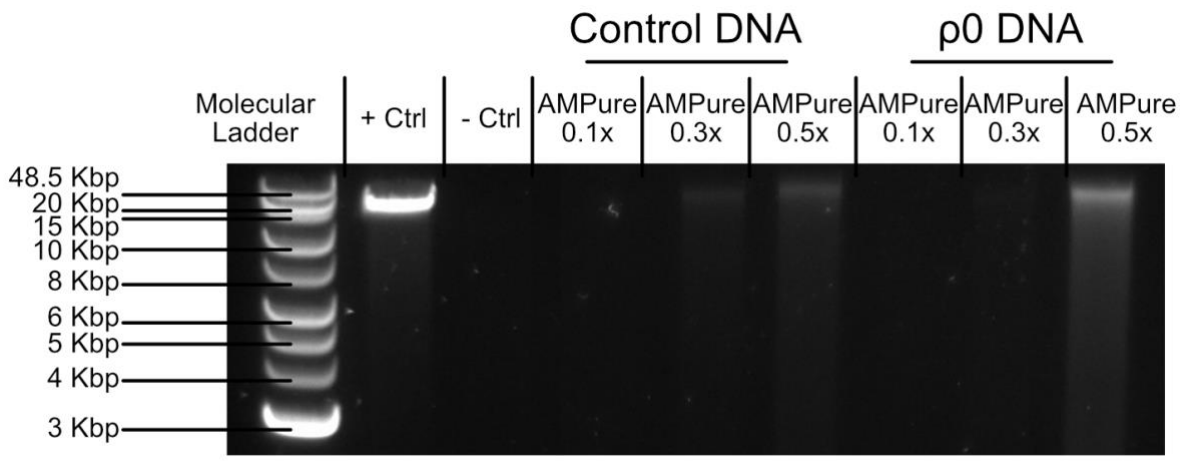


Figure 5.10: Results of BamHI-digested samples. **Results of electrophoresis on 0.7% agarose gel shows the presence of high molecular weight bands after enzymatic digestion and different AMPure beads purification. "+ Ctrl" = LR-PCR on untreated control DNA, "- Ctrl" = LR-PCR on untreated Rho0 cells DNA. The presence of intense high molecular weights bands in the BamHI-treated samples could also be attributed to incomplete enzymatic digestion, although this was not tested further in this study.**

We also noticed a faint band of the same molecular weight as the positive control in the 0.3x AMPure-purified Control DNA lane, while no products were visible in the rest of the conditions. We used the same positive and negative controls as the previous experiment.

The presence of a band in the digested DNA from Rho 0 cells prompted us to investigate *in silico* the possibility of the formation of digestion products ≥ 16.5 Kbp after BamHI digestion. Indeed, longer fragments can be formed from the digestion of nuclear DNA with BamHI, possibly explaining the formation of the band we observed in the Rho 0 sample (**Figure 5.11**).

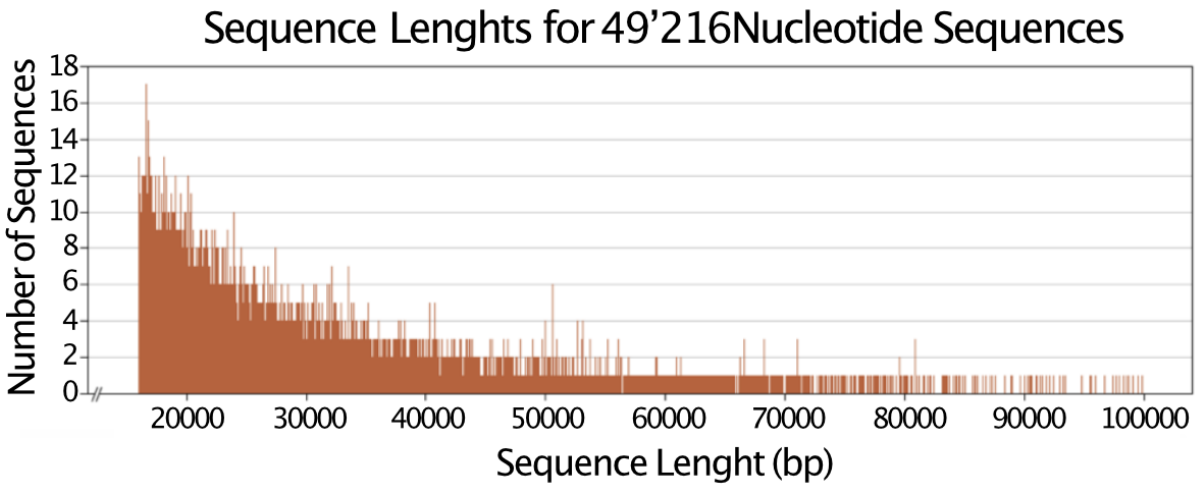
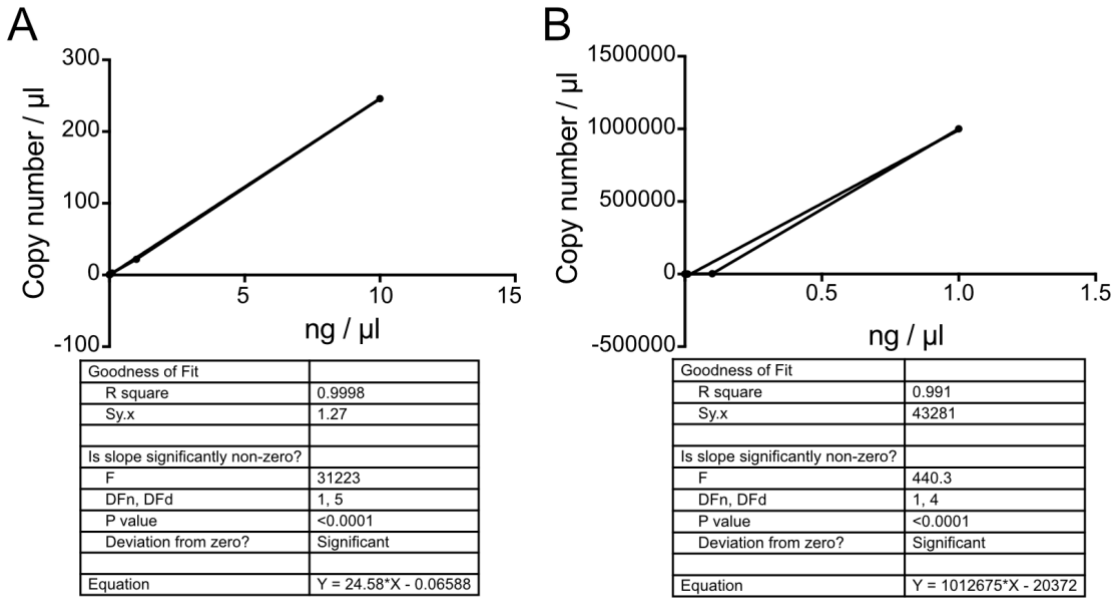


Figure 5.11: *In silico* digestion of the human genome reference hg38 with the enzyme BamHI. Graph shows the total number of the possible fragments ≥ 16.5 Kbp that can possibly be formed by digesting *in silico* a single hg38 human reference genome with the restriction enzyme BamHI.

This was further confirmed by the ddPCR results which showed *RNASE P* amplification and no signal of the mitochondrial probe in BamHI-digested Rho0 samples and amplification. In BamHI-digested control samples, both probes were amplified (**Figure 5.12**).



Goodness of Fit	
R square	0.9998
Sy.x	1.27
Is slope significantly non-zero?	
F	31223
DFn, DFd	1, 5
P value	<0.0001
Deviation from zero?	Significant
Equation	Y = 24.58*X - 0.06588

Goodness of Fit	
R square	0.991
Sy.x	43281
Is slope significantly non-zero?	
F	440.3
DFn, DFd	1, 4
P value	<0.0001
Deviation from zero?	Significant
Equation	Y = 1012675*X - 20372

Figure 5.12: ddPCR results of BamHI-digested samples. a) copy number/µl ratio of the nuclear RNASE P gene in samples treated with BamHI b) copy number/µl ratio of the

mitochondrial MT-ND1 gene in samples treated with BamHI. Correlation statistics are shown in the boxes below. Pearson correlation coefficient is shown. N = 1

The 0.5x DNA/beads ratio resulted in the highest recovery of DNA after digestion. To further improve on this, we tested the 0.5x DNA/beads ratio purification against an alternative method based on the use of NEB PCR purification columns. We chose to test this latter method because it only required a small modification of the standard purification protocol, and because we reasoned that we could obtain more consistent results with a column-based purification approach compared to one that requires handling of AMPure beads. Results showed that the purification method based on NEB columns purification was the best in both DNA recovery and high molecular weight fragment enrichment (**Figure 5.13**).

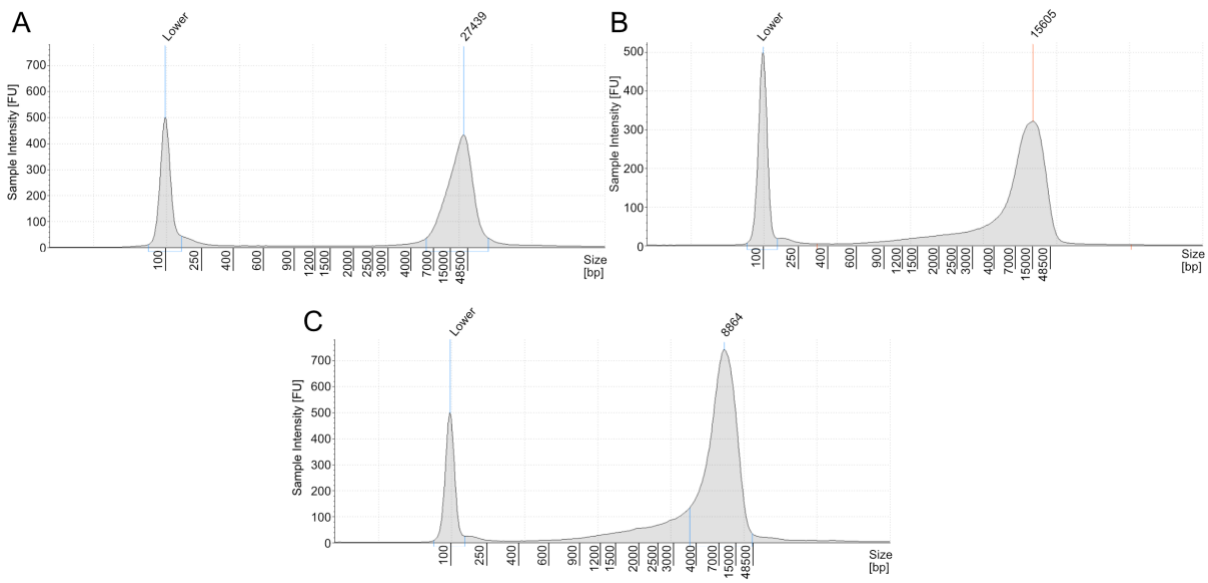


Figure 5.13: Comparison of BamHI purification methods. Fragment peak analysis results using Genomic ScreenTape. a) Fragment peak profile of untreated Control DNA. b) Fragment peak profile of Control DNA digested with BamHI and purified using 0.5x AMPure beads. c) Fragment peak profile of Control DNA digested with BamHI and purified using NEB PCR purification columns.

Comparing the two mtDNA enrichment approaches (Exonuclease V and BamHI-based), it was clear that although we could achieve a higher degree of mtDNA purification with the Exonuclease V method, we would require an amount of starting gDNA too high to achieve the 1 µg needed for ONS library preparation after digestion (**Table 2**). This, combined with the 48 hours of Exonuclease V incubation time,

prompted us to choose the BamHI approach (+ NEB column purification) as the method of choice for mtDNA enrichment from gDNA.

Protocol	Input	Output	Loss
Exonuclease V	1 µg	~2.5 ng	99.75%
BamHI	5 µg	4.72 µg	5.6%

Table 2: Comparison of the outputs of the Exonuclease V and BamHI digestions of Control DNA

5.2.5 Testing the improved ONS library preparation method

Next, we tested the efficiency of our modified protocol over the standard ONS library preparation (**Figure 5.14**), based on random fragmentation, by performing ONS on biological replicates of human DNA (N = 3 different gDNA, 5 technical replicates each, 15 in total).

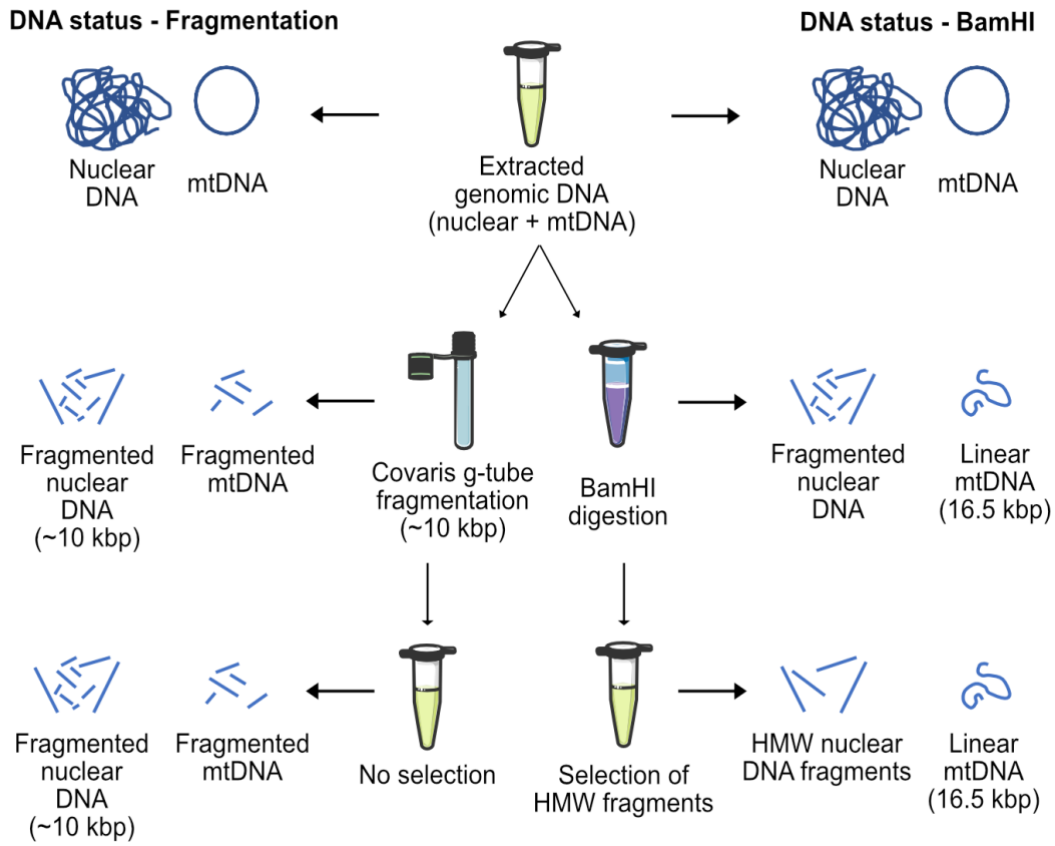


Figure 5.14: schematics of the fragmentation-based vs BamHI-based library preparation protocols. Overview of the workflow used to process samples using (left) standard ONT fragmentation protocol and (right) BamHI-based protocol.

Each gDNA was processed in parallel with both protocols. To reduce the possibility of NuMTs contamination, we performed a strict filtering on read lengths (selecting between 4 and 17 kbp) and per read quality (Phred ≥ 9) before the alignment, followed by secondary and supplementary alignments removal (**Figure 5.3**). While not altering quality parameters (percentage of identity and base quality per read; **Figure 5.15**), our filtering enriched for full length mtDNA sequences in all BamHI-treated samples.

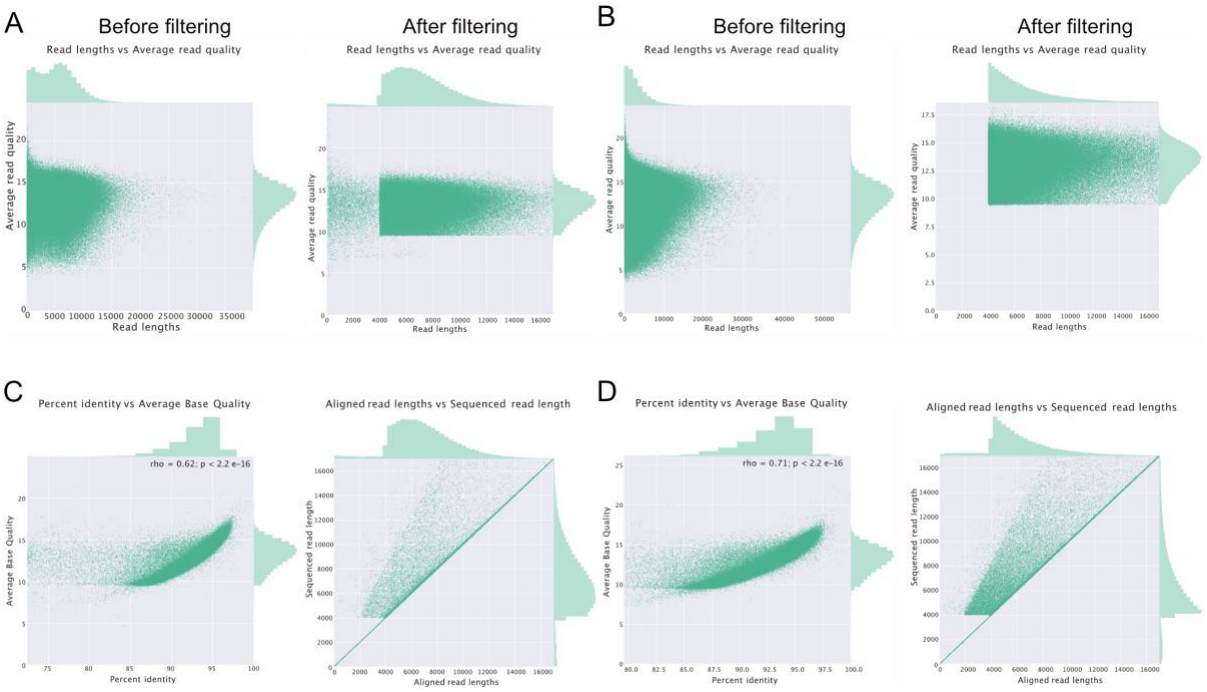


Figure 5.15: ONS quality controls parameters. a) Plots show the correlation between read lengths and read quality scores in one sample processed with the fragmentation protocol before filtering (left) and after filtering (right). b) Plots show the correlation between read lengths and read quality scores in one sample processed with the BamHI-based protocol before filtering (left) and after filtering (right). c) Plots show the correlation in one sample processed with the fragmentation protocol between percent identity to the reference sequence and average quality of the reads (left), and correlation between aligned read lengths and sequenced read lengths (right). d) Plots show the correlation in one sample processed with the BamHI protocol between percent identity to the reference sequence and average quality of the reads (left), and correlation between aligned read lengths and sequenced read lengths (right).

The enrichment of mtDNA in BamHI-treated samples was confirmed both by ddPCR and by analysis of the mapped reads. We found a higher percentage of mtDNA-

mapped reads in the BamHI samples compared to the fragmentation ones (Student's t-test P-value ≤ 0.05, **Figure 5.16**).

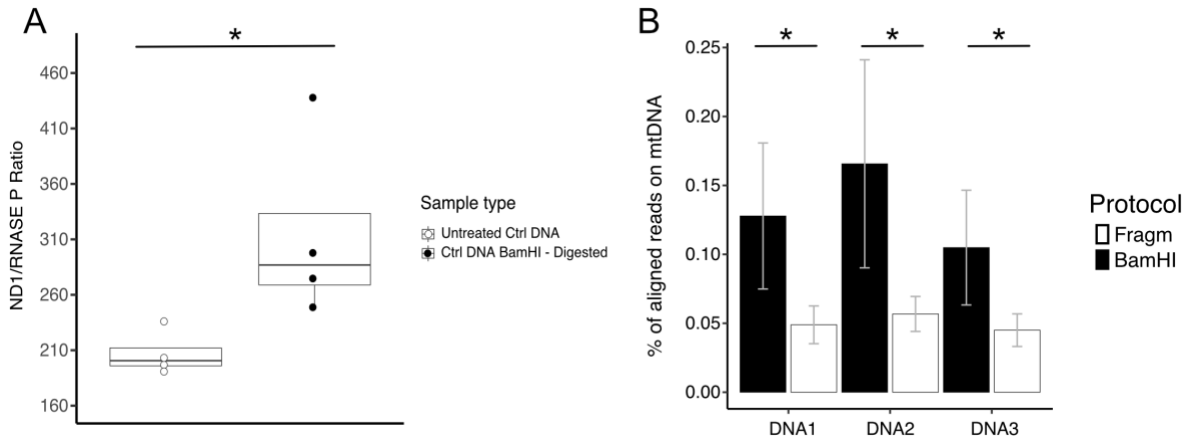


Figure 5.16: enrichment of mtDNA in BamHI-treated samples Vs fragmentation protocol. **a) Ratio of signal from the mitochondrial *MT-ND1* over *RNASE P* ddPCR probes in undigested genomic DNA and BamHI-digested genomic DNA. N = 4 for each protocol used. Star indicates significance (*: two-sided *P* ≤ 0.05, Wilcoxon test). b) Percentage of aligned reads on mtDNA and in fragmentation and BamHI sequenced samples (N=5 each). Stars indicate significance (*: two-sided P ≤ 0.05, Student's t-test).**

NuMTs contamination level was assessed by sequencing Rho 0 cells lacking mtDNA[278]. Results of this analysis showed that in 2 replicates of Rho 0 cells sequenced with ONS, of the 5488 mapped reads only 5 aligned to mtDNA (**Figure 5.17**). This suggested a low risk of mtDNA misalignment caused by NuMTs-derived sequences.

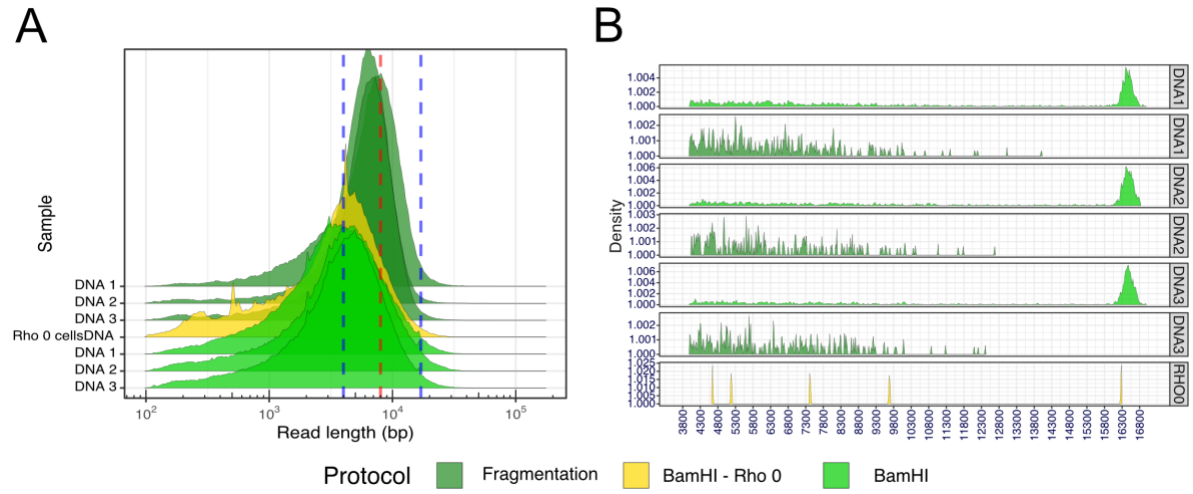


Figure 5.17: ONS reads distributions. a) Distributions of the total sequenced reads before alignment in 3 samples prepared with either fragmentation or BamHI-based protocols. Reads from Rho 0 cells (treated with BamHI) are highlighted in yellow. Blue dashed lines correspond to the cut-off for read filtering at 4000bp and 17000bp. The red dashed line corresponds to the human mtDNA genome length (16.5 Kbp). b) Distribution of the mtDNA reads filtered by length (4000bp -17000 bp) after alignment, in the same 3 samples prepared with either fragmentation or BamHI-based protocols. Reads from Rho 0 cells (treated with BamHI) are highlighted in yellow.

Under the conditions described, the fragmentation-based method showed an H-strand bias (L-strand$_{FRAG}$= 46.12% ± 5.13, H- strand$_{FRAG}$ = 53.87% ± 5.13, mean methylation ± sd; Anova one-way test P ≤ 0.001, **Figure 5.18 A** left panel, **Figure 5.18 B**, **Appendix 5**) with 6 samples having < 100% coverage (**Figure 5.18 C**). On the contrary, the BamHI-based protocol did not show any alignment bias (L-strand$_{BAMHI}$ = 50.67% ± 4.07, H- strand$_{BAMHI}$ = 49.32% ± 4.07, mean methylation ± sd; P = 0.36, Figure 5.18 A right panel, **Figure 5.18 B**, **Appendix 5**) or coverage bias (**Figure 5.18 D**).

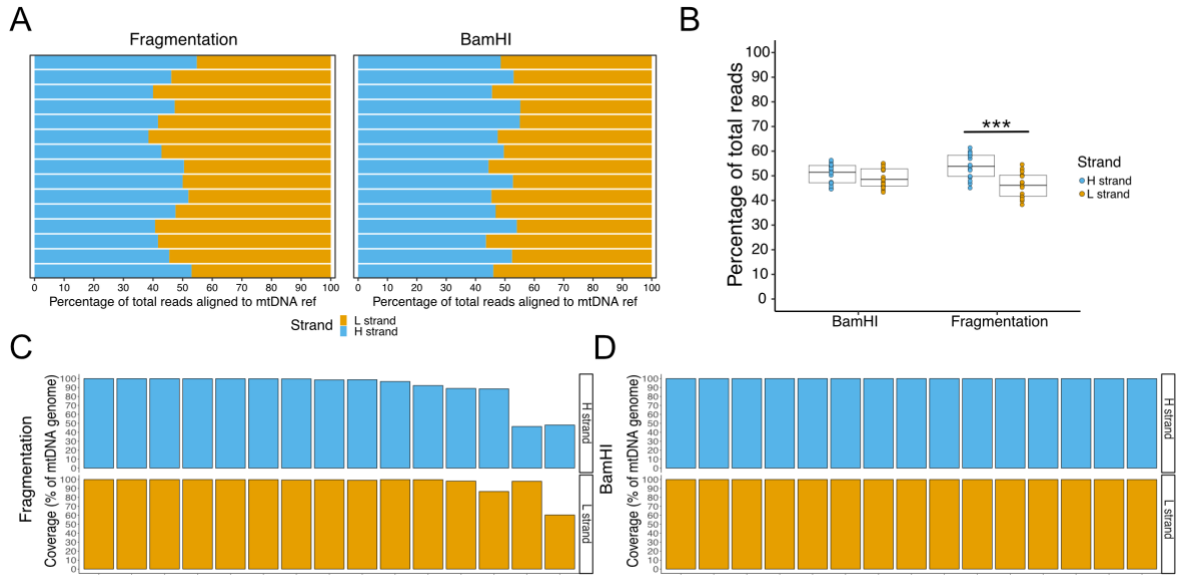


Figure 5.18: Alignment and coverage bias analysis of ONS samples. a) Percentage of reads aligned to the mtDNA reference per strand per biological replicate (N = 15 samples per protocol), in samples processed with (left) fragmentation protocol and (right) BamHI-based protocol. b) Percentage of reads aligned to mtDNA, divided by strand and library preparation protocol (N = 15 per protocol). Stars indicate significance (*: P ≤ 0.001, Anova one-way test). c,d) Percentage of mtDNA covered by at least 5 reads on the two mtDNA strands (H and L) per**

biological replicate (N = 15 samples per protocol), in samples processed with c) fragmentation protocol and d) BamHI-based protocol.

Average mtDNA read depth was higher in the samples processed with the BamHI-based protocol (Frag. = 23.83x ± 4.33, BamHI = 131.73x ± 8.15, mean ± sd; P = ≤ 0.0001, **Figure 5.19 A,B**, **Appendix 5**), with almost half of the mitochondrial reads mapped as full-length molecules (≥ 15 kbp; 42% ± 12 of BamHI reads Vs 2% ± 2 of Frag. reads, **Figure 5.17 B**).

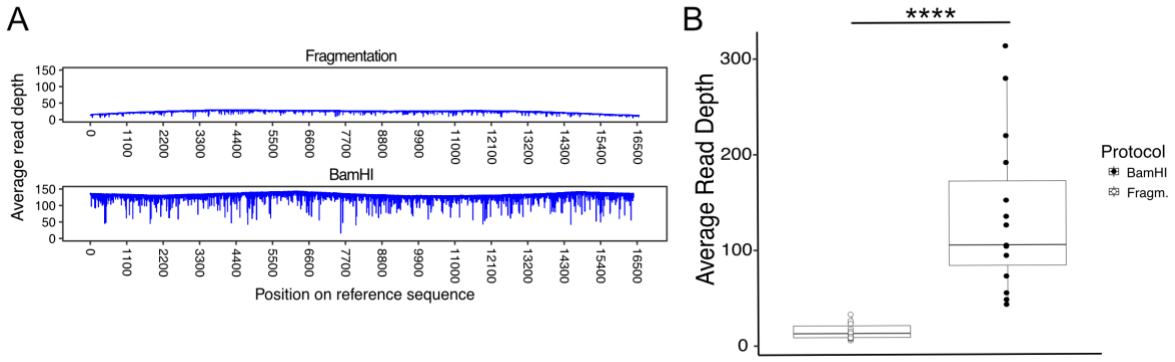


Figure 5.19: Read depth distribution of ONS sequenced samples. a) Distribution of the average read depth per mtDNA position in samples processed with (top) fragmentation protocol and (bottom) BamHI-based protocol (N = 15 samples per protocol). b) Average read depth per sample observed in the same sample pool processed using either fragmentation protocol (left) or BamHI-based protocol (right). N = 15 per protocol. Stars indicate significance (**: two-sided P = ≤ 0.0001, Wilcoxon test).**

5.2.6 Methylation calling results

Samples sequenced using a fragmentation-based protocol showed a greater range in average methylation levels (Min_{FRAG} : 5% - Max_{FRAG} : 33%) at low read depths levels (**Figure 5.20**; Min_{FRAG} : 9.16x – Max_{FRAG} : 55.62x), while the same samples processed with the BamHI-based protocol achieved similar methylation levels (Min_{BAMHI} : 2% - Min_{BAMHI}: 8%) at higher read depths (**Figure 5.20**; Min_{BAMHI}: 40.6x – Max_{BAMHI} : 306.2x).

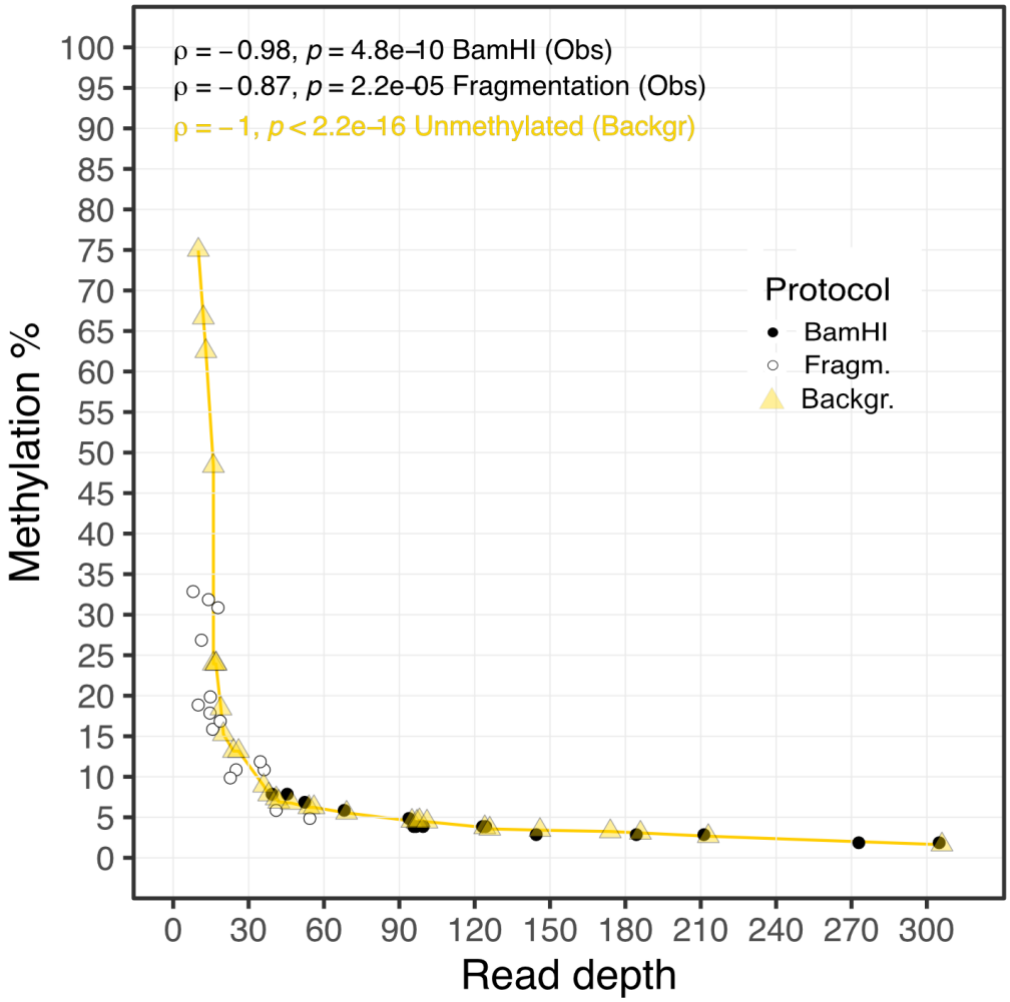


Figure 5.20: Average methylation and read depth profiles of ONS sequenced samples. **Correlation between average read depth and average methylation percentage in samples processed with fragmentation- and BamHI-based protocols (Obs) and in unmethylated datasets simulated from the negative control (Backgr). Circles represent a sample sequenced with either sequencing protocol (N = 15 per protocol). Triangles represent an unmethylated dataset simulated from the negative control. Spearman's rank correlation coefficient and two-sided P-values are shown.**

This inverse relationship between read depth and methylation levels was remarkably similar to the one obtained with WGBS data. However, this time we could not explain these results with an effect of bisulfite treatment on the DNA integrity. Instead, we hypothesised that these results may be explained by an effect of incorrect methylation calls introduced by Nanopolish (background methylation noise), which was more pronounced at lower read depths and got diluted out at higher read depths. To test this hypothesis, we first generated simulated unmethylated control data. To do that,

we selected reads at random from the NC BAM files, to generate 30 new BAM files ("Background") with a read depth matched to read depths obtained with both fragmentation and BamHI-based protocol-derived experiments ("Observed") (**Figure 5.20**). We then called the methylation again with Nanopolish on these simulated samples to determine the effect of the baseline methylation "noise". Indeed we could confirm that we observed a similar inverse relationship between average read depth and average methylation level (**Figure 5.20**). Based on this data we were then able to infer a model that best fitted the simulated data, which we used to estimate the background noise in methylation calling of all the ONS experiments performed in this study (R^2 = 0.94, **Figure 5.21**, **Methods**).

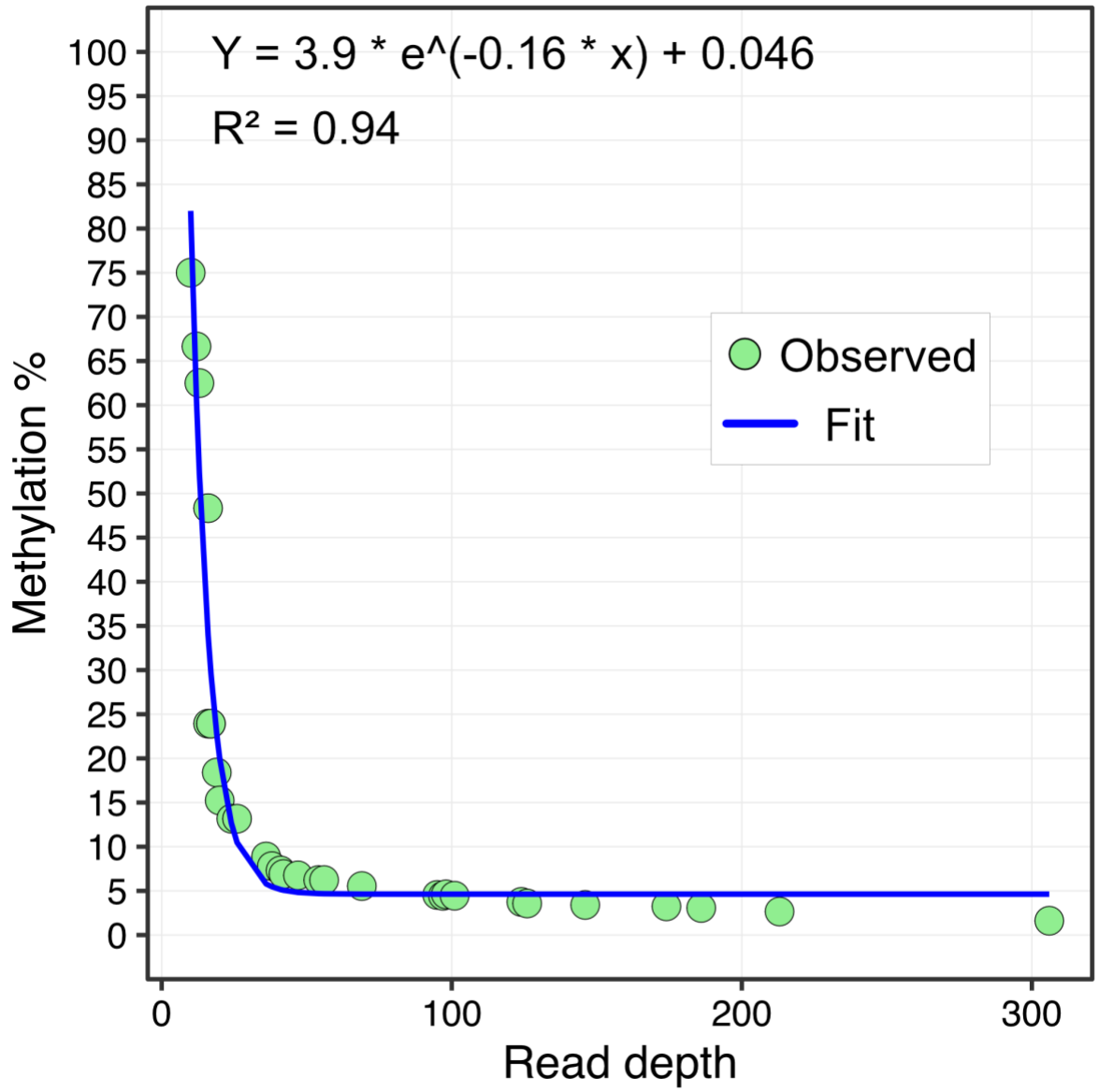


Figure 5.21: ONS methylation error modelling. **Green circles represent the average methylation percentages calculated from 30 ONS sequencing of simulated datasets (derived from negative controls, therefore expected to be unmethylated), in relation to their read depth. The blue solid line represents the fitted line (exponential decay function) that describes such distribution, corresponding to the methylation background noise (Backgr). The formula describing the fit and R^2 correlation calculated with the goodness of fit test are shown.**

5.3 Conclusion and discussion

In our effort to find a suitable alternative to WGBS and other single-base resolution methods for studying mtDNA methylation, we explored the suitability of the third-generation sequencing technology ONS. We chose ONS as it does not require neither bisulfite treatment nor PCR amplification to obtain information on the DNA methylation status, and because it was the most viable solution for us at the time this analysis was performed. Moreover, since ONS allows the sequencing of ultra-long reads (potentially > 100 kbp[275,277]), this allows us to sequence the full mtDNA molecule as a single read, and potentially obtain simultaneous information on SNVs and CpG methylation, as well as phasing of SNVs and methylated residues on the single molecules.

For this reason, we first tested first the accuracy of the methylation calling of the Nanopolish software by using custom negative and positive controls sequenced with ONS, as shown by Simpson and colleagues[198]. Results show that the accuracy level at the default methylation calling threshold of ≥ 2.5 LLR was 97.7%. Therefore, we decided to increase the calling threshold to ≥ 5 LLR, to obtain a better accuracy of 99%. We identified 13 false positives in the negative control. These positions were found to be methylated at the same level irrespective of the sample analysed. They were deemed false positives and removed from any further analysis.

We were also interested in avoiding the requirement of isolating mitochondria from biological samples to sequence pure mtDNA for sequencing. We wanted to do that firstly because this procedure usually requires a large amount of starting material (either tissues or cell lines), and secondly because of the great variability in the available procedures[279], which may affect the purity of the mtDNA preparation and hence the methylation calling. We tested two different methods for mtDNA enrichment from gDNA. The first one, published by Jayaprakash and colleagues[245], based on the digestion of linear nDNA by exonuclease V, resulted in a better purity of the remaining circular mtDNA, as measured by ddPCR. However, because of the low mtDNA

recovery yield from 1 µg of gDNA, the very long incubation times required and the requirement of one additional linearisation step to sequence linear mtDNA, we opted for the alternative method that we tested. This method is based on the digestion of gDNA with BamHI, which cuts mtDNA once and nDNA multiple times.

Finally, we tested our new library preparation method against the standard ONS library preparation protocol based on random fragmentation of gDNA (without any additional purification or enrichment steps before sequencing). Contrary to what we observed in WGBS samples, we found a low L-strand and coverage bias in samples sequenced with the fragmentation-based method, while no bias of any form was observed in the samples prepared with the BamHI-based protocol. Mitochondrial read depth was also significantly higher in BamHI-treated samples, with 42% of the mitochondrial reads longer than 15 kbp (versus only 2% of the mitochondrial reads in samples processed with the fragmentation protocol), as expected.

When comparing the average mitochondrial methylation level observed in samples sequenced using the two protocols with their relative mitochondrial read depth, we found an inverse correlation similar to the one observed in WGBS samples (**Figure 4.3 C**). However, it was clear that this time we could not explain this relationship with differences in the degree of bisulfite resistance, as we only sequenced native DNA. We therefore hypothesised that these data could be explained by the presence of a baseline methylation calls that are wrongly called as methylated by Nanopolish (thus representing background methylation error), and that could explain the average methylation frequency observed especially when the read depth is low. This could be shown by calculating the methylation levels of simulated unmethylated data at read depths matching the ones from the real samples analysed. The data derived from this simulation support the hypothesis that a baseline level of incorrect methylation calls must be taken into account when quantifying the mtDNA methylation on ONS-sequenced “real” biological samples, especially when read depth is low. This analysis also allowed us to infer a model describing the relationship of background methylation error with sequencing read depths, which we used in subsequent experiments to remove noise from the true methylation signal. Moreover, the fact that at higher “real” read depths (observed by sequencing with ONS control cell lines) we still observe an average methylation close to 0 adds to the evidence supporting the absence of mtDNA methylation (**Figure 5.20**). Moreover, we were able to infer a model of the expected

background error, to determine also in future experiments whether the observed signal could be ascribed as noise.

Chapter 6. BamHI-based method mtDNA sequencing with ONS for mitochondrial variant calling and CpG methylation analysis on human cell lines, primary fibroblasts, and tissue DNA

6.1 Introduction

6.1.1 mtDNA homoplasmy, heteroplasmy and mitochondrial haplogroups

One of the main characteristics of the mitochondrial genome of all organisms is to be present in multiple copies per cell [43,282]. The identity of all mtDNA sequences is a condition known as homoplasmy. Identification of homoplasmic variants in the human mtDNA allows the definition of different haplogroups: mtDNA molecules that share the same haplotype, are inherited by the maternal line, and derive by descent from the ancestral mtDNA molecule[283]. From a phylogenetic perspective, an haplogroup corresponds to a specific branch (or clade) in the human phylogenetic tree and its ancestral mutational pattern is situated on a branch node[284]. Mitochondrial haplogroups have been studied in population genetics to model the history of human migrations out of Africa[72]. The opposite condition to homoplasmy, known as heteroplasmy, is the presence of more than one mtDNA species, defined by their sequence[285]. Heteroplasmy of pathological variants may cause rare mitochondrial disease when their level surpasses a certain threshold[44]. Technologies used to study mitochondrial variation include PCR-based methods (southern blotting, qPCR, etc.[286–293]) microarrays (for a targeted approach where known variants are investigated[294]), pyrosequencing[291], and next-generation sequencing (NGS) technologies for the identification of unknown variants[295,296].

6.1.2 Origin of NGS error rates

The digital output of NGS is represented by FASTQ files containing short sequences, known as “reads”, of varying lengths, depending on the technology employed[296].

The FASTQ files are text-based files used to store information on both the nucleotide sequence itself and corresponding quality scores, expressed with a Phred-like formula[297]. This score is related logarithmically to the error probability (P) of the base-calling, and inversely correlated to the error rate. Therefore, high Phred scores correspond to more accurate base calls. This measure of the quality of the sequencing data is one of the main advantages of this technologies compared to others. In

addition, NGS technologies provide high scalability, as multiple samples can be sequenced at the same time thanks to the use of molecular indexes to tag specifically the individual samples.

However, every NGS technology has an intrinsic error late level that influences the heteroplasmy calling accuracy, that may depend on the specific type of NGS chemistry, or on the DNA sequence context. For examples, in the Illumina Miseq the fluorophores that label the A and C bases have the highest intensities and are identified through the same channel, thus leading to a higher susceptibility of A to C substitution errors[298]. This effect is more pronounced in case of homopolymeric stretches of the same repeated nucleotides, that may lead to incorrect insertions/deletions[299–301].

On top of this, incorrect proofreading of the DNA polymerase used for PCR can increase the miscalling rate. For Illumina sequencing, this could happen both during the final PCR amplification step during library preparation, during cluster generation, or during sequencing itself[302]. Various bioinformatic strategies can be implemented to correct for these PCR errors in NGS.

6.1.3 Mitochondrial variant calling

The first step in the bioinformatic analysis targeted at mitochondrial variant calling is the read alignment (or “mapping”) and genome assembly. In NGS, due to the nature of the reads, the alignment of short fragments only to either the mtDNA or nDNA reference is prone to the generation of misaligned reads and false positives in subsequent mtDNA variant calling[303,304]. Therefore, it is recommended to perform the read alignment on both genome sequences simultaneously, to detect and remove possible NuMTs sequences[304,305]. These strategies are already commonly used by most of the available mtDNA variant calling tools, included MToolBox (used in this study)[255].

The following step is mtDNA variant calling, aimed at identifying variants and quantify their heteroplasmy levels. The available pipelines may vary slightly in their quality procedures and calling methods of a valid set of mtDNA heteroplasmic variants. However in general, heteroplasmy is calculated as the ratio of the read depths of the allele on the total read depth per position. The cut-offs used to call a variant are arbitrary and may vary between the pipelines. The pipeline used in this study, of which results are presented in this chapter, MToolBox[255], adopts a default read depth value

of ≥5 and a per-base quality score of ≥25 to call a mitochondrial variant, and also excludes small insertions/deletions close to read ends, where quality tends to be lower.

There is no pre-defined set of rules or agreement on what specific heteroplasmy thresholds are to be used in mtDNA variant calling analysis. However, the sequencing metrics, like read depth, may help choosing the most appropriate cut-off for heteroplasmy analysis. In general, a high read depth (≥1000x) corresponds to higher sensitivity, and enhanced detection of low heteroplasmic variants (≤1%). Despite this, some false low heteroplasmy variants may arise from the presence of rare unknown NuMTs sequences not excluded in the alignment step or sequencing errors that have failed the quality filtering, despite stringent quality controls[304,306]. For this reason, to reduce false positives higher heteroplasmy cut-offs (≥1-5%) would be preferred, despite resulting in the exclusion of true low-frequency variants[303]. Skipping this quality control may in fact result to susceptibility to generation of flawed data[307].

6.1.4 MtDNA variant calling with third-generation sequencing technologies

As mentioned in **paragraph 1.6.5**, third-generation sequencing technologies such as PacBio or ONS can pose a valid alternative to short-read based NGS technologies, enabling the detection of mtDNA at the single-molecule level[296]. The opportunity offered by these technologies, which is being explored in this study, is to be able to capture and sequence full-length native mtDNA, to simultaneously perform variant calling, variant phasing and methylation calling of CpG residues. Alignment of the full-length sequence may lead to higher mitochondrial variant calling specificity, as the risk of aligning NuMTs is reduced. Moreover, by sequencing native DNA, the risk of introducing PCR artifact is virtually nullified, therefore increasing the mtDNA variants detection accuracy (especially for low-level heteroplasmy).

However, these approaches have several possible disadvantages. First, third-generation sequencing PCR-free protocols require a high amount of input DNA (1-5 µg) to achieve sensible read depths (although for mtDNA this requirement might be lower). Secondly, as we discussed in **paragraph 1.6.7.2** of the introduction, one of the major issues of ONS specifically is the high error rate arising from the interpretation of the electric signal variations over time. This is exacerbated in homopolymeric regions, and could represent a major setback in the usage of ONS for mtDNA variant

calling (although it could probably represent a good technology for the identification of large mtDNA re-arrangements).

6.1.5 Final remarks

The work detailed in **chapter 5** was aimed primarily at establishing a reliable method to consistently measure mtDNA methylation using ONS. In the first section of this chapter we tested for the first time to our knowledge the efficacy of ONS for mtDNA variant calling, using a modification of the MToolBox bioinformatics pipeline[255], and samples sequenced using gold-standard Illumina sequencing as control. This analysis was performed by Dr. Claudia Calabrese. With this analysis we want to unravel the potential of ONS for mtDNA variant calling, and outline its limitations.

Next, using the library preparation method developed in the previous chapter, based on BamHI-cutting, we assessed the presence of mtDNA methylation in human cancer cell lines with different mitochondrial haplogroups, primary fibroblasts with or without a pathological mitochondrial variant, and in different human tissues.

6.2 Results

6.2.1 Comparison of mtDNA ONS variant calling with Illumina Miseq

In this first section, we analyse and compare results from variant calling performed with Illumina sequencing Vs ONS. To test this, we sequenced DNA from cells with known mtDNA sequences. First, we sequenced DNA from 3 trans-mitochondrial osteosarcoma cybrids with mtDNAs belonging to 3 different human haplogroups[283] with an identical nuclear background[308] (N = 5 technical replicates of 3 independent DNA from the mitochondrial haplogroup H1, J1c and J2, respectively; “613H”, “128J”, “135J2; **Appendix 1**). Then, we sequenced mtDNA from primary fibroblasts including healthy control subjects without known mtDNA mutations (“Control 1”, “Control 2”), and one patient carrying the heteroplasmic m.8344A>G/MT-TK (“m.8344A>G”), causative of myoclonic epilepsy with ragged red fibers (MERRF) syndrome and 2 patients carrying the m.3243A>G/MT-TL1 mutation (“m.3244A>G”), known to cause Mitochondrial Myopathy, Encephalopathy, Lactic Acidosis, Stroke-Like Episodes (MELAS[309,310]; N = 3 technical replicates, **Appendix 1**). Additionally, we sequenced human tissues (Liver, Kidney, Heart, Muscle) from 7 different healthy individuals (**Appendix 1**).

To detect mtDNA variants and quantify their heteroplasmy, we identified mtDNA variants in ONS-sequenced samples and used high depth Illumina MiSeq sequencing of mtDNA for validation (mean read depth = 2,769x, min = 318x, max = 5,559x , **Appendix 6**). Illumina Miseq was performed by Dr. Zoe Golder and variant calling by Dr. Claudia Calabrese, and conducted on human cell lines and primary fibroblasts only (where either the mitochondrial haplogroup or the mtDNA mutation was known) to seek confirmation with both Illumina and ONS techniques. First, variant calling with ONS detected 99.5% (N = 739/743) of the homoplasmic variants (het. ≥ 95%) also detected by the Illumina sequencing. Then we confirmed that ONS was able to correctly predict haplogroups of all 613H, 128J and 135J2 biological replicates. It was also able to identify the known single nucleotide heteroplasmic variants in most biological replicates of primary cells derived from MERRF and MELAS patients (N= 2/3 for m.3244A>G (1), N=3/3 for m.3244A>G (2), N=2/3 for m.8344A>G; **Appendix 6**, **Figure 6.1**).

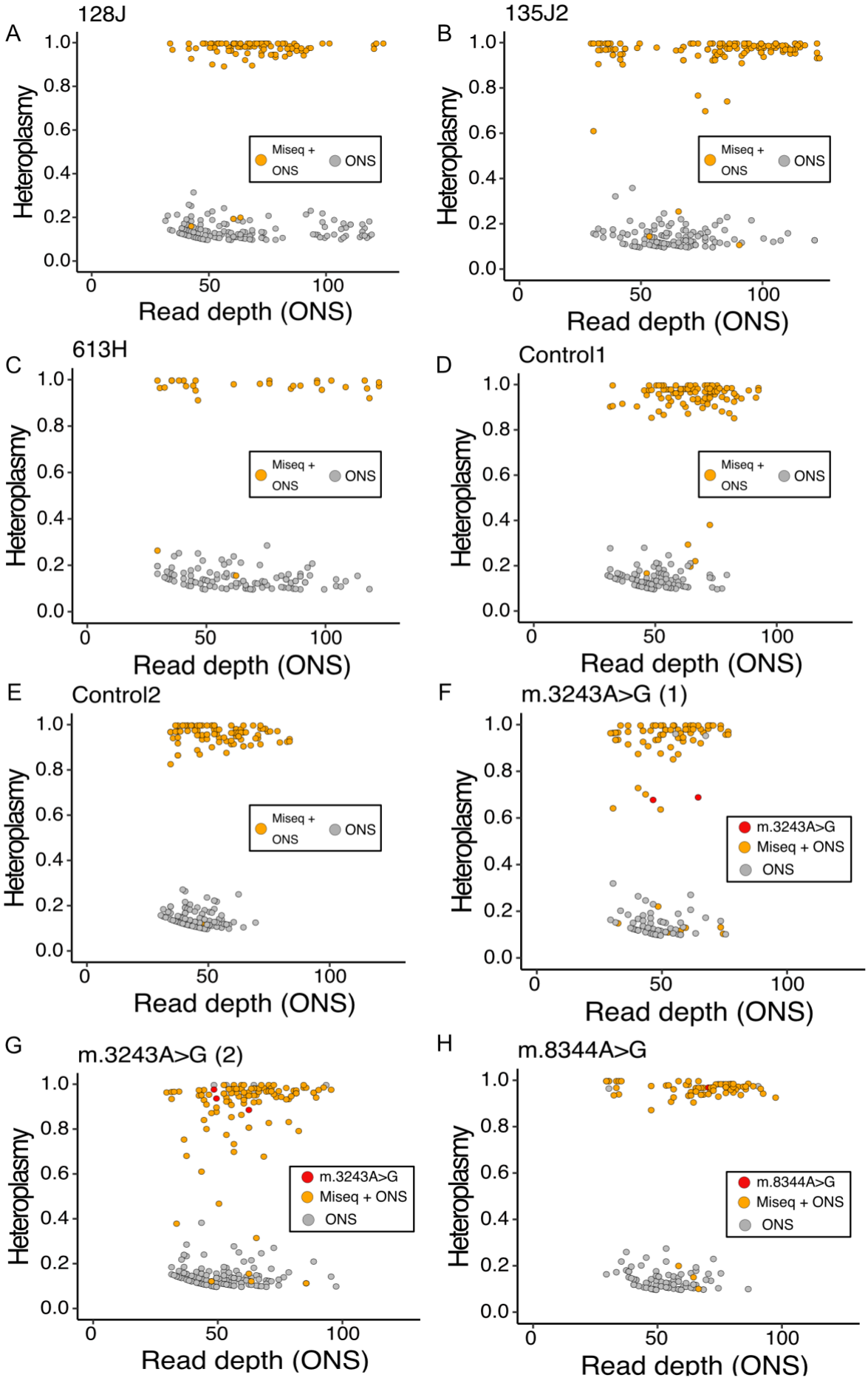

A
128J
B
135J2
C
613H
D
Control1
E
Control2
F
m.3243A>G (1)
G
m.3243A>G (2)
H
m.8344A>G
Heteroplasmy
Read depth (ONS)
Miseq + ONS
ONS
m.3243A>G
m.8344A>G
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100

Figure 6.1: Comparison of ONS and Illumina sequencing variant calling. a-h) Scatterplots show the mtDNA heteroplasmy quantified with ONS in each sample as a function of the read depth per position. Variants confirmed also by Illumina Miseq are highlighted in orange and in red (marking m.3243A>G and m.8344A>G mtDNA mutations).
MtDNA variants shown have been aggregated across biological replicates per sample sequenced with ONS (N = 5 for 613H/128J/135J2 and N = 3 for Control1/Control2/m.3243A>G(1)/m.3243A>G(2)/m.8344A>G). The m.3243A>G mutation was confirmed by Illumina Miseq but identified with ONS in two out of three biological replicates of the m.3243A>G (1) sample and in all the three replicates of the m.3243A>G (2) sample. The m.8344A>G mutation was confirmed by Illumina Miseq and identified by ONS in two out of the three biological replicates of the m.8344A>G sample.

Because we observed a base calling accuracy of ~90% in our samples sequenced with ONS (**Figure 5.15**), we set a stringent threshold of ≥10% heteroplasmy to call for mtDNA variants. On average, we identified 60 mtDNA variants with ≥ 10% heteroplasmy per sample with ONS alone, of which 28 (~47%) were confirmed with Illumina Miseq (**Figure 6.2** right plot). These were mostly highly heteroplasmic or homoplasmic variants ($heteroplasmy_{ONS}$ = 93% ± 17%; $heteroplasmy_{Miseq}$ = 96% ± 15%; mean ± sd, **Figure 6.2 A** left plot). The remaining ONS-only mtDNA variants were mostly low heteroplasmic ($heteroplasmy_{ONS}$ = 16% ± 11%, mean ± sd; **Figure 6.2 A** left plot). Heteroplasmies calculated with ONS overall tended to correlate better with Illumina at higher read depths (**Figure 6.2 B**).

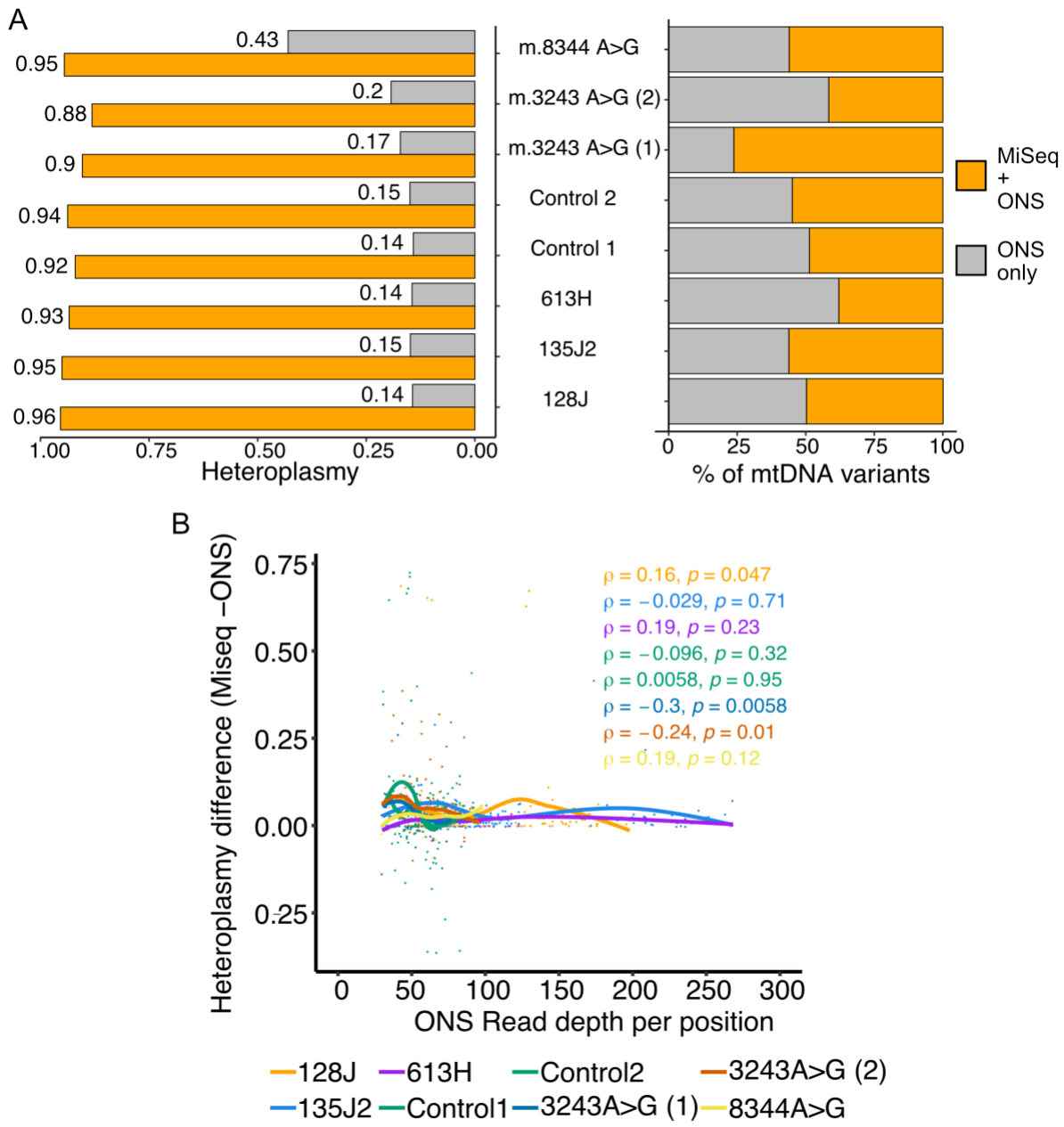


Figure 6.2: ONS-based variant calling of mtDNA. **a) Variant calling statistics per each cell line analysed. Heteroplasmy (left) and percentage of single nucleotide mtDNA variants (right) identified with either Illumina Miseq and ONS or ONS only. ONS values are means calculated across all biological replicates per each cell line analysed (N = 5). b) Scatterplot showing a correlation between differences in heteroplasmy values quantified with Illumina Miseq and ONS (calculated as Miseq heteroplasmy - ONS heteroplasmy), for each single nucleotide mtDNA variant detected with both techniques, and ONS read depth per position. Colours correspond to the different samples analysed (N = 5 for 613H/128J/135J2 and N = 3 for Control1/Control2/m.3243A>G (1)/m.3243A>G (2)/m.8344A>G), with lines indicating mean over all the data points in each sample (calculated using the "loess" *geom_smooth* R function). Spearman's rank two-sided P-values and rho coefficients are shown.**

6.2.2 ONS-based CpG methylation analysis of mtDNA in human cell lines and tissues reveals absence of CpG methylation

Next, we performed CpG methylation calling followed by differential methylation (DM) analysis[191], in all the cell lines and primary fibroblasts, using as baseline the 613H cell line and the control fibroblasts, respectively. We first checked the methylation levels of the 13 possible false positives we had identified in the NCs (chapter 5), and we found them to be methylated at the same level in the samples analysed (**Appendix 3**). Therefore, we removed the 13 false positives positions from all the ONS sequencing experiments results. In the cell lines, the DM analysis revealed 3 differentially methylated CpGs (DM-CpGs): one found only in the haplogroup J2 cells (m.16360), and two found in both J cell lines (m.10400 and m.16128; **Figure 6.3** top graph, **Appendix 7**). We also found 5 DM-CpGs in all the primary fibroblast (m.4919, m.9195, m.10400, m.15925, m.16128; **Appendix 7**). However, a comparison of these DM-CpG with the variants identified in our analysis, revealed that an haplogroup-defining variant always fell within a ± 5 bp window from a DM-CpG. This prompted us to hypothesize that these variants may alter the Nanopolish methylation calling. In fact, the software compares the signal from a 11 bp window of ± 5 bp around the CpG with the expected signal coming from the trained model. We hypothesised that the presence of a variant inside this 11 bp window may alter the signal in a way that could be misinterpreted by Nanopolish. To test this, we generated a new reference for methylation calling based on a mtDNA consensus sequence built on major mtDNA alleles identified with Illumina MiSeq sequencing (Methods). As expected, DM analysis repeated using consensus sequences-corrected methylation calls returned no significant differences in methylation levels between the samples, indicating that the previously identified possible DM-CpGs were artefacts of the Nanopolish calling algorithm (**Figure 6.3** bottom graph, **Appendix 7**).

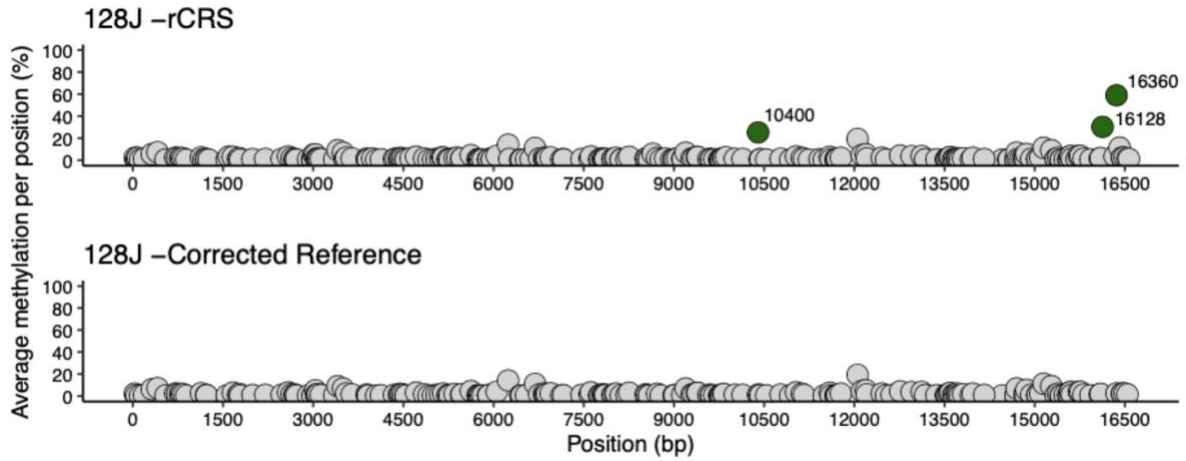


Figure 6.3: Differential methylation analysis results. Example of methylation calling artefacts introduced when using hg38 as reference (which includes the mitochondrial reference sequence rCRS) (top) instead of a sample-specific consensus sequence (bottom). In green are highlighted the sample-specific differentially methylated positions which disappear upon reference correction

Using a sample-specific mtDNA reference sequence for methylation calling, we measured consistently low methylation levels in all the cell lines and primary fibroblasts analysed ($Methylation_{C\_LINES/FIB}$= 1.3%-2%, min-max; **Figure 6.4**). All the measured methylation values were found to be below our estimated background noise (**Figure 6.4**).

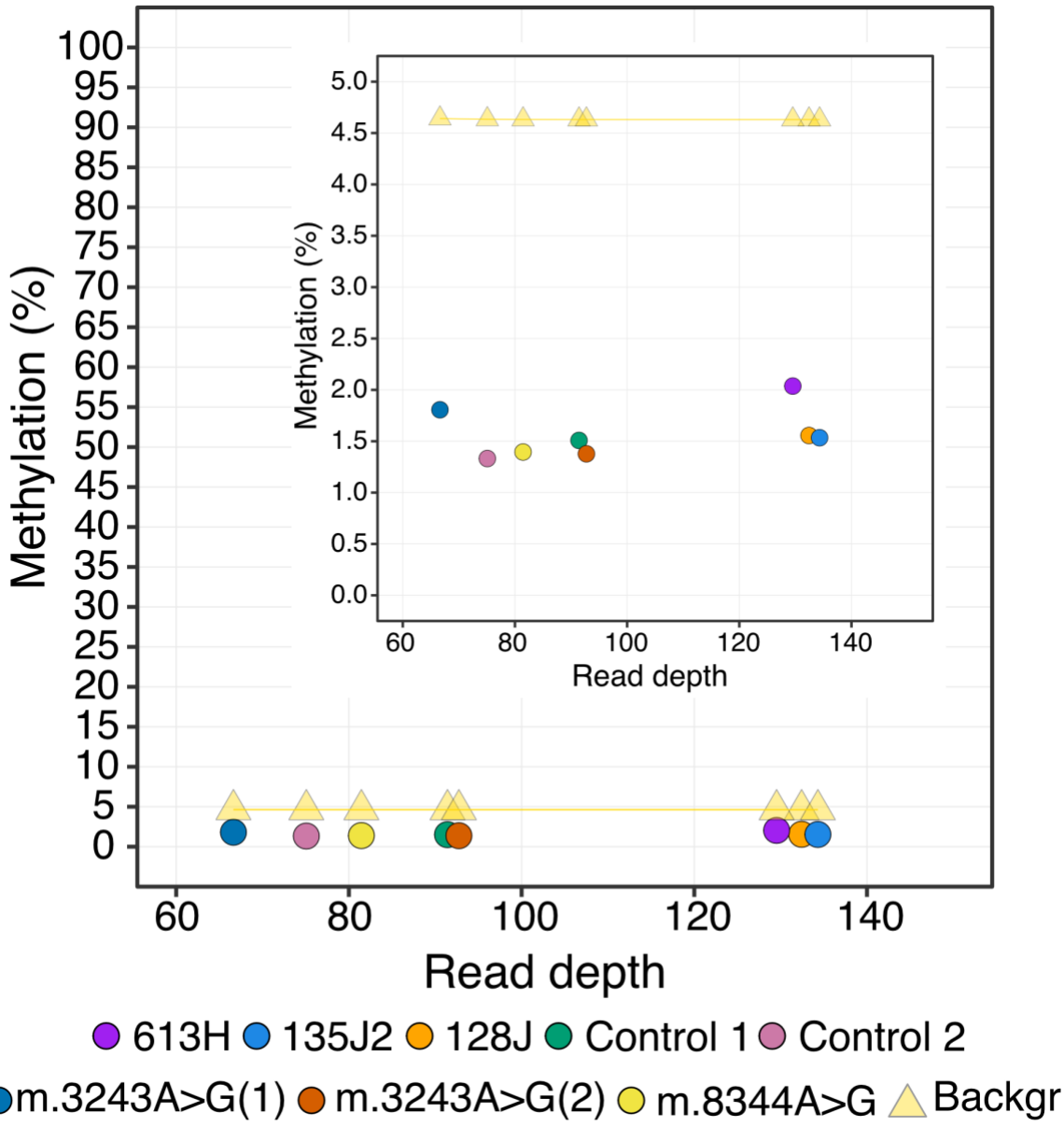


Figure 6.4: methylation analysis results on cell lines and primary fibroblasts. **Scatterplot showing the relationship between average read depth and average methylation percentage in samples processed with BamHI protocol. Circles represent an average of al mtDNA position in either 5 (cell lines) or 3 (primary fibroblasts) biological replicates. Yellow triangles represent the background noise. Inset plots show magnification of the data shown. Spearman's test, *p-value* and Rho are shown.**

The biological samples utilised for the analysis showed so far were generated in our laboratory and were cultivated exclusively *in vitro*. We therefore decided to look for additional evidence of mtDNA methylation presence in a more physiological context, sequencing mtDNA from post-mortem human tissues of 7 different healthy individuals (**Appendix 1**). However, once again we observed that the methylation levels were

extremely low and below our estimated background noise, even at higher read depths compared to the cell lines (**Figure 6.5**; $Methylation_{TISSUES}$= 0.6%-0.8%, min-max).

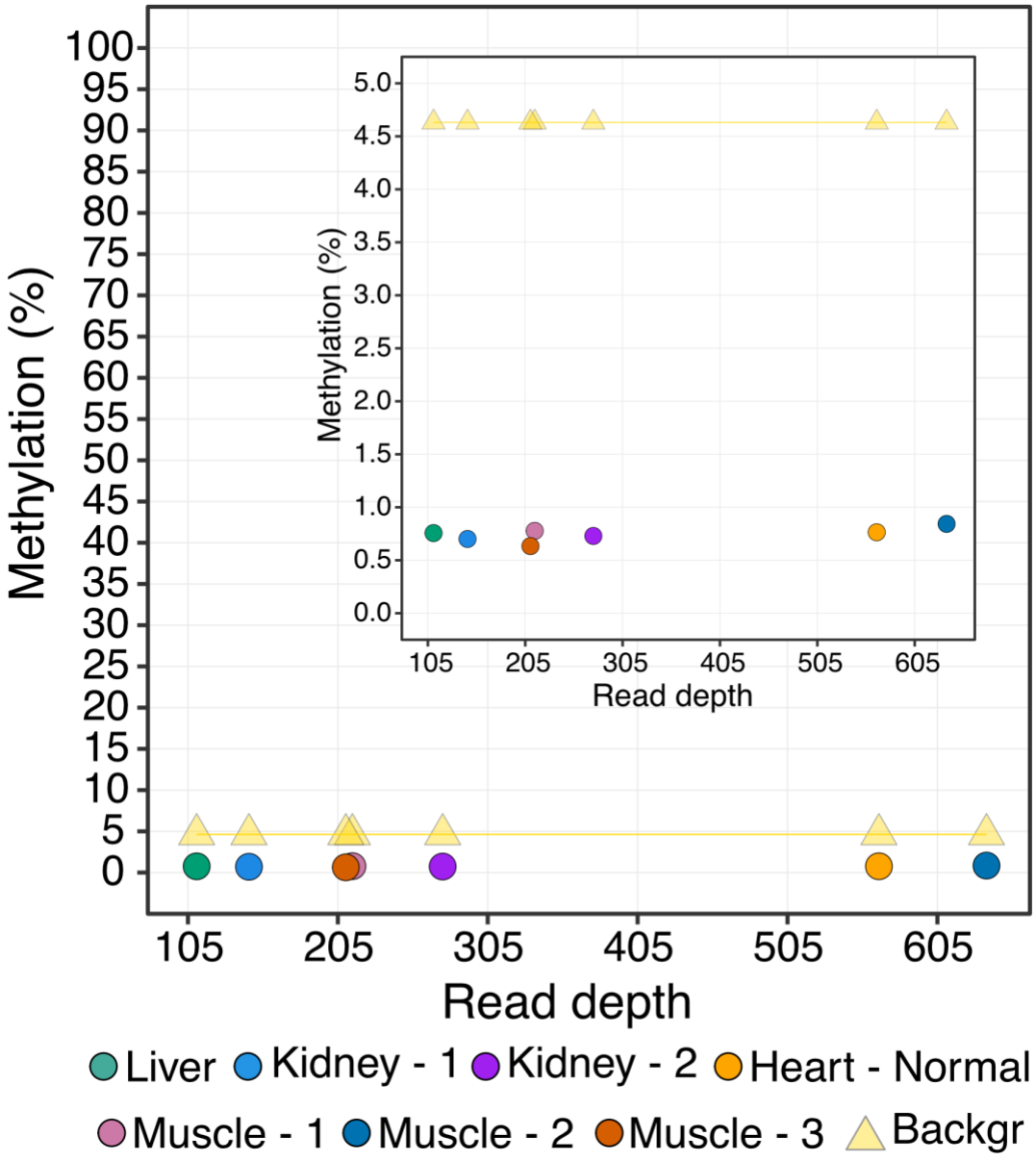


Figure 6.5: methylation analysis results on human tissues. **Scatterplot showing the relationship between average read depth and average methylation percentage in samples processed with BamHI protocol. Circles represent an average of al mtDNA position in human tissues of 7 different individuals. Yellow triangles represent the background noise. Inset plots show magnification of the data shown. Spearman's test, *p-value* and Rho are shown.**

Since we could not identify any significant methylation in the human samples analysed so far, we sought for conclusive evidence that ONS can identify methylation above the background level. To do that, we generated and sequenced with ONS 4 additional

positive control samples with expected methylation levels of 5%, 25%, 50% and 75%, generated by mixing the PC and NC at different percentages. Results of this analysis revealed that the expected methylation levels could be indeed correctly detected with ONS (Rho=1, P=0.003, Spearman's rank test, **Figure 6.6**). Therefore, we concluded that the low methylation levels observed in all the biological contexts analysed were artefacts, and that CpG methylation is not present in human mtDNA.

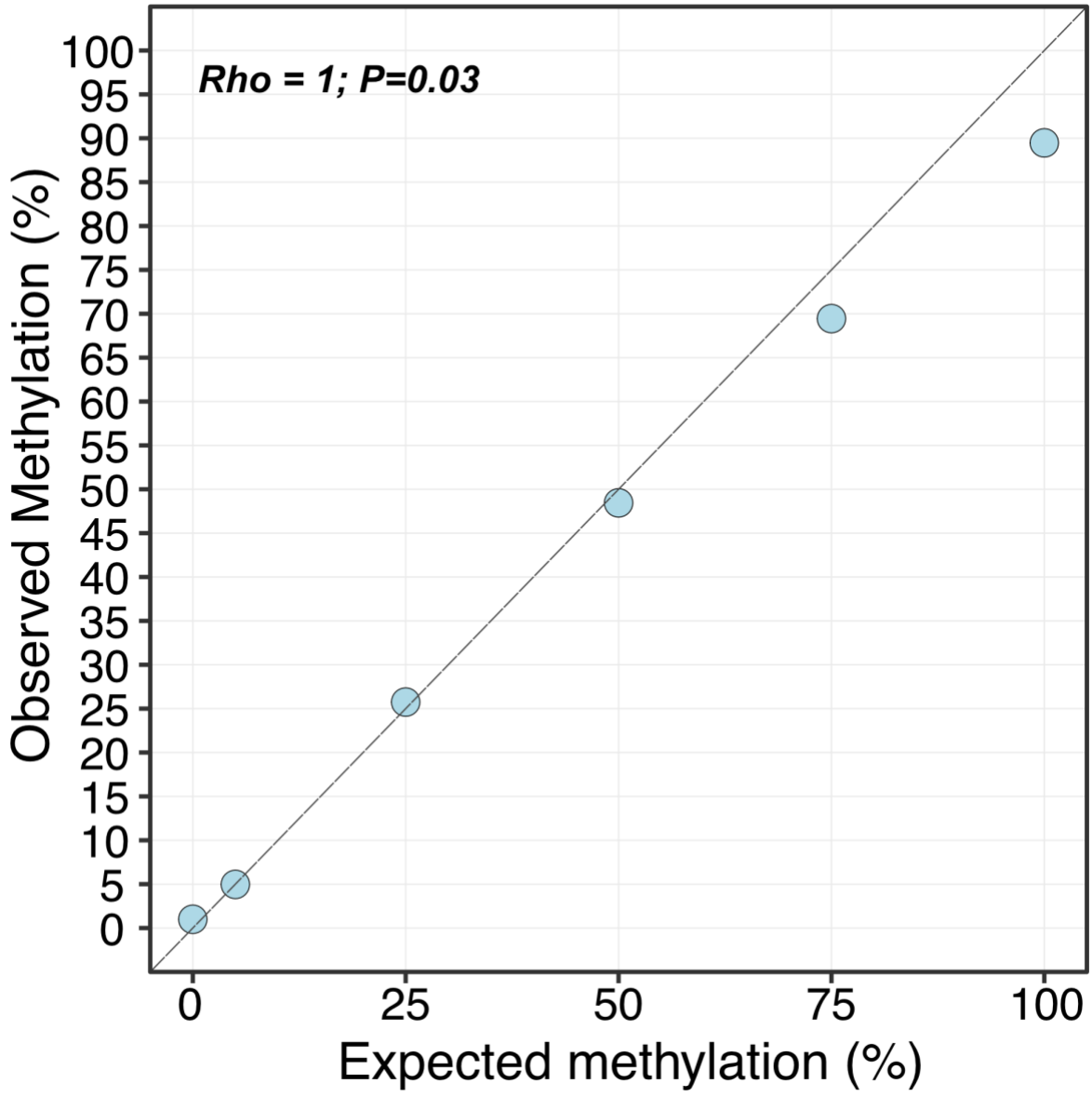


Figure 6.6: Positive controls methylation results. **Correlation between the expected and observed methylation levels calculated on methylated controls generated by mixing PC and NC. Spearman's test, p-value and Rho are shown.**

6.3 Conclusions and discussion

Variant calling is crucial for our understanding of the human genome variation[311]. Compared to gold-standard Illumina sequencing (and other short-read based technologies), long-read sequencing-based ONS has the advantage of having a reduced GC-bias and does not require PCR amplification[276]. This makes this technology useful to analyse DNA sequences that were previously notoriously difficult to study[275]. Because of its polyploid status and the presence of NuMTs, variant calling on mtDNA is particularly challenging[312]. Here, for the first time to our knowledge, we tested the efficacy of ONS in identifying mitochondrial genome variants. Using samples with known haplogroups and/or known mitochondrial disease-causing mutations, our analysis performed against gold-standard Illumina sequencing revealed that ONS can correctly identify variants with an heteroplasmy ≥90%, such as homoplasmic haplogroup-defining variants or the pathological mutations of our test cell lines and primary fibroblasts, respectively. However, ONS revealed to be unreliable in assessing low heteroplasmic variants, as proven by the estimated levels of heteroplasmy of the SNV variants identified in both ONS and Illumina sequencing (**Figure 6.1**), and the fact that low heteroplasmic variants were for the majority only found with ONS and not confirmed by Illumina Miseq. This latter finding was somewhat expected, given the high rate of incorrect insertions/deletions known to be systematically introduced by ONS near homopolymeric stretches[313]. Also, the observation of an inverse relationship between the difference in heteroplasmy quantified at the same time with Illumina and ONS and the mitochondrial read depth measured with ONS suggests that with increasing mtDNA read depths the ONS errors introduced in variant calling is smoothed out, and the accuracy of the heteroplasmy quantification increases (**Figure 6.2 C**). This implies that adjustments to the ONS protocol aimed at reaching higher read depths (e.g. longer sequencing times, higher starting sample material, etc) can improve heteroplasmic mtDNA variants identification. These observations reveal that ONS technology is still lacking behind Illumina sequencing regarding at least mtDNA variant calling. It is clear from our analysis that not only ONS sequencing of native DNA cannot compare to Illumina for the correct estimation of high heteroplasmic variants (≥ 90%), but it is very unreliable for low heteroplasmic variants. However, it also seems likely that high ONS read depths could help to ameliorate mtDNA variant calling, and this point may be crucial:

the analysis we performed in this study was specifically aimed at estimating methylation, with the variant calling being tested as a secondary aim. Because of this, we were obligated to sequence native DNA (and to try to enrich as much as possible for native mtDNA molecules). Therefore, while this limited us to achieving high read depths, on the other hand does not preclude that if mtDNA variant calling analysis is the main objective of the research, PCR amplification before sequencing may be used to achieve higher read depths that could lead to more accurate results.

DM calling analysis on human cell lines and primary fibroblasts was initially performed following the guidelines from Gigante and colleagues[191]. In this study the authors could successfully identify different methylation patterns in genes inherited by either of the two parental nuclear chromosomes, demonstrating that a combination of phasing of DNA variants and methylation calling can be simultaneously achieved with ONS. Similarly, our analysis initially revealed a few DM-CpGs in the samples studied. However, a deeper scrutiny of the variants positions revealed the possibility that homoplasmic variants around the DM position could influence the results. This hypothesis was substantiated by repeating the methylation calling analysis using a sample-specific mitochondrial reference sequence (generated using Illumina sequencing variant calling performed on the same samples), where all the previously identified DM-CpGs disappeared, thus confirming the artefactual nature of the previously-identified DM results. Additionally, the levels of mtDNA methylation quantified in our samples were extremely low and decreasing at increasing read depths (particularly in the human tissues), and consistently lower than the methylation background noise calculated using the model inferred from simulated data and described in **chapter 5** (and the **Methods** section).

As a final confirmation that ONS was indeed capable of identifying mtDNA methylation when present (even at low levels), we sequenced a series of positive controls with expected intermediate methylation levels (0%, 5%, 25%, 50%, 75%, 100%) and high read depths (~2000x - ~5000x, min-max). Our results showed that indeed we could correctly identify the expected methylation levels of our positive controls, confirming on one side that ONS technology is able to correctly detect mtDNA methylation levels as low as 5% and, on the other side, the absence of any substantial CpG methylation on the mtDNA in the biological contexts we analysed.

Chapter 7. Summary and conclusions

While every chapter has been discussed in turn, in this chapter we will provide a discussion on the whole work and its relevance in the field of mtDNA epigenetics.

7.1 Summary

Soon after the discovery of mtDNA, researchers began to be interested in discovering whether this DNA molecule could be epigenetically regulated as well[170]. Although the early reports did not find substantial amounts of methylated residues on mtDNA, the recent discovery of a mitochondrially-targeted DNA methyltransferase[210] (*DNMT1*), rekindled the interest in the field. In the span of a few years many articles were published reporting specific mtDNA methylation patterns in a variety of fields ranging from cancer to neurological research (see **chapter 1**). However, a few studies published in parallel to these new reports focussed their attention on the techniques used to measure mtDNA methylation[240,242], pointing out fundamental flaws in bisulfite technology which laid at the basis of most of the published studies.

In our work, we addressed again this fundamental issue, and in **chapter 4** we analysed 55 human WGBS samples part of the Roadmap Epigenomic Project[247]. We focussed on highlighting the underlying biases intrinsic to WGBS and we could indeed separate the samples in 2 groups based on the amount of alignment bias they showed. More than half of the samples (58.2%) had a strong alignment bias on the H strand (≥ 55% of the mitochondrial reads aligned on the H strand), and uneven coverage between the mtDNA strands. The methylation between the strands was also significantly higher in the L-strand compared to the H-strand. As reported by Mechta and colleagues[240], we too identified a strong inverse relationship between the mtDNA methylation levels and the read depth of the individual cytosines on all mtDNA strands in all the sample groups analysed. This is in line with what Olova and colleagues reported[242], about the strong bias being likely linked with the bisulfite-mediated degradation of the cytosine-rich mtDNA L-strand. This also explains the differences of methylation between the two mtDNA strands: the intrinsic bisulfite-mediated degradation of the L-strand resulted in a lower read depth, which influenced in turn the methylation level on that strand (since methylation is measured as a number of methylated calls over the total calls). In conclusion, in this chapter we demonstrated the full extension of the problems

intrinsic to the gold-standard technology currently used to study methylation, and we then sought for an alternative method to look at mtDNA methylation.

This was the focus of **chapter 5**, where we started exploring the possibility of using ONS to analyse methylation on mtDNA. Firstly, we assessed the accuracy level of this technology for methylation detection. Using *ad-hoc* mitochondrial negative and positive controls (NC and PC) and a modified bioinformatic pipeline, we established that using the Nanopolish software[198] we were able to reach an accuracy of 99% in detecting methylated residues on the mitochondrial molecule. The next step in our work was then to modify the standard ONS library preparation procedure to specifically enrich mitochondrial sequences out of gDNA. To do this, we tested two enrichment methods: one based on digestion of gDNA by Exonuclease V[245], and the other based on digestion by BamHI followed by enrichment of long fragments. The latter proved to be more practically convenient for our purposes, and we then tested this new method against the standard ONS library preparation, based on random fragmentation of gDNA. Our results showed that our new protocol was better than the standard library preparation method in achieving higher mitochondrial read depths with no alignment or coverage biases observed in the samples treated with BamHI. Regarding methylation estimations, once again we observed the same inverse relationship between read depth and methylation levels in all samples sequenced with ONS. We hypothesised that this phenomenon was due to random noise introduced by incorrect methylation calls introduced by Nanopolish (or due to random fluctuations in the raw electrical ONS signal). To assess this and to describe a baseline methylation error level, we randomly sampled from the NC control BAM file sequences in order to form new simulated datasets with read depths corresponding to the observed samples. This then allowed us to calculate a mathematical model of the methylation background noise distribution, based on these simulated samples. When we compared the observed methylation levels to the background, we found that all the signal was below the noise level, even at higher read depths.

The potential of our new protocol was explored in **chapter 6**, where we initially performed variant calling analysis on cybrid cell lines and primary fibroblasts with either a specific haplogroup (H, J or J2) or a mtDNA pathological mutation, respectively. Comparing results obtained with ONS to gold-standard Illumina

sequencing we were able to identify the correct haplogroup in all of the cybrids cell lines and the pathologic mutation in the fibroblasts. On average, we were able to identify 60 mtDNA variants with ≥ 10% heteroplasmy per sample using only ONS. Of these, only ~47% were then confirmed with Illumina Miseq, and they were mostly highly heteroplasmic or homoplasmic variants (≥ 90%). The remaining mtDNA variants identified only with ONS were mostly low heteroplasmic, probably derived from the random insertion/deletion that are common in ONS. In general, heteroplasmies calculated with ONS overall tended to correlate better with Illumina at higher read depths.

Having explored the potential of ONS in calling mitochondrial variants, we set out to analyse methylation in the same sample groups on which variant calling analysis was performed. An initial differential methylation analysis revealed a few possible methylation residues that were differentially methylated between H and J haplogroups cell lines, and between control and mutation primary fibroblasts, respectively. However, we discovered that every single one of these residues had an haplogroup-defining variant in a $\pm$ 5 bp window around the methylated position. Because for calling methylation Nanopolish compares the raw ONS signal with an expected signal that it calculates based on the reference sequence that is provided by the user, we reasoned that the presence of these variants could affect methylation calling. Therefore we repeated the differential methylation analysis using new sample-specific reference sequences. In this case, all the methylated position identified previously disappeared, confirming our theory. All methylation calling was then performed using sample-specific mitochondrial sequences.

Methylation analysis revealed very low levels of average apparent methylation in all the samples analysed (cybrids and fibroblasts methylation = 1.3%-2%; min-max), including in 7 human samples where read depth was higher than the cell lines (methylation tissues = 0.6%-0.8%, min-max). Moreover, all of the measured apparent methylation levels were below the background level calculated at each read depth using the formula we inferred in **chapter 5** using simulated negative controls. As a final proof that ONS was indeed able to measure methylation, we generated new positive controls possessing intermediate methylation levels (5%, 10%, 25%, 50%, 75%) and sequenced them with ONS. Methylation analysis revealed that in this case we were indeed able to observe all the expected average methylation levels in our controls, confirming that methylation is absent in biological samples.

7.2 Conclusions

In conclusion, the work presented in this dissertation is following up a line of research initiated early after the discovery of mtDNA itself. The challenge of demonstrating the presence of mtDNA methylation is hard, but the potential is huge. If mtDNA was indeed methylated a range of possibilities would open up when determining its role in either gene expression (mirroring its role in nDNA), or even in mtDNA replication, that could potentially change mitochondrial biology in a major way. Mitochondrial DNA is not only involved in regulating OXPHOS and in debilitating mitochondrial disease syndromes, but its role in regulating metabolism is emerging[314]. For example, knowing whether the expression of mtDNA proteins is regulated epigenetically could imply knowing whether it is possible to ameliorate the effects of the pathological mtDNA mutations in mitochondrial syndromes. Because of this potential, many studies have over the years tried to identify mtDNA methylation and/or which protein is responsible for its establishment. However, of the many works that were published, no consensus has ever been reached on which patterns of methylation are actually correlated with diseased states (in cancer or other diseases) or with physiological conditions (such as cell senescence or aging). There is no consensus either on which of the three DNMT proteins are the ones responsible for the establishment and replication of mtDNA methylation patters, and no study has ever explored how practically mtDNA methylation could affect either mtDNA gene expression or replication. This uncertainty, coupled with the few thorough studies that pointed out the flaws intrinsic to the technologies used to detect mtDNA methylation have contributed to give the impression that, at best, the field of mitochondria epigenetics is confused or, at worst, that mtDNA methylation does not exist or is not relevant even if it existed. We believe that this confusion is in part justified, and it could be explained by the fact that most of the published studies do not apply the level of thoroughness that should instead be standard in mtDNA research in general. For example, in very few of the published works we observed efforts to reduce NuMTs contamination in the sequenced samples, something that should be fundamental when trying to decipher whether the observed methylation comes from *bona fide* mitochondrial sequences. Also, after the publication of the reports from Mechta and colleagues[240] and Olova and colleagues[242], only a few groups followed their guidelines to reduce sources of biases when using bisulfite sequencing to study mtDNA methylation.

Our work is intended as both a follow up of those works highlighting flaws in the current way mtDNA methylation is studied, and to propose a future direction that has applications even beyond mtDNA methylation analysis. Our analysis of publicly available WGBS samples confirmed the intrinsic problems identified by Olova and colleagues[242], and pointed out how they affect mtDNA methylation calling. In the following chapters we tested the potential of using a cutting edge technology, ONS, to surpass the problems intrinsic to bisulfite-based technologies. We showed that our modified protocol, intended for streamlined analysis of multiple samples, was better than the standard ONS library preparation method, based on random fragmentation, to sequence native full-length mtDNA sequences. We tried to exploit the potential of this multi-modal technology by assessing the accuracy of the variant calling on ONS samples, showing good correlation with Illumina data for high heteroplasmy and high depth variants. When assessing mtDNA methylation we tried to reduce as much as possible any sources of false positives, such as NuMTs contamination, alignment biases or sequence-specific artefacts introduced by Nanopolish software. Using very stringent parameters we called methylation on a series of biological samples with good read depths, but failed to identify any methylation above the background level calculated on *ad-hoc* negative controls. As we could on the other hand identify "artificial" methylation on positive controls, we concluded that methylation on the cytosine is absent in mtDNA.

However, the future of mitochondria epigenetic seems already on the brink of another major breakthrough, for which our method could represent an ideal tool. Very recent works have identified adenine methylation (6mA) as the methylated residue mainly present on mtDNA[315]. In the future it could be interesting to go back to the data created for this study and look for 6mA instead of 5mC, provided a good detection model is generated with Nanopolish using specific controls.

7.3 Future plans

Although in our opinion the work presented in this dissertation has been more thorough than what has been published on the matter of mtDNA methylation, at least for what concerns the estimation of the accuracy of our results, there are a series of line of evidence that would have been interesting to explore. In general, as it was clear from

the results presented, it would be useful to add replicates to the biological samples analysed in this study with the aim to improve on the read depths achieved. This would give us the opportunity to strengthen the conclusions that we had found regarding the absence of mtDNA methylation on one side, and ameliorate the variant calling.

Secondly, we would have liked the possibility to test the presence of mtDNA methylation using additional controls. One of this would have been using cell lines where the mtDNA methylation is artificially induced using a recombinant methyltransferase directed to the mitochondria. We had done preliminary experiments in that sense, where we engineered a few plasmid constructs where we attached the bacterial methyltransferase M.SssI sequence to a mitochondrial target sequence (MTS), and transfected the plasmid into HeLa cells. While the efficacy of the MTS to target mitochondria was demonstrated by swapping M.SssI with GFP, we were never able to see expression of the methyltransferase to mitochondria. These experiments would have required some additional time to fine-tune the technical issues usually involved with experiments of this kind, and would have resulted in the generation of positive controls which would have probably strengthen our conclusions regarding the absence of mtDNA in biological samples.

Regarding this last point, another good line of research that we started was done in collaboration with the laboratory of Dr James Stewart at the university of Newcastle. Following on what we also thought in parallel for cell lines, they were trying to develop a mouse model that could target both a CpG and a GpC methyltransferase to mitochondria. This latter strategy was used in a recent publication, where cytosine methylation on GpC residues was used as a method to assess nucleosome occupancy, using long-read single-molecule ONS sequencing, in addition to "standard" CpG methylation[316]. They developed mouse models at both the homozygous and heterozygous level of expression of their constructs, and we received some organs from those models. It would be interesting to test the validity of their mouse models using our BamHI-based model to assess the methylation both at the CpG and on the GpC level (given a re-training of the Nanopolish model or using other software).

Finally, as mentioned in the previous paragraph, since a characteristic of ONS data is that they can be re-analysed multiple times to look for additional modifications, it would be interesting to perform a complete re-analysis of the biological samples used in this work to determine the presence of 6mA methylation. We had started this line of

research by generating positive and negative controls in a similar way as we did for those used for CpG methylation, but we incurred in technical problems that would require more time to smooth out.

References

(1) Berg, J.; Tymoczko, J.; Stryer, L. *Biochemistry, 5th Edition*; 2002.

(2) Timmis, J. N.; Ayliff, M. A.; Huang, C. Y.; Martin, W. Endosymbiotic Gene Transfer: Organelle Genomes Forge Eukaryotic Chromosomes. *Nature Reviews Genetics*. 2004. https://doi.org/10.1038/nrg1271.

(3) Archibald, J. M. Origin of Eukaryotic Cells: 40 Years On. *Symbiosis* **2011**. https://doi.org/10.1007/s13199-011-0129-z.

(4) Martijn, J.; Vosseberg, J.; Guy, L.; Offre, P.; Ettema, T. J. G. Deep Mitochondrial Origin Outside the Sampled Alphaproteobacteria. *Nature* **2018**. https://doi.org/10.1038/s41586-018-0059-5.

(5) Hazkani-Covo, E.; Zeller, R. M.; Martin, W. Molecular Poltergeists: Mitochondrial DNA Copies (Numts) in Sequenced Nuclear Genomes. *PLoS Genetics*. 2010. https://doi.org/10.1371/journal.pgen.1000834.

(6) Freya, T. G.; Mannellab, C. A. The Internal Structure of Mitochondria. *Trends in Biochemical Sciences*. 2000. https://doi.org/10.1016/S0968-0004(00)01609-1.

(7) Prudent, J.; Zunino, R.; Sugiura, A.; Mattie, S.; Shore, G. C.; McBride, H. M. MAPL SUMOylation of Drp1 Stabilizes an ER/Mitochondrial Platform Required for Cell Death. *Mol. Cell* **2015**. https://doi.org/10.1016/j.molcel.2015.08.001.

(8) Friedman, J. R.; Nunnari, J. Mitochondrial Form and Function. *Nature*. 2014. https://doi.org/10.1038/nature12985.

(9) Daems, W. T.; Wisse, E. Shape and Attachment of the Cristae Mitochondriales in Mouse Hepatic Cell Mitochondria. *J. Ultrasructure Res.* **1966**. https://doi.org/10.1016/S0022-5320(66)80027-8.

(10) Lea, P. J.; Hollenberg, M. J. Mitochondrial Structure Revealed by High-resolution Scanning Electron Microscopy. *Am. J. Anat.* **1989**. https://doi.org/10.1002/aja.1001840308.

(11) Wiedemann, N.; Pfanner, N. Mitochondrial Machineries for Protein Import and Assembly. *Annu. Rev. Biochem.* **2017**. https://doi.org/10.1146/annurev-biochem-060815-014352.

(12) Guo, X. W.; Mannella, C. A. Conformational Change in the Mitochondrial Channel, VDAC, Detected by Electron Cryo-Microscopy. *Biophys. J.* **1993**. https://doi.org/10.1016/S0006-3495(93)81399-7.

(13) Pernas, L.; Scorrano, L. Mito-Morphosis: Mitochondrial Fusion, Fission, and Cristae Remodeling as Key Mediators of Cellular Function. *Annual Review of Physiology*. 2016. https://doi.org/10.1146/annurev-physiol-021115-105011.

(14) Li, P.; Nijhawan, D.; Budihardjo, I.; Srinivasula, S. M.; Ahmad, M.; Alnemri, E. S.; Wang, X. Cytochrome c and DATP-Dependent Formation of Apaf-1/Caspase-9 Complex Initiates an Apoptotic Protease Cascade. *Cell* **1997**. https://doi.org/10.1016/S0092-8674(00)80434-1.

(15) Susin, S. A.; Lorenzo, H. K.; Zamzami, N.; Marzo, I.; Snow, B. E.; Brothers, G. M.; Mangion, J.; Jacotot, E.; Costantini, P.; Loeffler, M.; Larochette, N.; Goodlett, D. R.; Aebersold, R.; Siderovski, D. P.; Penninger, J. M.; Kroemer, G. Molecular Characterization of Mitochodrial Apoptosis-Inducing Factor. *Nature* **1999**. https://doi.org/10.1038/17135.

(16) Baughman, J. M.; Perocchi, F.; Girgis, H. S.; Plovanich, M.; Belcher-Timme, C. A.; Sancak, Y.; Bao, X. R.; Strittmatter, L.; Goldberger, O.; Bogorad, R. L.; Koteliansky, V.; Mootha, V. K. Integrative Genomics Identifies MCU as an Essential Component of the Mitochondrial Calcium Uniporter. *Nature* **2011**. https://doi.org/10.1038/nature10234.

(17) DELUCA, H. F.; ENGSTROM, G. W. Calcium Uptake by Rat Kidney Mitochondria. *Proc. Natl. Acad. Sci. U. S. A.* **1961**. https://doi.org/10.1073/pnas.47.11.1744.
(18) Paupe, V.; Prudent, J. New Insights into the Role of Mitochondrial Calcium Homeostasis in Cell Migration. *Biochem. Biophys. Res. Commun.* **2018**. https://doi.org/10.1016/j.bbrc.2017.05.039.
(19) Perocchi, F.; Gohil, V. M.; Girgis, H. S.; Bao, X. R.; McCombs, J. E.; Palmer, A. E.; Mootha, V. K. MICU1 Encodes a Mitochondrial EF Hand Protein Required for Ca2+ Uptake. *Nature* **2010**. https://doi.org/10.1038/nature09358.
(20) Matheoud, D.; Sugiura, A.; Bellemare-Pelletier, A.; Laplante, A.; Rondeau, C.; Chemali, M.; Fazel, A.; Bergeron, J. J.; Trudeau, L. E.; Burelle, Y.; Gagnon, E.; McBride, H. M.; Desjardins, M. Parkinson's Disease-Related Proteins PINK1 and Parkin Repress Mitochondrial Antigen Presentation. *Cell* **2016**. https://doi.org/10.1016/j.cell.2016.05.039.
(21) Weinberg, S. E.; Sena, L. A.; Chandel, N. S. Mitochondria in the Regulation of Innate and Adaptive Immunity. *Immunity*. 2015. https://doi.org/10.1016/j.immuni.2015.02.002.
(22) West, A. P.; Shadel, G. S.; Ghosh, S. Mitochondria in Innate Immune Responses. *Nature Reviews Immunology*. 2011. https://doi.org/10.1038/nri2975.
(23) Reyes, J. L.; Aldana, I.; Barbier, O.; Parrales, A. A.; Melendez, E. Indomethacin Decreases Furosemide-Induced Natriuresis and Diuresis on the Neonatal Kidney. *Pediatr. Nephrol.* **2006**. https://doi.org/10.1007/s00467-006-0224-1.
(24) Inaba, K.; Oda, T. Phosphorylation of Purine and Pyrimidine Nucleosides by Isolated Rat Liver Mitochondria. *Acta Med. Okayama* **1975**.
(25) Adeva-Andany, M. M.; Carneiro-Freire, N.; Seco-Filgueira, M.; Fernández-Fernández, C.; Mouriño-Bayolo, D. Mitochondrial β-Oxidation of Saturated Fatty Acids in Humans. *Mitochondrion*. 2019. https://doi.org/10.1016/j.mito.2018.02.009.
(26) Murphy, M. P. How Mitochondria Produce Reactive Oxygen Species. *Biochemical Journal*. 2009. https://doi.org/10.1042/BJ20081386.
(27) Kadenbach, B. Introduction to Mitochondrial Oxidative Phosphorylation. *Adv. Exp. Med. Biol.* **2012**. https://doi.org/10.1007/978-1-4614-3573-0_1.
(28) Van Den Heuvel, L.; Smeitink, J. The Oxidative Phosphorylation (OXPHOS) System: Nuclear Genes and Human Genetic Diseases. *BioEssays*. 2001. https://doi.org/10.1002/bies.1071.
(29) Zhu, J.; Vinothkumar, K. R.; Hirst, J. Structure of Mammalian Respiratory Complex I. *Nature* **2016**. https://doi.org/10.1038/nature19095.
(30) Sun, F.; Huo, X.; Zhai, Y.; Wang, A.; Xu, J.; Su, D.; Bartlam, M.; Rao, Z. Crystal Structure of Mitochondrial Respiratory Membrane Protein Complex II. *Cell* **2005**. https://doi.org/10.1016/j.cell.2005.05.025.
(31) Iwata, S.; Lee, J. W.; Okada, K.; Lee, J. K.; Iwata, M.; Rasmussen, B.; Link, T. A.; Ramaswamy, S.; Jap, B. K. Complete Structure of the 11-Subunit Bovine Mitochondrial Cytochrome Bc1 Complex. *Science (80-.).* **1998**. https://doi.org/10.1126/science.281.5373.64.
(32) Tsukihara, T.; Aoyama, H.; Yamashita, E.; Tomizaki, T.; Yamaguchi, H.; Shinzawa-Itoh, K.; Nakashima, R.; Yaono, R.; Yoshikawa, S. Structures of Metal Sites of Oxidized Bovine Heart Cytochrome c Oxidase at 2.8 Å. *Science (80-.).* **1995**. https://doi.org/10.1126/science.7652554.
(33) Walker, J. E.; Dickson, V. K. The Peripheral Stalk of the Mitochondrial ATP

Synthase. *Biochimica et Biophysica Acta - Bioenergetics*. 2006. https://doi.org/10.1016/j.bbabio.2006.01.001.
(34) Martínez-Reyes, I.; Chandel, N. S. Mitochondrial TCA Cycle Metabolites Control Physiology and Disease. *Nature Communications*. 2020. https://doi.org/10.1038/s41467-019-13668-3.
(35) Kim, H. J.; Khalimonchuk, O.; Smith, P. M.; Winge, D. R. Structure, Function, and Assembly of Heme Centers in Mitochondrial Respiratory Complexes. *Biochimica et Biophysica Acta - Molecular Cell Research*. 2012. https://doi.org/10.1016/j.bbamcr.2012.04.008.
(36) Mitchell, P.; Moyle, J. Respiration-Driven Proton Translocation in Rat Liver Mitochondria. *Biochem. J.* **1967**. https://doi.org/10.1042/bj1051147.
(37) Shadel, G. S.; Horvath, T. L. Mitochondrial ROS Signaling in Organismal Homeostasis. *Cell*. 2015. https://doi.org/10.1016/j.cell.2015.10.001.
(38) Wallace, D. C. Mitochondrial DNA Mutations in Disease and Aging. *Environmental and Molecular Mutagenesis*. 2010. https://doi.org/10.1002/em.20586.
(39) He, J.; Ford, H. C.; Carroll, J.; Douglas, C.; Gonzales, E.; Ding, S.; Fearnley, I. M.; Walker, J. E. Assembly of the Membrane Domain of ATP Synthase in Human Mitochondria. *Proc. Natl. Acad. Sci. U. S. A.* **2018**. https://doi.org/10.1073/pnas.1722086115.
(40) Kitazaki, K.; Kubo, T. Cost of Having the Largest Mitochondrial Genome: Evolutionary Mechanism of Plant Mitochondrial Genome. *J. Bot.* **2010**, *2010*. https://doi.org/10.1155/2010/620137.
(41) Boore, J. L. Animal Mitochondrial Genomes. *Nucleic Acids Res.* **1999**, *27* (8). https://doi.org/10.1093/nar/27.8.1767.
(42) Berk, A. J.; Clayton, D. A. Mechanism of Mitochondrial DNA Replication in Mouse L-Cells: Asynchronous Replication of Strands, Segregation of Circular Daughter Molecules, Aspects of Topology and Turnover of an Initiation Sequence. *J. Mol. Biol.* **1974**. https://doi.org/10.1016/0022-2836(74)90355-6.
(43) Chinnery, P. F.; Hudson, G. Mitochondrial Genetics. *British Medical Bulletin*. 2013. https://doi.org/10.1093/bmb/ldt017.
(44) Stewart, J. B.; Chinnery, P. F. The Dynamics of Mitochondrial DNA Heteroplasmy: Implications for Human Health and Disease. *Nature Reviews Genetics*. 2015. https://doi.org/10.1038/nrg3966.
(45) Nicholls, T. J.; Minczuk, M. In D-Loop: 40 Years of Mitochondrial 7S DNA. *Exp. Gerontol.* **2014**. https://doi.org/10.1016/j.exger.2014.03.027.
(46) Temperley, R.; Richter, R.; Dennerlein, S.; Lightowlers, R. N.; Chrzanowska-Lightowlers, Z. M. Hungry Codons Promote Frameshifting in Human Mitochondrial Ribosomes. *Science*. 2010. https://doi.org/10.1126/science.1180674.
(47) Sutovsky, P. Ubiquitin-Dependent Proteolysis in Mammalian Spermatogenesis, Fertilization, and Sperm Quality Control: Killing Three Birds with One Stone. *Microsc. Res. Tech.* **2003**. https://doi.org/10.1002/jemt.10319.
(48) Speranzini, V.; Pilotto, S.; Sixma, T. K.; Mattevi, A. Touch, Act and Go: Landing and Operating on Nucleosomes. *EMBO J.* **2016**. https://doi.org/10.15252/embj.201593377.
(49) Brown, T. A.; Tkachuk, A. N.; Shtengel, G.; Kopek, B. G.; Bogenhagen, D. F.; Hess, H. F.; Clayton, D. A. Superresolution Fluorescence Imaging of Mitochondrial Nucleoids Reveals Their Spatial Range, Limits, and Membrane Interaction. *Mol. Cell. Biol.* **2011**. https://doi.org/10.1128/mcb.05694-11.

(50) Hayashi, J. I.; Ohta, S.; Kikuchi, A.; Takemitsu, M.; Goto, Y. I.; Nonaka, I. Introduction of Disease-Related Mitochondrial DNA Deletions into HeLa Cells Lacking Mitochondrial DNA Results in Mitochondrial Dysfunction. *Proc. Natl. Acad. Sci. U. S. A.* **1991**. https://doi.org/10.1073/pnas.88.23.10614.
(51) Garrido, N.; Griparic, L.; Jokitalo, E.; Wartiovaara, J.; Van der Bliek, A. M.; Spelbrink, J. N. Composition and Dynamics of Human Mitochondrial Nucleoids. *Mol. Biol. Cell* **2003**. https://doi.org/10.1091/mbc.E02-07-0399.
(52) Kukat, C.; Davies, K. M.; Wurm, C. A.; Spåhr, H.; Bonekamp, N. A.; Kühl, I.; Joos, F.; Polosa, P. L.; Park, C. B.; Posse, V.; Falkenberg, M.; Jakobs, S.; Kühlbrandt, W.; Larsson, N. G. Cross-Strand Binding of TFAM to a Single MtDNA Molecule Forms the Mitochondrial Nucleoid. *Proc. Natl. Acad. Sci. U. S. A.* **2015**. https://doi.org/10.1073/pnas.1512131112.
(53) Kukat, C.; Wurm, C. A.; Spåhr, H.; Falkenberg, M.; Larsson, N. G.; Jakobs, S. Super-Resolution Microscopy Reveals That Mammalian Mitochondrial Nucleoids Have a Uniform Size and Frequently Contain a Single Copy of MtDNA. *Proc. Natl. Acad. Sci. U. S. A.* **2011**. https://doi.org/10.1073/pnas.1109263108.
(54) Kaufman, B. A.; Durisic, N.; Mativetsky, J. M.; Costantino, S.; Hancock, M. A.; Grutter, P.; Shoubridge, E. A. The Mitochondrial Transcription Factor TFAM Coordinates the Assembly of Multiple DNA Molecules into Nucleoid-like Structures. *Mol. Biol. Cell* **2007**. https://doi.org/10.1091/mbc.E07-05-0404.
(55) Bogenhagen, D. F.; Rousseau, D.; Burke, S. The Layered Structure of Human Mitochondrial DNA Nucleoids. *J. Biol. Chem.* **2008**. https://doi.org/10.1074/jbc.M708444200.
(56) Miller, F. J.; Rosenfeldt, F. L.; Zhang, C.; Linnane, A. W.; Nagley, P. Precise Determination of Mitochondrial DNA Copy Number in Human Skeletal and Cardiac Muscle by a PCR-Based Assay: Lack of Change of Copy Number with Age. *Nucleic Acids Res.* **2003**. https://doi.org/10.1093/nar/gng060.
(57) D'Erchia, A. M.; Atlante, A.; Gadaleta, G.; Pavesi, G.; Chiara, M.; De Virgilio, C.; Manzari, C.; Mastropasqua, F.; Prazzoli, G. M.; Picardi, E.; Gissi, C.; Horner, D.; Reyes, A.; Sbisà, E.; Tullo, A.; Pesole, G. Tissue-Specific MtDNA Abundance from Exome Data and Its Correlation with Mitochondrial Transcription, Mass and Respiratory Activity. *Mitochondrion* **2015**. https://doi.org/10.1016/j.mito.2014.10.005.
(58) Kelly, R. D. W.; Mahmud, A.; McKenzie, M.; Trounce, I. A.; St John, J. C. Mitochondrial DNA Copy Number Is Regulated in a Tissue Specific Manner by DNA Methylation of the Nuclear-Encoded DNA Polymerase Gamma A. *Nucleic Acids Res.* **2012**. https://doi.org/10.1093/nar/gks770.
(59) Tyynismaa, H.; Sembongi, H.; Bokori-Brown, M.; Granycome, C.; Ashley, N.; Poulton, J.; Jalanko, A.; Spelbrink, J. N.; Holt, I. J.; Suomalainen, A. Twinkle Helicase Is Essential for MtDNA Maintenance and Regulates MtDNA Copy Number. *Hum. Mol. Genet.* **2004**. https://doi.org/10.1093/hmg/ddh342.
(60) Van Dyck, E.; Foury, F.; Stillman, B.; Brill, S. J. A Single-Stranded DNA Binding Protein Required for Mitochondrial DNA Replication in S. Cerevisiae Is Homologous to E. Coli SSB. *EMBO J.* **1992**. https://doi.org/10.1002/j.1460-2075.1992.tb05421.x.
(61) Stewart, J. D.; Schoeler, S.; Sitarz, K. S.; Horvath, R.; Hallmann, K.; Pyle, A.; Yu-Wai-Man, P.; Taylor, R. W.; Samuels, D. C.; Kunz, W. S.; Chinnery, P. F. POLG Mutations Cause Decreased Mitochondrial DNA Repopulation Rates Following Induced Depletion in Human Fibroblasts. *Biochim. Biophys. Acta* -

Mol. Basis Dis. **2011**. https://doi.org/10.1016/j.bbadis.2010.11.012.
(62) Larsson, N. G.; Wang, J.; Wilhelmsson, H.; Oldfors, A.; Rustin, P.; Lewandoski, M.; Barsh, G. S.; Clayton, D. A. Mitochondrial Transcription Factor A Is Necessary for MtDNA Maintenance and Embryogenesis in Mice. *Nat. Genet.* **1998**. https://doi.org/10.1038/ng0398-231.
(63) Ikeda, M.; Ide, T.; Fujino, T.; Arai, S.; Saku, K.; Kakino, T.; Tyynismaa, H.; Yamasaki, T.; Yamada, K. I.; Kang, D.; Suomalainen, A.; Sunagawa, K. Overexpression of TFAM or Twinkle Increases MtDNA Copy Number and Facilitates Cardioprotection Associated with Limited Mitochondrial Oxidative Stress. *PLoS One* **2015**. https://doi.org/10.1371/journal.pone.0119687.
(64) Saada, A.; Bar-Meir, M.; Belaiche, C.; Miller, C.; Elpeleg, O. Evaluation of Enzymatic Assays and Compounds Affecting ATP Production in Mitochondrial Respiratory Chain Complex I Deficiency. *Anal. Biochem.* **2004**. https://doi.org/10.1016/j.ab.2004.08.015.
(65) Jones, B. A.; Fangman, W. L. Mitochondrial DNA Maintenance in Yeast Requires a Protein Containing a Region Related to the GTP-Binding Domain of Dynamin. *Genes Dev.* **1992**. https://doi.org/10.1101/gad.6.3.380.
(66) Reyes, A.; He, J.; Mao, C. C.; Bailey, L. J.; Di Re, M.; Sembongi, H.; Kazak, L.; Dzionek, K.; Holmes, J. B.; Cluett, T. J.; Harbour, M. E.; Fearnley, I. M.; Crouch, R. J.; Conti, M. A.; Adelstein, R. S.; Walker, J. E.; Holt, I. J. Actin and Myosin Contribute to Mammalian Mitochondrial DNA Maintenance. *Nucleic Acids Res.* **2011**. https://doi.org/10.1093/nar/gkr052.
(67) Srivastava, S.; Diaz, F.; Iommarini, L.; Aure, K.; Lombes, A.; Moraes, C. T. PGC-1α/β Induced Expression Partially Compensates for Respiratory Chain Defects in Cells from Patients with Mitochondrial Disorders. *Hum. Mol. Genet.* **2009**. https://doi.org/10.1093/hmg/ddp093.
(68) Ciesielski, G. L.; Plotka, M.; Manicki, M.; Schilke, B. A.; Dutkiewicz, R.; Sahi, C.; Marszalek, J.; Craig, E. A. Nucleoid Localization of Hsp40 Mdj1 Is Important for Its Function in Maintenance of Mitochondrial DNA. *Biochim. Biophys. Acta - Mol. Cell Res.* **2013**. https://doi.org/10.1016/j.bbamcr.2013.05.012.
(69) Matsushima, Y.; Goto, Y. I.; Kaguni, L. S. Mitochondrial Lon Protease Regulates Mitochondrial DNA Copy Number and Transcription by Selective Degradation of Mitochondrial Transcription Factor A (TFAM). *Proc. Natl. Acad. Sci. U. S. A.* **2010**. https://doi.org/10.1073/pnas.1008924107.
(70) Sesaki, H.; Southard, S. M.; Aiken Hobbs, A. E.; Jensen, R. E. Cells Lacking Pcp1p/Ugo2p, a Rhomboid-like Protease Required for Mgm1p Processing, Lose MtDNA and Mitochondrial Structure in a Dnm1p-Dependent Manner, but Remain Competent for Mitochondrial Fusion. *Biochem. Biophys. Res. Commun.* **2003**. https://doi.org/10.1016/S0006-291X(03)01348-2.
(71) Kraytsberg, Y.; Schwartz, M.; Brown, T. A.; Ebralidse, K.; Kunz, W. S.; Clayton, D. A.; Vissing, J.; Khrapko, K. Recombination of Human Mitochondrial DNA. *Science (80-.).* **2004**, *304* (5673). https://doi.org/10.1126/science.1096342.
(72) Torroni, A.; Achilli, A.; Macaulay, V.; Richards, M.; Bandelt, H. J. Harvesting the Fruit of the Human MtDNA Tree. *Trends in Genetics*. 2006. https://doi.org/10.1016/j.tig.2006.04.001.
(73) van Oven, M.; Kayser, M. Updated Comprehensive Phylogenetic Tree of Global Human Mitochondrial DNA Variation. *Hum. Mutat.* **2009**. https://doi.org/10.1002/humu.20921.
(74) Mishmar, D.; Ruiz-Pesini, E.; Golik, P.; Macaulay, V.; Clark, A. G.; Hosseini, S.; Brandon, M.; Easleyf, K.; Chen, E.; Brown, M. D.; Sukernik, R. I.; Olckers, A.;

Wallace, D. C. Natural Selection Shaped Regional MtDNA Variation in Humans. *Proc. Natl. Acad. Sci. U. S. A.* **2003**. https://doi.org/10.1073/pnas.0136972100.
(75) Macaulay, V.; Hill, C.; Achilli, A.; Rengo, C.; Clarke, D.; Meehan, W.; Blackburn, J.; Semino, O.; Scozzari, R.; Cruciani, F.; Taha, A.; Shaari, N. K.; Raja, J. M.; Ismail, P.; Zainuddin, Z.; Goodwin, W.; Bulbeck, D.; Bandelt, H. J.; Oppenheimer, S.; Torroni, A.; Richards, M. Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes. *Science (80-.).* **2005**. https://doi.org/10.1126/science.1109792.
(76) Behar, D. M.; Van Oven, M.; Rosset, S.; Metspalu, M.; Loogväli, E. L.; Silva, N. M.; Kivisild, T.; Torroni, A.; Villems, R. A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from Its Root. *Am. J. Hum. Genet.* **2012**. https://doi.org/10.1016/j.ajhg.2012.03.002.
(77) Wallace, D. C. A Mitochondrial Paradigm of Metabolic and Degenerative Diseases, Aging, and Cancer: A Dawn for Evolutionary Medicine. *Annual Review of Genetics*. 2005. https://doi.org/10.1146/annurev.genet.39.110304.095751.
(78) Chinnery, P. F.; Elliott, H. R.; Hudson, G.; Samuels, D. C.; Relton, C. L. Epigenetics, Epidemiology and Mitochondrial DNA Diseases. *Int. J. Epidemiol.* **2012**. https://doi.org/10.1093/ije/dyr232.
(79) Ghezzi, D.; Zeviani, M. Assembly Factors of Human Mitochondrial Respiratory Chain Complexes: Physiology and Pathophysiology. *Adv. Exp. Med. Biol.* **2012**. https://doi.org/10.1007/978-1-4614-3573-0_4.
(80) Iommarini, L.; Calvaruso, M. A.; Kurelac, I.; Gasparre, G.; Porcelli, A. M. Complex i Impairment in Mitochondrial Diseases and Cancer: Parallel Roads Leading to Different Outcomes. *Int. J. Biochem. Cell Biol.* **2013**. https://doi.org/10.1016/j.biocel.2012.05.016.
(81) Rossignol, R.; Faustin, B.; Rocher, C.; Malgat, M.; Mazat, J. P.; Letellier, T. Mitochondrial Threshold Effects. *Biochemical Journal*. 2003. https://doi.org/10.1042/BJ20021594.
(82) Gorman, G. S.; Chinnery, P. F.; DiMauro, S.; Hirano, M.; Koga, Y.; McFarland, R.; Suomalainen, A.; Thorburn, D. R.; Zeviani, M.; Turnbull, D. M. Mitochondrial Diseases. *Nat. Rev. Dis. Prim.* **2016**. https://doi.org/10.1038/nrdp.2016.80.
(83) Nunnari, J.; Suomalainen, A. Mitochondria: In Sickness and in Health. *Cell*. 2012. https://doi.org/10.1016/j.cell.2012.02.035.
(84) Bernstein, B. E.; Meissner, A.; Lander, E. S. The Mammalian Epigenome. *Cell*. 2007. https://doi.org/10.1016/j.cell.2007.01.033.
(85) Greenberg, M. V. C.; Bourc'his, D. The Diverse Roles of DNA Methylation in Mammalian Development and Disease. *Nature Reviews Molecular Cell Biology*. 2019. https://doi.org/10.1038/s41580-019-0159-6.
(86) Venkatesh, S.; Workman, J. L. Histone Exchange, Chromatin Structure and the Regulation of Transcription. *Nature Reviews Molecular Cell Biology*. 2015. https://doi.org/10.1038/nrm3941.
(87) Bird, A. DNA Methylation Patterns and Epigenetic Memory. *Genes and Development*. 2002. https://doi.org/10.1101/gad.947102.
(88) Hemberger, M.; Dean, W.; Reik, W. Epigenetic Dynamics of Stem Cells and Cell Lineage Commitment: Digging Waddington's Canal. *Nature Reviews Molecular Cell Biology*. 2009. https://doi.org/10.1038/nrm2727.
(89) Shipony, Z.; Mukamel, Z.; Cohen, N. M.; Landan, G.; Chomsky, E.; Zeliger, S. R.; Fried, Y. C.; Ainbinder, E.; Friedman, N.; Tanay, A. Dynamic and Static Maintenance of Epigenetic Memory in Pluripotent and Somatic Cells. *Nature*

2014. https://doi.org/10.1038/nature13458.
(90) Ramsahoye, B. H.; Biniszkiewicz, D.; Lyko, F.; Clark, V.; Bird, A. P.; Jaenisch, R. Non-CpG Methylation Is Prevalent in Embryonic Stem Cells and May Be Mediated by DNA Methyltransferase 3a. *Proc. Natl. Acad. Sci. U. S. A.* **2000**. https://doi.org/10.1073/pnas.97.10.5237.
(91) Ziller, M. J.; Müller, F.; Liao, J.; Zhang, Y.; Gu, H.; Bock, C.; Boyle, P.; Epstein, C. B.; Bernstein, B. E.; Lengauer, T.; Gnirke, A.; Meissner, A. Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types. *PLoS Genet.* **2011**. https://doi.org/10.1371/journal.pgen.1002389.
(92) Wu, T. P.; Wang, T.; Seetin, M. G.; Lai, Y.; Zhu, S.; Lin, K.; Liu, Y.; Byrum, S. D.; Mackintosh, S. G.; Zhong, M.; Tackett, A.; Wang, G.; Hon, L. S.; Fang, G.; Swenberg, J. A.; Xiao, A. Z. DNA Methylation on N6-Adenine in Mammalian Embryonic Stem Cells. *Nature* **2016**. https://doi.org/10.1038/nature17640.
(93) Douvlataniotis, K.; Bensberg, M.; Lentini, A.; Gylemo, B.; Nestor, C. E. No Evidence for DNA N6-Methyladenine in Mammals. *Sci. Adv.* **2020**. https://doi.org/10.1126/sciadv.aay3335.
(94) Lister, R.; Pelizzola, M.; Dowen, R. H.; Hawkins, R. D.; Hon, G.; Tonti-Filippini, J.; Nery, J. R.; Lee, L.; Ye, Z.; Ngo, Q. M.; Edsall, L.; Antosiewicz-Bourget, J.; Stewart, R.; Ruotti, V.; Millar, A. H.; Thomson, J. A.; Ren, B.; Ecker, J. R. Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences. *Nature* **2009**. https://doi.org/10.1038/nature08514.
(95) Deaton, A. M.; Bird, A. CpG Islands and the Regulation of Transcription. *Genes Dev.* **2011**. https://doi.org/10.1101/gad.2037511.
(96) Saxonov, S.; Berg, P.; Brutlag, D. L. A Genome-Wide Analysis of CpG Dinucleotides in the Human Genome Distinguishes Two Distinct Classes of Promoters. *Proc. Natl. Acad. Sci. U. S. A.* **2006**. https://doi.org/10.1073/pnas.0510310103.
(97) Smith, Z. D.; Meissner, A. DNA Methylation: Roles in Mammalian Development. *Nat. Rev. Genet.* **2013**, *14* (3), 204–220. https://doi.org/10.1038/nrg3354.
(98) Edwards, J. R.; Yarychkivska, O.; Boulard, M.; Bestor, T. H. DNA Methylation and DNA Methyltransferases. *Epigenetics and Chromatin*. 2017. https://doi.org/10.1186/s13072-017-0130-8.
(99) Ooi, S. K. T.; Qiu, C.; Bernstein, E.; Li, K.; Jia, D.; Yang, Z.; Erdjument-Bromage, H.; Tempst, P.; Lin, S. P.; Allis, C. D.; Cheng, X.; Bestor, T. H. DNMT3L Connects Unmethylated Lysine 4 of Histone H3 to de Novo Methylation of DNA. *Nature* **2007**. https://doi.org/10.1038/nature05987.
(100) Jia, D.; Jurkowska, R. Z.; Zhang, X.; Jeltsch, A.; Cheng, X. Structure of Dnmt3a Bound to Dnmt3L Suggests a Model for de Novo DNA Methylation. *Nature* **2007**. https://doi.org/10.1038/nature06146.
(101) Epsztejn-Litman, S.; Feldman, N.; Abu-Remaileh, M.; Shufaro, Y.; Gerson, A.; Ueda, J.; Deplus, R.; Fuks, F.; Shinkai, Y.; Cedar, H.; Bergman, Y. De Novo DNA Methylation Promoted by G9a Prevents Reprogramming of Embryonically Silenced Genes. *Nat. Struct. Mol. Biol.* **2008**. https://doi.org/10.1038/nsmb.1476.
(102) Probst, A. V.; Dunleavy, E.; Almouzni, G. Epigenetic Inheritance during the Cell Cycle. *Nature Reviews Molecular Cell Biology*. 2009. https://doi.org/10.1038/nrm2640.
(103) Hermann, A.; Goyal, R.; Jeltsch, A. The Dnmt1 DNA-(Cytosine-C5)-Methyltransferase Methylates DNA Processively with High Preference for Hemimethylated Target Sites. *J. Biol. Chem.* **2004**.

https://doi.org/10.1074/jbc.M403427200.
(104) Arita, K.; Ariyoshi, M.; Tochio, H.; Nakamura, Y.; Shirakawa, M. Recognition of Hemi-Methylated DNA by the SRA Protein UHRF1 by a Base-Flipping Mechanism. *Nature* **2008**. https://doi.org/10.1038/nature07249.
(105) Wang, J.; Hevi, S.; Kurash, J. K.; Lei, H.; Gay, F.; Bajko, J.; Su, H.; Sun, W.; Chang, H.; Xu, G.; Gaudet, F.; Li, E.; Chen, T. The Lysine Demethylase LSD1 (KDM1) Is Required for Maintenance of Global DNA Methylation. *Nat. Genet.* **2009**. https://doi.org/10.1038/ng.268.
(106) Rothbart, S. B.; Krajewski, K.; Nady, N.; Tempel, W.; Xue, S.; Badeaux, A. I.; Barsyte-Lovejoy, D.; Martinez, J. Y.; Bedford, M. T.; Fuchs, S. M.; Arrowsmith, C. H.; Strahl, B. D. Association of UHRF1 with Methylated H3K9 Directs the Maintenance of DNA Methylation. *Nat. Struct. Mol. Biol.* **2012**. https://doi.org/10.1038/nsmb.2391.
(107) Kohli, R. M.; Zhang, Y. TET Enzymes, TDG and the Dynamics of DNA Demethylation. *Nature*. 2013. https://doi.org/10.1038/nature12750.
(108) Ito, S.; Dalessio, A. C.; Taranova, O. V.; Hong, K.; Sowers, L. C.; Zhang, Y. Role of Tet Proteins in 5mC to 5hmC Conversion, ES-Cell Self-Renewal and Inner Cell Mass Specification. *Nature* **2010**. https://doi.org/10.1038/nature09303.
(109) Ito, S.; Shen, L.; Dai, Q.; Wu, S. C.; Collins, L. B.; Swenberg, J. A.; He, C.; Zhang, Y. Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science (80-.).* **2011**. https://doi.org/10.1126/science.1210597.
(110) He, Y. F.; Li, B. Z.; Li, Z.; Liu, P.; Wang, Y.; Tang, Q.; Ding, J.; Jia, Y.; Chen, Z.; Li, N.; Sun, Y.; Li, X.; Dai, Q.; Song, C. X.; Zhang, K.; He, C.; Xu, G. L. Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science (80-.).* **2011**. https://doi.org/10.1126/science.1210944.
(111) Maiti, A.; Drohat, A. C. Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine: Potential Implications for Active Demethylation of CpG Sites. *J. Biol. Chem.* **2011**. https://doi.org/10.1074/jbc.C111.284620.
(112) Michaud, E. J.; van Vugt, M. J.; Bultman, S. J.; Sweet, H. O.; Davisson, M. T.; Woychik, R. P. Differential Expression of a New Dominant Agouti Allele (A(Iapy)) Is Correlated with Methylation State and Is Influenced by Parental Lineage. *Genes Dev.* **1994**. https://doi.org/10.1101/gad.8.12.1463.
(113) Wu, M.; Rinchik, E. M.; Wilkinson, E.; Johnson, D. K. Inherited Somatic Mosaicism Caused by an Intracisternal A Particle Insertion in the Mouse Tyrosinase Gene. *Proc. Natl. Acad. Sci. U. S. A.* **1997**. https://doi.org/10.1073/pnas.94.3.890.
(114) Kuster, J. E.; Guarnieri, M. H.; Ault, J. G.; Flaherty, L.; Swiatek, P. J. IAP Insertion in the Murine LamB3 Gene Results in Junctional Epidermolysis Bullosa. *Mamm. Genome* **1997**. https://doi.org/10.1007/s003359900535.
(115) Gwynn, B.; Lueders, K.; Sands, M. S.; Birkenmeier, E. H. Intracisternal A-Particle Element Transposition into the Murine β-Glucuronidase Gene Correlates with Loss of Enzyme Activity: A New Model for β-Glucuronidase Deficiency in the C3H Mouse. *Mol. Cell. Biol.* **1998**. https://doi.org/10.1128/mcb.18.11.6474.
(116) Ukai, H.; Ishii-Oba, H.; Ukai-Tadenuma, M.; Ogiu, T.; Tsuji, H. Formation of an Active Form of the Interleukin-2/ 15 Receptor β-Chain by Insertion of the Intracisternal A Particle in a Radiation-Induced Mouse Thymic Lymphoma and

Its Role in Tumorigenesis. *Mol. Carcinog.* **2003**. https://doi.org/10.1002/mc.10128.
(117) Schulz, W. A.; Steinhoff, C.; Florl, A. R. Methylation of Endogenous Human Retroelements in Health and Disease. *Current Topics in Microbiology and Immunology*. 2006. https://doi.org/10.1007/3-540-31181-5_11.
(118) Walsh, C. P.; Chaillet, J. R.; Bestor, T. H. Transcription of IAP Endogenous Retroviruses Is Constrained by Cytosine Methylation [4]. *Nature Genetics*. 1998. https://doi.org/10.1038/2413.
(119) Gaudet, F.; Rideout, W. M.; Meissner, A.; Dausman, J.; Leonhardt, H.; Jaenisch, R. Dnmt1 Expression in Pre- and Postimplantation Embryogenesis and the Maintenance of IAP Silencing. *Mol. Cell. Biol.* **2004**. https://doi.org/10.1128/mcb.24.4.1640-1648.2004.
(120) Hutnick, L. K.; Golshani, P.; Namihira, M.; Xue, Z.; Matynia, A.; Yang, X. W.; Silva, A. J.; Schweizer, F. E.; Fan, G. DNA Hypomethylation Restricted to the Murine Forebrain Induces Cortical Degeneration and Impairs Postnatal Neuronal Maturation. *Hum. Mol. Genet.* **2009**. https://doi.org/10.1093/hmg/ddp222.
(121) Gardiner-Garden, M.; Frommer, M. CpG Islands in Vertebrate Genomes. *J. Mol. Biol.* **1987**. https://doi.org/10.1016/0022-2836(87)90689-9.
(122) Illingworth, R. S.; Gruenewald-Schneider, U.; Webb, S.; Kerr, A. R. W.; James, K. D.; Turner, D. J.; Smith, C.; Harrison, D. J.; Andrews, R.; Bird, A. P. Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome. *PLoS Genet.* **2010**. https://doi.org/10.1371/journal.pgen.1001134.
(123) Tazi, J.; Bird, A. Alternative Chromatin Structure at CpG Islands. *Cell* **1990**. https://doi.org/10.1016/0092-8674(90)90339-G.
(124) Ramirez-Carrozzi, V. R.; Braas, D.; Bhatt, D. M.; Cheng, C. S.; Hong, C.; Doty, K. R.; Black, J. C.; Hoffmann, A.; Carey, M.; Smale, S. T. A Unifying Model for the Selective Regulation of Inducible Transcription by CpG Islands and Nucleosome Remodeling. *Cell* **2009**. https://doi.org/10.1016/j.cell.2009.04.020.
(125) Choi, J. D.; Underkoffler, L. A.; Wood, A. J.; Collins, J. N.; Williams, P. T.; Golden, J. A.; Schuster, E. F.; Loomes, K. M.; Oakey, R. J. A Novel Variant of Inpp5f Is Imprinted in Brain, and Its Expression Is Correlated with Differential Methylation of an Internal CpG Island . *Mol. Cell. Biol.* **2005**. https://doi.org/10.1128/mcb.25.13.5514-5522.2005.
(126) Mikkelsen, T. S.; Ku, M.; Jaffe, D. B.; Issac, B.; Lieberman, E.; Giannoukos, G.; Alvarez, P.; Brockman, W.; Kim, T. K.; Koche, R. P.; Lee, W.; Mendenhall, E.; O'Donovan, A.; Presser, A.; Russ, C.; Xie, X.; Meissner, A.; Wernig, M.; Jaenisch, R.; Nusbaum, C.; Lander, E. S.; Bernstein, B. E. Genome-Wide Maps of Chromatin State in Pluripotent and Lineage-Committed Cells. *Nature* **2007**. https://doi.org/10.1038/nature06008.
(127) Carninci, P.; Sandelin, A.; Lenhard, B.; Katayama, S.; Shimokawa, K.; Ponjavic, J.; Semple, C. A. M.; Taylor, M. S.; Engström, P. G.; Frith, M. C.; Forrest, A. R. R.; Alkema, W. B.; Tan, S. L.; Plessy, C.; Kodzius, R.; Ravasi, T.; Kasukawa, T.; Fukuda, S.; Kanamori-Katayama, M.; Kitazume, Y.; Kawaji, H.; Kai, C.; Nakamura, M.; Konno, H.; Nakano, K.; Mottagui-Tabar, S.; Arner, P.; Chesi, A.; Gustincich, S.; Persichetti, F.; Suzuki, H.; Grimmond, S. M.; Wells, C. A.; Orlando, V.; Wahlestedt, C.; Liu, E. T.; Harbers, M.; Kawai, J.; Bajic, V. B.; Hume, D. A.; Hayashizaki, Y. Genome-Wide Analysis of Mammalian Promoter Architecture and Evolution. *Nat. Genet.* **2006**. https://doi.org/10.1038/ng1789.
(128) Mohn, F.; Weber, M.; Rebhan, M.; Roloff, T. C.; Richter, J.; Stadler, M. B.; Bibel,

M.; Schübeler, D. Lineage-Specific Polycomb Targets and De Novo DNA Methylation Define Restriction and Potential of Neuronal Progenitors. *Mol. Cell* **2008**. https://doi.org/10.1016/j.molcel.2008.05.007.
(129) Wutz, A.; Smrzka, O. W.; Schweifer, N.; Schellander, K.; Wagner, E. F.; Barlow, D. P. Imprinted Expression of the Igf2r Gene Depends on an Intronic CpG Island. *Nature* **1997**. https://doi.org/10.1038/39631.
(130) Caspary, T.; Cleary, M. A.; Baker, C. C.; Guan, X.-J.; Tilghman, S. M. Multiple Mechanisms Regulate Imprinting of the Mouse Distal Chromosome 7 Gene Cluster. *Mol. Cell. Biol.* **1998**. https://doi.org/10.1128/mcb.18.6.3466.
(131) Zwart, R.; Sleutels, F.; Wutz, A.; Schinkel, A. H.; Barlow, D. P. Bidirectional Action of the Igf2r Imprint Control Element on Upstream and Downstream Imprinted Genes. *Genes Dev.* **2001**. https://doi.org/10.1101/gad.206201.
(132) Kantor, B.; Kaufman, Y.; Makedonski, K.; Razin, A.; Shemer, R. Establishing the Epigenetic Status of the Prader-Willi/Angelman Imprinting Center in the Gametes and Embryo. *Hum. Mol. Genet.* **2004**. https://doi.org/10.1093/hmg/ddh290.
(133) Shen, L.; Kondo, Y.; Guo, Y.; Zhang, J.; Zhang, L.; Ahmed, S.; Shu, J.; Chen, X.; Waterland, R. A.; Issa, J. P. J. Genome-Wide Profiling of DNA Methylation Reveals a Class of Normally Methylated CpG Island Promoters. *PLoS Genet.* **2007**. https://doi.org/10.1371/journal.pgen.0030181.
(134) Weber, M.; Hellmann, I.; Stadler, M. B.; Ramos, L.; Pääbo, S.; Rebhan, M.; Schübeler, D. Distribution, Silencing Potential and Evolutionary Impact of Promoter DNA Methylation in the Human Genome. *Nat. Genet.* **2007**. https://doi.org/10.1038/ng1990.
(135) Fouse, S. D.; Shen, Y.; Pellegrini, M.; Cole, S.; Meissner, A.; Van Neste, L.; Jaenisch, R.; Fan, G. Promoter CpG Methylation Contributes to ES Cell Gene Regulation in Parallel with Oct4/Nanog, PcG Complex, and Histone H3 K4/K27 Trimethylation. *Cell Stem Cell* **2008**. https://doi.org/10.1016/j.stem.2007.12.011.
(136) Meissner, A.; Mikkelsen, T. S.; Gu, H.; Wernig, M.; Hanna, J.; Sivachenko, A.; Zhang, X.; Bernstein, B. E.; Nusbaum, C.; Jaffe, D. B.; Gnirke, A.; Jaenisch, R.; Lander, E. S. Genome-Scale DNA Methylation Maps of Pluripotent and Differentiated Cells. *Nature* **2008**. https://doi.org/10.1038/nature07107.
(137) Rakyan, V. K.; Hildmann, T.; Novik, K. L.; Lewin, J.; Tost, J.; Cox, A. V.; Andrews, T. D.; Howe, K. L.; Otto, T.; Olek, A.; Fischer, J.; Gut, I. G.; Berlin, K.; Beck, S. DNA Methylation Profiling of the Human Major Histocompatibility Complex: A Pilot Study for the Human Epigenome Project. *PLoS Biol.* **2004**. https://doi.org/10.1371/journal.pbio.0020405.
(138) Eckhardt, F.; Lewin, J.; Cortese, R.; Rakyan, V. K.; Attwood, J.; Burger, M.; Burton, J.; Cox, T. V.; Davies, R.; Down, T. A.; Haefliger, C.; Horton, R.; Howe, K.; Jackson, D. K.; Kunde, J.; Koenig, C.; Liddle, J.; Niblett, D.; Otto, T.; Pettett, R.; Seemann, S.; Thompson, C.; West, T.; Rogers, J.; Olek, A.; Berlin, K.; Beck, S. DNA Methylation Profiling of Human Chromosomes 6, 20 and 22. *Nat. Genet.* **2006**. https://doi.org/10.1038/ng1909.
(139) Maunakea, A. K.; Nagarajan, R. P.; Bilenky, M.; Ballinger, T. J.; Dsouza, C.; Fouse, S. D.; Johnson, B. E.; Hong, C.; Nielsen, C.; Zhao, Y.; Turecki, G.; Delaney, A.; Varhol, R.; Thiessen, N.; Shchors, K.; Heine, V. M.; Rowitch, D. H.; Xing, X.; Fiore, C.; Schillebeeckx, M.; Jones, S. J. M.; Haussler, D.; Marra, M. A.; Hirst, M.; Wang, T.; Costello, J. F. Conserved Role of Intragenic DNA Methylation in Regulating Alternative Promoters. *Nature* **2010**. https://doi.org/10.1038/nature09165.

(140) Irizarry, R. A.; Ladd-Acosta, C.; Wen, B.; Wu, Z.; Montano, C.; Onyango, P.; Cui, H.; Gabo, K.; Rongione, M.; Webster, M.; Ji, H.; Potash, J. B.; Sabunciyan, S.; Feinberg, A. P. The Human Colon Cancer Methylome Shows Similar Hypo- and Hypermethylation at Conserved Tissue-Specific CpG Island Shores. *Nat. Genet.* **2009**. https://doi.org/10.1038/ng.298.
(141) Brenet, F.; Moh, M.; Funk, P.; Feierstein, E.; Viale, A. J.; Socci, N. D.; Scandura, J. M. DNA Methylation of the First Exon Is Tightly Linked to Transcriptional Silencing. *PLoS One* **2011**. https://doi.org/10.1371/journal.pone.0014524.
(142) Hellman, A.; Chess, A. Gene Body-Specific Methylation on the Active X Chromosome. *Science (80-.).* **2007**. https://doi.org/10.1126/science.1136352.
(143) Ball, M. P.; Li, J. B.; Gao, Y.; Lee, J. H.; Leproust, E. M.; Park, I. H.; Xie, B.; Daley, G. Q.; Church, G. M. Targeted and Genome-Scale Strategies Reveal Gene-Body Methylation Signatures in Human Cells. *Nat. Biotechnol.* **2009**. https://doi.org/10.1038/nbt.1533.
(144) Aran, D.; Toperoff, G.; Rosenberg, M.; Hellman, A. Replication Timing-Related and Gene Body-Specific Methylation of Active Human Genes. *Hum. Mol. Genet.* **2011**. https://doi.org/10.1093/hmg/ddq513.
(145) Guo, J. U.; Ma, D. K.; Mo, H.; Ball, M. P.; Jang, M. H.; Bonaguidi, M. A.; Balazer, J. A.; Eaves, H. L.; Xie, B.; Ford, E.; Zhang, K.; Ming, G. L.; Gao, Y.; Song, H. Neuronal Activity Modifies the DNA Methylation Landscape in the Adult Brain. *Nat. Neurosci.* **2011**. https://doi.org/10.1038/nn.2900.
(146) Guo, J. U.; Su, Y.; Zhong, C.; Ming, G. L.; Song, H. Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain. *Cell* **2011**. https://doi.org/10.1016/j.cell.2011.03.022.
(147) Nan, X.; Meehan, R. R.; Bird, A. Dissection of the Methyl-CpG Binding Domain from the Chromosomal Protein MeCP2. *Nucleic Acids Res.* **1993**. https://doi.org/10.1093/nar/21.21.4886.
(148) Meehan, R. R.; Lewis, J. D.; McKay, S.; Kleiner, E. L.; Bird, A. P. Identification of a Mammalian Protein That Binds Specifically to DNA Containing Methylated CpGs. *Cell* **1989**. https://doi.org/10.1016/0092-8674(89)90430-3.
(149) Lewis, J. D.; Meehan, R. R.; Henzel, W. J.; Maurer-Fogy, I.; Jeppesen, P.; Klein, F.; Bird, A. Purification, Sequence, and Cellular Localization of a Novel Chromosomal Protein That Binds to Methylated DNA. *Cell* **1992**. https://doi.org/10.1016/0092-8674(92)90610-O.
(150) HENDRICH, B.; BIRD, A. Identification and Characterization of a Family of Mammalian Methyl CpG-Binding Proteins. *Genet. Res.* **1998**. https://doi.org/10.1017/s0016672398533307.
(151) Amir, R. E.; Van Den Veyver, I. B.; Wan, M.; Tran, C. Q.; Francke, U.; Zoghbi, H. Y. Rett Syndrome Is Caused by Mutations in X-Linked MECP2, Encoding Methyl- CpG-Binding Protein 2. *Nat. Genet.* **1999**. https://doi.org/10.1038/13810.
(152) Ng, H. H.; Zhang, Y.; Hendrich, B.; Johnson, C. A.; Turner, B. M.; Erdjument-Bromage, H.; Tempst, P.; Reinberg, D.; Bird, A. MBD2 Is a Transcriptional Repressor Belonging to the MeCP1 Histone Deacetylase Complex. *Nat. Genet.* **1999**. https://doi.org/10.1038/12659.
(153) Sarraf, S. A.; Stancheva, I. Methyl-CpG Binding Protein MBD1 Couples Histone H3 Methylation at Lysine 9 by SETDB1 to DNA Replication and Chromatin Assembly. *Molecular Cell*. 2004. https://doi.org/10.1016/j.molcel.2004.06.043.
(154) Kimura, H.; Shiota, K. Methyl-CpG-Binding Protein, MeCP2, Is a Target Molecule for Maintenance DNA Methyltransferase, Dnmt1. *J. Biol. Chem.* **2003**.

https://doi.org/10.1074/jbc.M209923200.
(155) Hashimoto, H.; Horton, J. R.; Zhang, X.; Bostick, M.; Jacobsen, S. E.; Cheng, X. The SRA Domain of UHRF1 Flips 5-Methylcytosine out of the DNA Helix. *Nature* **2008**. https://doi.org/10.1038/nature07280.
(156) Sharif, J.; Muto, M.; Takebayashi, S. I.; Suetake, I.; Iwamatsu, A.; Endo, T. A.; Shinga, J.; Mizutani-Koseki, Y.; Toyoda, T.; Okamura, K.; Tajima, S.; Mitsuya, K.; Okano, M.; Koseki, H. The SRA Protein Np95 Mediates Epigenetic Inheritance by Recruiting Dnmt1 to Methylated DNA. *Nature* **2007**. https://doi.org/10.1038/nature06397.
(157) Bostick, M.; Jong, K. K.; Estève, P. O.; Clark, A.; Pradhan, S.; Jacobsen, S. E. UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells. *Science (80-.).* **2007**. https://doi.org/10.1126/science.1147939.
(158) Achour, M.; Jacq, X.; Rondé, P.; Alhosin, M.; Charlot, C.; Chataigneau, T.; Jeanblanc, M.; Macaluso, M.; Giordano, A.; Hughes, A. D.; Schini-Kerth, V. B.; Bronner, C. The Interaction of the SRA Domain of ICBP90 with a Novel Domain of DNMT1 Is Involved in the Regulation of VEGF Gene Expression. *Oncogene* **2008**. https://doi.org/10.1038/sj.onc.1210855.
(159) Prokhortchouk, A.; Hendrich, B.; Jørgensen, H.; Ruzov, A.; Wilm, M.; Georgiev, G.; Bird, A.; Prokhortchouk, E. The P120 Catenin Partner Kaiso Is a DNA Methylation-Dependent Transcriptional Repressor. *Genes Dev.* **2001**. https://doi.org/10.1101/gad.198501.
(160) Filion, G. J. P.; Zhenilo, S.; Salozhin, S.; Yamada, D.; Prokhortchouk, E.; Defossez, P.-A. A Family of Human Zinc Finger Proteins That Bind Methylated DNA and Repress Transcription. *Mol. Cell. Biol.* **2006**. https://doi.org/10.1128/mcb.26.1.169-181.2006.
(161) Yoon, H. G.; Chan, D. W.; Reynolds, A. B.; Qin, J.; Wong, J. N-CoR Mediates DNA Methylation-Dependent Repression through a Methyl CpG Binding Protein Kaiso. *Mol. Cell* **2003**. https://doi.org/10.1016/j.molcel.2003.08.008.
(162) Lopes, E. C.; Valls, E.; Figueroa, M. E.; Mazur, A.; Meng, F. G.; Chiosis, G.; Laird, P. W.; Schreiber-Agus, N.; Greally, J. M.; Prokhortchouk, E.; Melnick, A. Kaiso Contributes to DNA Methylation-Dependent Silencing of Tumor Suppressor Genes in Colon Cancer Cell Lines. *Cancer Res.* **2008**. https://doi.org/10.1158/0008-5472.CAN-08-0344.
(163) Nan, X.; Ng, H. H.; Johnson, C. A.; Laherty, C. D.; Turner, B. M.; Eisenman, R. N.; Bird, A. Transcriptional Repression by the Methyl-CpG-Binding Protein MeCP2 Involves a Histone Deacetylase Complex. *Nature* **1998**. https://doi.org/10.1038/30764.
(164) Citterio, E.; Papait, R.; Nicassio, F.; Vecchi, M.; Gomiero, P.; Mantovani, R.; Di Fiore, P. P.; Bonapace, I. M. Np95 Is a Histone-Binding Protein Endowed with Ubiquitin Ligase Activity. *Mol. Cell. Biol.* **2004**. https://doi.org/10.1128/mcb.24.6.2526-2535.2004.
(165) Karagianni, P.; Amazit, L.; Qin, J.; Wong, J. ICBP90, a Novel Methyl K9 H3 Binding Protein Linking Protein Ubiquitination with Heterochromatin Formation. *Mol. Cell. Biol.* **2008**. https://doi.org/10.1128/mcb.01598-07.
(166) Jones, P. L.; Veenstra, G. J. C.; Wade, P. A.; Vermaak, D.; Kass, S. U.; Landsberger, N.; Strouboulis, J.; Wolffe, A. P. Methylated DNA and MeCP2 Recruit Histone Deacetylase to Repress Transcription. *Nat. Genet.* **1998**. https://doi.org/10.1038/561.
(167) Fuks, F.; Hurd, P. J.; Deplus, R.; Kouzarides, T. The DNA Methyltransferases Associate with HP1 and the SUV39H1 Histone Methyltransferase. *Nucleic Acids*

Res. **2003**. https://doi.org/10.1093/nar/gkg332.
(168) Ehrich, M.; Nelson, M. R.; Stanssens, P.; Zabeau, M.; Liloglou, T.; Xinarianos, G.; Cantor, C. R.; Field, J. K.; Van Den Boom, D. Quantitative High-Throughput Analysis of DNA Methylation Patterns by Base-Specific Cleavage and Mass Spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **2005**. https://doi.org/10.1073/pnas.0507816102.
(169) Matsuda, S.; Yasukawa, T.; Sakaguchi, Y.; Ichiyanagi, K.; Unoki, M.; Gotoh, K.; Fukuda, K.; Sasaki, H.; Suzuki, T.; Kang, D. Accurate Estimation of 5-Methylcytosine in Mammalian Mitochondrial DNA. *Sci. Rep.* **2018**, *8* (1), 5801. https://doi.org/10.1038/s41598-018-24251-z.
(170) Nass, M. M. K. Differential Methylation of Mitochondrial and Nuclear DNA in Cultured Mouse, Hamster and Virus-Transformed Hamster Cells In Vivo and in Vitro Methylation. *J. Mol. Biol.* **1973**. https://doi.org/10.1016/0022-2836(73)90239-8.
(171) Vanyushin, B. F.; Kirnos, M. D. Structure of Animal Mitochondrial DNA (Base Composition, Pyrimidine Clusters, Character of Methylation). *BBA Sect. Nucleic Acids Protein Synth.* **1977**. https://doi.org/10.1016/0005-2787(77)90023-5.
(172) Treangen, T. J.; Salzberg, S. L. Repetitive DNA and Next-Generation Sequencing: Computational Challenges and Solutions. *Nature Reviews Genetics*. 2012. https://doi.org/10.1038/nrg3117.
(173) Masser, D. R.; Berg, A. S.; Freeman, W. M. Focused, High Accuracy 5-Methylcytosine Quantitation with Base Resolution by Benchtop next-Generation Sequencing. *Epigenetics and Chromatin* **2013**. https://doi.org/10.1186/1756-8935-6-33.
(174) Lee, E. J.; Pei, L.; Srivastava, G.; Joshi, T.; Kushwaha, G.; Choi, J. H.; Robertson, K. D.; Wang, X.; Colbourne, J. K.; Zhang, L.; Schroth, G. P.; Xu, D.; Zhang, K.; Shi, H. Targeted Bisulfite Sequencing by Solution Hybrid Selection and Massively Parallel Sequencing. *Nucleic Acids Res.* **2011**. https://doi.org/10.1093/nar/gkr598.
(175) Li, Q.; Suzuki, M.; Wendt, J.; Patterson, N.; Eichten, S. R.; Hermanson, P. J.; Green, D.; Jeddeloh, J.; Richmond, T.; Rosenbaum, H.; Burgess, D.; Springer, N. M.; Greally, J. M. Post-Conversion Targeted Capture of Modified Cytosines in Mammalian and Plant Genomes. *Nucleic Acids Res.* **2015**. https://doi.org/10.1093/nar/gkv244.
(176) Masser, D. R.; Stanford, D. R.; Hadad, N.; Giles, C. B.; Wren, J. D.; Sonntag, W. E.; Richardson, A.; Freeman, W. M. Bisulfite Oligonucleotide-Capture Sequencing for Targeted Base- and Strand-Specific Absolute 5-Methylcytosine Quantitation. *Age (Omaha).* **2016**. https://doi.org/10.1007/s11357-016-9914-1.
(177) Wendt, J.; Rosenbaum, H.; Richmond, T. A.; Jeddeloh, J. A.; Burgess, D. L. Targeted Bisulfite Sequencing Using the SeqCap Epi Enrichment System. In *Methods in Molecular Biology*; 2018. https://doi.org/10.1007/978-1-4939-7481-8_20.
(178) Kacmarczyk, T. J.; Fall, M. P.; Zhang, X.; Xin, Y.; Li, Y.; Alonso, A.; Betel, D. "same Difference": Comprehensive Evaluation of Four DNA Methylation Measurement Platforms. *Epigenetics and Chromatin* **2018**. https://doi.org/10.1186/s13072-018-0190-4.
(179) Meissner, A.; Gnirke, A.; Bell, G. W.; Ramsahoye, B.; Lander, E. S.; Jaenisch, R. Reduced Representation Bisulfite Sequencing for Comparative High-Resolution DNA Methylation Analysis. *Nucleic Acids Res.* **2005**. https://doi.org/10.1093/nar/gki901.

(180) Gu, H.; Smith, Z. D.; Bock, C.; Boyle, P.; Gnirke, A.; Meissner, A. Preparation of Reduced Representation Bisulfite Sequencing Libraries for Genome-Scale DNA Methylation Profiling. *Nat. Protoc.* **2011**. https://doi.org/10.1038/nprot.2010.190.
(181) Akalin, A.; Garrett-Bakelman, F. E.; Kormaksson, M.; Busuttil, J.; Zhang, L.; Khrebtukova, I.; Milne, T. A.; Huang, Y.; Biswas, D.; Hess, J. L.; Allis, C. D.; Roeder, R. G.; Valk, P. J. M.; Löwenberg, B.; Delwel, R.; Fernandez, H. F.; Paietta, E.; Tallman, M. S.; Schroth, G. P.; Mason, C. E.; Melnick, A.; Figueroa, M. E. Base-Pair Resolution DNA Methylation Sequencing Reveals Profoundly Divergent Epigenetic Landscapes in Acute Myeloid Leukemia. *PLoS Genet.* **2012**. https://doi.org/10.1371/journal.pgen.1002781.
(182) Weber, M.; Davies, J. J.; Wittig, D.; Oakeley, E. J.; Haase, M.; Lam, W. L.; Schübeler, D. Chromosome-Wide and Promoter-Specific Analyses Identify Sites of Differential DNA Methylation in Normal and Transformed Human Cells. *Nat. Genet.* **2005**. https://doi.org/10.1038/ng1598.
(183) Yong, W. S.; Hsu, F. M.; Chen, P. Y. Profiling Genome-Wide DNA Methylation. *Epigenetics and Chromatin*. 2016. https://doi.org/10.1186/s13072-016-0075-3.
(184) Staunstrup, N. H.; Starnawska, A.; Nyegaard, M.; Christiansen, L.; Nielsen, A. L.; Børglum, A.; Mors, O. Genome-Wide DNA Methylation Profiling with MeDIP-Seq Using Archived Dried Blood Spots. *Clin. Epigenetics* **2016**. https://doi.org/10.1186/s13148-016-0242-1.
(185) Devall, M.; Smith, R. G.; Jeffries, A.; Hannon, E.; Davies, M. N.; Schalkwyk, L.; Mill, J.; Weedon, M.; Lunnon, K. Regional Differences in Mitochondrial DNA Methylation in Human Post-Mortem Brain Tissue. *Clin. Epigenetics* **2017**. https://doi.org/10.1186/s13148-017-0337-3.
(186) Wolters, J. E. J.; Van Breda, S. G. J.; Caiment, F.; Claessen, S. M.; De Kok, T. M. C. M.; Kleinjans, J. C. S. Nuclear and Mitochondrial DNA Methylation Patterns Induced by Valproic Acid in Human Hepatocytes. *Chem. Res. Toxicol.* **2017**. https://doi.org/10.1021/acs.chemrestox.7b00171.
(187) Amarasinghe, S. L.; Su, S.; Dong, X.; Zappia, L.; Ritchie, M. E.; Gouil, Q. Opportunities and Challenges in Long-Read Sequencing Data Analysis. *Genome Biology*. 2020. https://doi.org/10.1186/s13059-020-1935-5.
(188) Flusberg, B. A.; Webster, D. R.; Lee, J. H.; Travers, K. J.; Olivares, E. C.; Clark, T. A.; Korlach, J.; Turner, S. W. Direct Detection of DNA Methylation during Single-Molecule, Real-Time Sequencing. *Nat. Methods* **2010**. https://doi.org/10.1038/nmeth.1459.
(189) Xu, L.; Seki, M. Recent Advances in the Detection of Base Modifications Using the Nanopore Sequencer. *Journal of Human Genetics*. 2020. https://doi.org/10.1038/s10038-019-0679-0.
(190) Beaulaurier, J.; Zhang, X. S.; Zhu, S.; Sebra, R.; Rosenbluh, C.; Deikus, G.; Shen, N.; Munera, D.; Waldor, M. K.; Chess, A.; Blaser, M. J.; Schadt, E. E.; Fang, G. Single Molecule-Level Detection and Long Read-Based Phasing of Epigenetic Variations in Bacterial Methylomes. *Nat. Commun.* **2015**. https://doi.org/10.1038/ncomms8438.
(191) Gigante, S.; Gouil, Q.; Lucattini, A.; Keniry, A.; Beck, T.; Tinning, M.; Gordon, L.; Woodruff, C.; Speed, T. P.; Blewitt, M. E.; Ritchie, M. E. Using Long-Read Sequencing to Detect Imprinted DNA Methylation. *Nucleic Acids Res.* **2019**. https://doi.org/10.1093/nar/gkz107.
(192) Rhoads, A.; Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*. 2015. https://doi.org/10.1016/j.gpb.2015.08.002.

(193) Clark, T. A.; Lu, X.; Luong, K.; Dai, Q.; Boitano, M.; Turner, S. W.; He, C.; Korlach, J. Enhanced 5-Methylcytosine Detection in Single-Molecule, Real-Time Sequencing via Tet1 Oxidation. *BMC Biol.* **2013**. https://doi.org/10.1186/1741-7007-11-4.
(194) Delahaye, C.; Nicolas, J. Sequencing DNA with Nanopores: Troubles and Biases. *PLoS One* **2021**, *16* (10 October). https://doi.org/10.1371/journal.pone.0257521.
(195) Wick, R. R.; Judd, L. M.; Holt, K. E. Performance of Neural Network Basecalling Tools for Oxford Nanopore Sequencing. *Genome Biol.* **2019**. https://doi.org/10.1186/s13059-019-1727-y.
(196) Rang, F. J.; Kloosterman, W. P.; de Ridder, J. From Squiggle to Basepair: Computational Approaches for Improving Nanopore Sequencing Read Accuracy. *Genome Biology*. 2018. https://doi.org/10.1186/s13059-018-1462-9.
(197) Stoler, N.; Nekrutenko, A. Sequencing Error Profiles of Illumina Sequencing Instruments. *NAR Genomics Bioinforma.* **2021**, *3* (1). https://doi.org/10.1093/nargab/lqab019.
(198) Simpson, J. T.; Workman, R. E.; Zuzarte, P. C.; David, M.; Dursi, L. J.; Timp, W. Detecting DNA Cytosine Methylation Using Nanopore Sequencing. *Nat. Methods* **2017**. https://doi.org/10.1038/nmeth.4184.
(199) Jain, M.; Koren, S.; Miga, K. H.; Quick, J.; Rand, A. C.; Sasani, T. A.; Tyson, J. R.; Beggs, A. D.; Dilthey, A. T.; Fiddes, I. T.; Malla, S.; Marriott, H.; Nieto, T.; O'Grady, J.; Olsen, H. E.; Pedersen, B. S.; Rhie, A.; Richardson, H.; Quinlan, A. R.; Snutch, T. P.; Tee, L.; Paten, B.; Phillippy, A. M.; Simpson, J. T.; Loman, N. J.; Loose, M. Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads. *Nat. Biotechnol.* **2018**. https://doi.org/10.1038/nbt.4060.
(200) Rand, A. C.; Jain, M.; Eizenga, J. M.; Musselman-Brown, A.; Olsen, H. E.; Akeson, M.; Paten, B. Mapping DNA Methylation with High-Throughput Nanopore Sequencing. *Nat. Methods* **2017**. https://doi.org/10.1038/nmeth.4189.
(201) McIntyre, A. B. R.; Alexander, N.; Grigorev, K.; Bezdan, D.; Sichtig, H.; Chiu, C. Y.; Mason, C. E. Single-Molecule Sequencing Detection of N6-Methyladenine in Microbial Reference Materials. *Nat. Commun.* **2019**. https://doi.org/10.1038/s41467-019-08289-9.
(202) Ni, P.; Huang, N.; Zhang, Z.; Wang, D. P.; Liang, F.; Miao, Y.; Xiao, C. Le; Luo, F.; Wang, J. DeepSignal: Detecting DNA Methylation State from Nanopore Sequencing Reads Using Deep-Learning. *Bioinformatics* **2019**. https://doi.org/10.1093/bioinformatics/btz276.
(203) Liu, Q.; Fang, L.; Yu, G.; Wang, D.; Xiao, C. Le; Wang, K. Detection of DNA Base Modifications by Deep Recurrent Neural Network on Oxford Nanopore Sequencing Data. *Nat. Commun.* **2019**. https://doi.org/10.1038/s41467-019-10168-2.
(204) Stoiber, M.; Quick, J.; Egan, R.; Eun Lee, J.; Celniker, S.; Neely, R.; Loman, N.; Pennacchio, L.; Brown, J. De Novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv* **2016**. https://doi.org/10.1101/094672.
(205) Vanyushin, B. F.; Kiryanov, G. I.; Kudryashova, I. B.; Belozersky, A. N. DNA-Methylase in Loach Embryos (Misgurnus Fossilis). *FEBS Lett.* **1971**. https://doi.org/10.1016/0014-5793(71)80646-4.
(206) Dawid, I. B. 5-Methylcytidylic Acid: Absence from Mitochondrial DNA of Frogs and HeLa Cells. *Science (80-.).* **1974**.

https://doi.org/10.1126/science.184.4132.80.
(207) Vanyushin, B. F.; Kirnos, M. D. The Nucleotide Composition and Pyrimidine Clusters in DNA from Beef Heart Mitochondria. *FEBS Lett.* **1974**. https://doi.org/10.1016/0014-5793(74)80049-9.
(208) Mushkambarov, N. N.; Votrin, I. I.; Debov, S. S. Methylation of Preformed DNA in Cell Nuclei and Mitochondria of the Rat Liver. *Doklady Biochemistry*. 1977.
(209) Shmookler Reis, R. J.; Goldstein, S. Mitochondrial DNA in Mortal and Immortal Human Cells. Genome Number, Integrity, and Methylation. *J. Biol. Chem.* **1983**. https://doi.org/10.1016/s0021-9258(17)44633-3.
(210) Shock, L. S.; Thakkar, P. V.; Peterson, E. J.; Moran, R. G.; Taylor, S. M. DNA Methyltransferase 1, Cytosine Methylation, and Cytosine Hydroxymethylation in Mammalian Mitochondria. *Proc. Natl. Acad. Sci. U. S. A.* **2011**. https://doi.org/10.1073/pnas.1012311108.
(211) Saini, S. K.; Mangalhara, K. C.; Prakasam, G.; Bamezai, R. N. K. DNA Methyltransferase1 (DNMT1) Isoform3 Methylates Mitochondrial Genome and Modulates Its Biology. *Sci. Rep.* **2017**. https://doi.org/10.1038/s41598-017-01743-y.
(212) Janssen, B. G.; Madhloum, N.; Gyselaers, W.; Bijnens, E.; Clemente, D. B.; Cox, B.; Hogervorst, J.; Luyten, L.; Martens, D. S.; Peusens, M.; Plusquin, M.; Provost, E. B.; Roels, H. A.; Saenen, N. D.; Tsamou, M.; Vriens, A.; Winckelmans, E.; Vrijens, K.; Nawrot, T. S. Cohort Profile: The ENVIRonmental Influence on Early AGEing (ENVIRONAGE): A Birth Cohort Study. *Int. J. Epidemiol.* **2017**. https://doi.org/10.1093/ije/dyw269.
(213) Janssen, B. G.; Gyselaers, W.; Byun, H. M.; Roels, H. A.; Cuypers, A.; Baccarelli, A. A.; Nawrot, T. S. Placental Mitochondrial DNA and CYP1A1 Gene Methylation as Molecular Signatures for Tobacco Smoke Exposure in Pregnant Women and the Relevance for Birth Weight. *J. Transl. Med.* **2017**. https://doi.org/10.1186/s12967-016-1113-4.
(214) Byun, H. M.; Panni, T.; Motta, V.; Hou, L.; Nordio, F.; Apostoli, P.; Bertazzi, P. A.; Baccarelli, A. A. Effects of Airborne Pollutants on Mitochondrial DNA Methylation. *Part. Fibre Toxicol.* **2013**. https://doi.org/10.1186/1743-8977-10-18.
(215) Janssen, B. G.; Byun, H. M.; Cox, B.; Gyselaers, W.; Izzi, B.; Baccarelli, A. A.; Nawrot, T. S. Variation of DNA Methylation in Candidate Age-Related Targets on the Mitochondrial-Telomere Axis in Cord Blood and Placenta. *Placenta* **2014**. https://doi.org/10.1016/j.placenta.2014.06.371.
(216) Byun, H. M.; Barrow, T. M. Analysis of Pollutant-Induced Changes in Mitochondrial DNA Methylation. In *Mitochondrial Medicine*; 2015. https://doi.org/10.1007/978-1-4939-2288-8_19.
(217) Janssen, B. G.; Byun, H. M.; Gyselaers, W.; Lefebvre, W.; Baccarelli, A. A.; Nawrot, T. S. Placental Mitochondrial Methylation and Exposure to Airborne Particulate Matter in the Early Life Environment: An ENVIRONAGE Birth Cohort Study. *Epigenetics* **2015**. https://doi.org/10.1080/15592294.2015.1048412.
(218) Byun, H. M.; Colicino, E.; Trevisi, L.; Fan, T.; Christiani, D. C.; Baccarelli, A. A. Effects of Air Pollution and Blood Mitochondrial DNA Methylation on Markers of Heart Rate Variability. *J. Am. Heart Assoc.* **2016**. https://doi.org/10.1161/JAHA.116.003218.
(219) Vos, S.; Nawrot, T. S.; Martens, D. S.; Byun, H. M.; Janssen, B. G. Mitochondrial DNA Methylation in Placental Tissue: A Proof of Concept Study by Means of Prenatal Environmental Stressors. *Epigenetics* **2020**.

https://doi.org/10.1080/15592294.2020.1790923.
(220) Novielli, C.; Mandò, C.; Tabano, S.; Anelli, G. M.; Fontana, L.; Antonazzo, P.; Miozzo, M.; Cetin, I. Mitochondrial DNA Content and Methylation in Fetal Cord Blood of Pregnancies with Placental Insufficiency. *Placenta* **2017**. https://doi.org/10.1016/j.placenta.2017.05.008.
(221) Maekawa, M.; Taniguchi, T.; Higashi, H.; Sugimura, H.; Sugano, K.; Kanno, T. Methylation of Mitochondrial DNA Is Not a Useful Marker for Cancer Detection [6]. *Clinical Chemistry*. 2004. https://doi.org/10.1373/clinchem.2004.035139.
(222) Sun, C.; Reimers, L. L.; Burk, R. D. Methylation of HPV16 Genome CpG Sites Is Associated with Cervix Precancer and Cancer. *Gynecol. Oncol.* **2011**. https://doi.org/10.1016/j.ygyno.2011.01.013.
(223) Feng, S.; Xiong, L.; Ji, Z.; Cheng, W.; Yang, H. Correlation between Increased ND2 Expression and Demethylated Displacement Loop of MtDNA in Colorectal Cancer. *Mol. Med. Rep.* **2012**. https://doi.org/10.3892/mmr.2012.870.
(224) Gao, J.; Wen, S.; Zhou, H.; Feng, S. De-Methylation of Displacement Loop of Mitochondrial DNA Is Associated with Increased Mitochondrial Copy Number and Nicotinamide Adenine Dinucleotide Subunit 2 Expression in Colorectal Cancer. *Mol. Med. Rep.* **2015**. https://doi.org/10.3892/mmr.2015.4256.
(225) Tong, H.; Zhang, L.; Gao, J.; Wen, S.; Zhou, H.; Feng, S. Methylation of Mitochondrial DNA Displacement Loop Region Regulates Mitochondrial Copy Number in Colorectal Cancer. *Mol. Med. Rep.* **2017**. https://doi.org/10.3892/mmr.2017.7264.
(226) Sun, X.; Vaghjiani, V.; Jayasekara, W. S. N.; Cain, J. E.; St John, J. C. The Degree of Mitochondrial DNA Methylation in Tumor Models of Glioblastoma and Osteosarcoma. *Clin. Epigenetics* **2018**. https://doi.org/10.1186/s13148-018-0590-0.
(227) D'Aquila, P.; Giordano, M.; Montesanto, A.; De Rango, F.; Passarino, G.; Bellizzi, D. Age-and Gender-Related Pattern of Methylation in the MT-RNR1 Gene. *Epigenomics* **2015**. https://doi.org/10.2217/epi.15.30.
(228) Mawlood, S. K.; Dennany, L.; Watson, N.; Dempster, J.; Pickard, B. S. Quantification of Global Mitochondrial DNA Methylation Levels and Inverse Correlation with Age at Two CpG Sites. *Aging (Albany. NY).* **2016**. https://doi.org/10.18632/aging.100892.
(229) Chestnut, B. A.; Chang, Q.; Price, A.; Lesuisse, C.; Wong, M.; Martin, L. J. Epigenetic Regulation of Motor Neuron Cell Death through DNA Methylation. *J. Neurosci.* **2011**. https://doi.org/10.1523/JNEUROSCI.1639-11.2011.
(230) Wong, M.; Gertz, B.; Chestnut, B. A.; Martin, L. J. Mitochondrial DNMT3A and DNA Methylation in Skeletal Muscle and CNS of Transgenic Mouse Models of ALS. *Front. Cell. Neurosci.* **2013**. https://doi.org/10.3389/fncel.2013.00279.
(231) Stoccoro, A.; Siciliano, G.; Migliore, L.; Coppedè, F. Decreased Methylation of the Mitochondrial D-Loop Region in Late-Onset Alzheimer's Disease. *J. Alzheimer's Dis.* **2017**. https://doi.org/10.3233/JAD-170139.
(232) Blanch, M.; Mosquera, J. L.; Ansoleaga, B.; Ferrer, I.; Barrachina, M. Altered Mitochondrial DNA Methylation Pattern in Alzheimer Disease-Related Pathology and in Parkinson Disease. *Am. J. Pathol.* **2016**. https://doi.org/10.1016/j.ajpath.2015.10.004.
(233) Sun, Z.; Terragni, J.; Borgaro, J. G.; Liu, Y.; Yu, L.; Guan, S.; Wang, H.; Sun, D.; Cheng, X.; Zhu, Z.; Pradhan, S.; Zheng, Y. High-Resolution Enzymatic Mapping of Genomic 5-Hydroxymethylcytosine in Mouse Embryonic Stem Cells. *Cell Rep.* **2013**. https://doi.org/10.1016/j.celrep.2013.01.001.

(234) Bellizzi, D.; D'aquila, P.; Scafone, T.; Giordano, M.; Riso, V.; Riccio, A.; Passarino, G. The Control Region of Mitochondrial DNA Shows an Unusual CpG and Non-CpG Methylation Pattern. *DNA Res.* **2013**. https://doi.org/10.1093/dnares/dst029.
(235) Ren, L.; Zhang, C.; Tao, L.; Hao, J.; Tan, K.; Miao, K.; Yu, Y.; Sui, L.; Wu, Z.; Tian, J.; An, L. High-Resolution Profiles of Gene Expression and DNA Methylation Highlight Mitochondrial Modifications during Early Embryonic Development. *J. Reprod. Dev.* **2017**. https://doi.org/10.1262/jrd.2016-168.
(236) Mishra, M.; Kowluru, R. A. Epigenetic Modification of Mitochondrial DNA in the Development of Diabetic Retinopathy. *Investig. Ophthalmol. Vis. Sci.* **2015**. https://doi.org/10.1167/iovs.15-16937.
(237) Bianchessi, V.; Vinci, M. C.; Nigro, P.; Rizzi, V.; Farina, F.; Capogrossi, M. C.; Pompilio, G.; Gualdi, V.; Lauri, A. Methylation Profiling by Bisulfite Sequencing Analysis of the MtDNA Non-Coding Region in Replicative and Senescent Endothelial Cells. *Mitochondrion* **2016**. https://doi.org/10.1016/j.mito.2016.02.004.
(238) Corsi, S.; Iodice, S.; Vigna, L.; Cayir, A.; Mathers, J. C.; Bollati, V.; Byun, H. M. Platelet Mitochondrial DNA Methylation Predicts Future Cardiovascular Outcome in Adults with Overweight and Obesity. *Clin. Epigenetics* **2020**. https://doi.org/10.1186/s13148-020-00825-5.
(239) Hong, E. E.; Okitsu, C. Y.; Smith, A. D.; Hsieh, C.-L. Regionally Specific and Genome-Wide Analyses Conclusively Demonstrate the Absence of CpG Methylation in Human Mitochondrial DNA. *Mol. Cell. Biol.* **2013**. https://doi.org/10.1128/mcb.00220-13.
(240) Mechta, M.; Ingerslev, L. R.; Fabre, O.; Picard, M.; Barrès, R. Evidence Suggesting Absence of Mitochondrial DNA Methylation. *Front. Genet.* **2017**. https://doi.org/10.3389/fgene.2017.00166.
(241) Kolesar, J. E.; Wang, C. Y.; Taguchi, Y. V.; Chou, S. H.; Kaufman, B. A. Two-Dimensional Intact Mitochondrial DNA Agarose Electrophoresis Reveals the Structural Complexity of the Mammalian Mitochondrial Genome. *Nucleic Acids Res.* **2013**. https://doi.org/10.1093/nar/gks1324.
(242) Olova, N.; Krueger, F.; Andrews, S.; Oxley, D.; Berrens, R. V.; Branco, M. R.; Reik, W. Comparison of Whole-Genome Bisulfite Sequencing Library Preparation Strategies Identifies Sources of Biases Affecting DNA Methylation Data. *Genome Biol.* **2018**, *19* (1), 33. https://doi.org/10.1186/s13059-018-1408-2.
(243) Tanaka, K.; Okamoto, A. Degradation of DNA by Bisulfite Treatment. *Bioorganic Med. Chem. Lett.* **2007**. https://doi.org/10.1016/j.bmcl.2007.01.040.
(244) Iacobazzi, V.; Castegna, A.; Infantino, V.; Andria, G. Mitochondrial DNA Methylation as a Next-Generation Biomarker and Diagnostic Tool. *Molecular Genetics and Metabolism*. 2013. https://doi.org/10.1016/j.ymgme.2013.07.012.
(245) Jayaprakash, A. D.; Benson, E. K.; Gone, S.; Liang, R.; Shim, J.; Lambertini, L.; Toloue, M. M.; Wigler, M.; Aaronson, S. A.; Sachidanandam, R. Stable Heteroplasmy at the Single-Cell Level Is Facilitated by Intercellular Exchange of MtDNA. *Nucleic Acids Res.* **2015**. https://doi.org/10.1093/nar/gkv052.
(246) O'Hara, R.; Tedone, E.; Ludlow, A.; Huang, E.; Arosio, B.; Mari, D.; Shay, J. W. Quantitative Mitochondrial DNA Copy Number Determination Using Droplet Digital PCR with Single-Cell Resolution. *Genome Res.* **2019**. https://doi.org/10.1101/gr.250480.119.
(247) Roadmap Epigenomics Consortium; Kundaje, A.; Meuleman, W.; Ernst, J.;

Bilenky, M.; Yen, A.; Heravi-Moussavi, A.; Kheradpour, P.; Zhang, Z.; Wang, J.; Ziller, M. J.; Amin, V.; Whitaker, J. W.; Schultz, M. D.; Ward, L. D.; Sarkar, A.; Quon, G.; Sandstrom, R. S.; Eaton, M. L.; Wu, Y. C.; Pfenning, A. R.; Wang, X.; Claussnitzer, M.; Liu, Y.; Coarfa, C.; Harris, R. A.; Shoresh, N.; Epstein, C. B.; Gjoneska, E.; Leung, D.; Xie, W.; Hawkins, R. D.; Lister, R.; Hong, C.; Gascard, P.; Mungall, A. J.; Moore, R.; Chuah, E.; Tam, A.; Canfield, T. K.; Hansen, R. S.; Kaul, R.; Sabo, P. J.; Bansal, M. S.; Carles, A.; Dixon, J. R.; Farh, K. H.; Feizi, S.; Karlic, R.; Kim, A. R.; Kulkarni, A.; Li, D.; Lowdon, R.; Elliott, G.; Mercer, T. R.; Neph, S. J.; Onuchic, V.; Polak, P.; Rajagopal, N.; Ray, P.; Sallari, R. C.; Siebenthall, K. T.; Sinnott-Armstrong, N. A.; Stevens, M.; Thurman, R. E.; Wu, J.; Zhang, B.; Zhou, X.; Beaudet, A. E.; Boyer, L. A.; De Jager, P. L.; Farnham, P. J.; Fisher, S. J.; Haussler, D.; Jones, S. J. M.; Li, W.; Marra, M. A.; McManus, M. T.; Sunyaev, S.; Thomson, J. A.; Tlsty, T. D.; Tsai, L. H.; Wang, W.; Waterland, R. A.; Zhang, M. Q.; Chadwick, L. H.; Bernstein, B. E.; Costello, J. F.; Ecker, J. R.; Hirst, M.; Meissner, A.; Milosavljevic, A.; Ren, B.; Stamatoyannopoulos, J. A.; Wang, T.; Kellis, M. Integrative Analysis of 111 Reference Human Epigenomes. *Nature* **2015**. https://doi.org/10.1038/nature14248.

(248) Andrews, S. FASTQC A Quality Control Tool for High Throughput Sequence Data. *Babraham Inst.* **2015**.

(249) Krueger, F. Trim Galore https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

(250) Langmead, B.; Salzberg, S. L. Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* **2012**. https://doi.org/10.1038/nmeth.1923.

(251) Krueger, F.; Andrews, S. R. Bismark: A Flexible Aligner and Methylation Caller for Bisulfite-Seq Applications. *Bioinformatics* **2011**. https://doi.org/10.1093/bioinformatics/btr167.

(252) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **2009**. https://doi.org/10.1093/bioinformatics/btp352.

(253) De Coster, W.; D'Hert, S.; Schultz, D. T.; Cruts, M.; Van Broeckhoven, C. NanoPack: Visualizing and Processing Long-Read Sequencing Data. *Bioinformatics* **2018**. https://doi.org/10.1093/bioinformatics/bty149.

(254) Li, H. Minimap2: Pairwise Alignment for Nucleotide Sequences. *Bioinformatics* **2018**. https://doi.org/10.1093/bioinformatics/bty191.

(255) Calabrese, C.; Simone, D.; Diroma, M. A.; Santorsola, M.; Gutta, C.; Gasparre, G.; Picardi, E.; Pesole, G.; Attimonelli, M. MToolBox: A Highly Automated Pipeline for Heteroplasmy Annotation and Prioritization Analysis of Human Mitochondrial Variants in High-Throughput Sequencing. *Bioinformatics* **2014**. https://doi.org/10.1093/bioinformatics/btu483.

(256) van Oven, M. PhyloTree Build 17: Growing the Human Mitochondrial DNA Tree. *Forensic Sci. Int. Genet. Suppl. Ser.* **2015**. https://doi.org/10.1016/j.fsigss.2015.09.155.

(257) Weissensteiner, H.; Pacher, D.; Kloss-Brandstätter, A.; Forer, L.; Specht, G.; Bandelt, H. J.; Kronenberg, F.; Salas, A.; Schönherr, S. HaploGrep 2: Mitochondrial Haplogroup Classification in the Era of High-Throughput Sequencing. *Nucleic Acids Res.* **2016**. https://doi.org/10.1093/nar/gkw233.

(258) Park, Y.; Wu, H. Differential Methylation Analysis for BS-Seq Data under General Experimental Design. *Bioinformatics* **2016**. https://doi.org/10.1093/bioinformatics/btw026.

(259) Siegfried, Z.; Simon, I. DNA Methylation and Gene Expression. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2010**, *2* (3), 362–371. https://doi.org/10.1002/wsbm.64.
(260) Dong, Z.; Pu, L.; Cui, H. Mitoepigenetics and Its Emerging Roles in Cancer. *Frontiers in Cell and Developmental Biology*. 2020. https://doi.org/10.3389/fcell.2020.00004.
(261) Liu, B.; Du, Q.; Chen, L.; Fu, G.; Li, S.; Fu, L.; Zhang, X.; Ma, C.; Bin, C. CpG Methylation Patterns of Human Mitochondrial DNA. *Sci. Rep.* **2016**. https://doi.org/10.1038/srep23421.
(262) Sirard, M. A. Distribution and Dynamics of Mitochondrial DNA Methylation in Oocytes, Embryos and Granulosa Cells. *Sci. Rep.* **2019**. https://doi.org/10.1038/s41598-019-48422-8.
(263) Krueger, F.; Kreck, B.; Franke, A.; Andrews, S. R. DNA Methylome Analysis Using Short Bisulfite Sequencing Data. *Nature Methods*. 2012. https://doi.org/10.1038/nmeth.1828.
(264) Dou, X.; Boyd-Kirkup, J. D.; McDermott, J.; Zhang, X.; Li, F.; Rong, B.; Zhang, R.; Miao, B.; Chen, P.; Cheng, H.; Xue, J.; Bennett, D.; Wong, J.; Lan, F.; Han, J. D. J. The Strand-Biased Mitochondrial DNA Methylome and Its Regulation by DNMT3A. *Genome Res.* **2019**. https://doi.org/10.1101/gr.234021.117.
(265) Anderson, S.; Bankier, A. T.; Barrell, B. G.; De Bruijn, M. H. L.; Coulson, A. R.; Drouin, J.; Eperon, I. C.; Nierlich, D. P.; Roe, B. A.; Sanger, F.; Schreier, P. H.; Smith, A. J. H.; Staden, R.; Young, I. G. Sequence and Organization of the Human Mitochondrial Genome. *Nature* **1981**. https://doi.org/10.1038/290457a0.
(266) Van Der Wijst, M. G. P.; Van Tilburg, A. Y.; Ruiters, M. H. J.; Rots, M. G. Experimental Mitochondria-Targeted DNA Methylation Identifies GpC Methylation, Not CpG Methylation, as Potential Regulator of Mitochondrial Gene Expression. *Sci. Rep.* **2017**. https://doi.org/10.1038/s41598-017-00263-z.
(267) Park, S. H.; Lee, S. Y.; Kim, S. A. Mitochondrial DNA Methylation Is Higher in Acute Coronary Syndrome than in Stable Coronary Artery Disease. *In Vivo (Brooklyn).* **2021**. https://doi.org/10.21873/INVIVO.12247.
(268) Sun, X.; Wang, Z.; Cong, X.; Lv, Y.; Li, Z.; Rong, L.; Yang, T.; Yu, D. Mitochondrial Gene COX2 Methylation and Downregulation Is a Biomarker of Aging in Heart Mesenchymal Stem Cells. *Int. J. Mol. Med.* **2021**. https://doi.org/10.3892/ijmm.2020.4799.
(269) Tao, X.; Zhan, Y.; Scott, R. T.; Seli, E. ASSESSMENT OF MITOCHONDRIAL DNA METHYLATION IN HUMAN BLASTOCYSTS. *Fertil. Steril.* **2020**. https://doi.org/10.1016/j.fertnstert.2020.08.1042.
(270) Patil, V.; Cuenin, C.; Chung, F.; Aguilera, J. R. R.; Fernandez-Jimenez, N.; Romero-Garmendia, I.; Bilbao, J. R.; Cahais, V.; Rothwell, J.; Herceg, Z. Human Mitochondrial DNA Is Extensively Methylated in a Non-CpG Context. *Nucleic Acids Res.* **2019**. https://doi.org/10.1093/nar/gkz762.
(271) Corsi, S.; Iodice, S.; Shannon, O.; Siervo, M.; Mathers, J.; Bollati, V.; Byun, H.-M. Mitochondrial DNA Methylation Is Associated with Mediterranean Diet Adherence in a Population of Older Adults with Overweight and Obesity. *Proc. Nutr. Soc.* **2020**. https://doi.org/10.1017/s0029665120000439.
(272) Delaney, C.; Garg, S. K.; Yung, R. Analysis of DNA Methylation by Pyrosequencing. In *Methods in Molecular Biology*; 2015. https://doi.org/10.1007/978-1-4939-2963-4_19.
(273) Song, L.; James, S. R.; Kazim, L.; Karpf, A. R. Specific Method for the Determination of Genomic DNA Methylation by Liquid Chromatography-

Electrospray Ionization Tandem Mass Spectrometry. *Anal. Chem.* **2005**. https://doi.org/10.1021/ac0489420.
(274) Neary, J. L.; Carless, M. A. Methylated DNA Immunoprecipitation Sequencing (MeDIP-Seq): Principles and Applications. In *Epigenetics Methods*; 2020. https://doi.org/10.1016/b978-0-12-819414-0.00009-4.
(275) Logsdon, G. A.; Vollger, M. R.; Eichler, E. E. Long-Read Human Genome Sequencing and Its Applications. *Nature Reviews Genetics*. 2020. https://doi.org/10.1038/s41576-020-0236-x.
(276) Jain, M.; Olsen, H. E.; Paten, B.; Akeson, M. The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community. *Genome Biol.* **2016**. https://doi.org/10.1186/s13059-016-1103-0l.
(277) Prall, T. M.; Neumann, E. K.; Karl, J. A.; Shortreed, C. G.; Baker, D. A.; Bussan, H. E.; Wiseman, R. W.; O'Connor, D. H. Consistent Ultra-Long DNA Sequencing with Automated Slow Pipetting. *BMC Genomics* **2021**. https://doi.org/10.1186/s12864-021-07500-w.
(278) King, M. P.; Attardi, G. Human Cells Lacking MtDNA: Repopulation with Exogenous Mitochondria by Complementation. *Science (80-.).* **1989**. https://doi.org/10.1126/science.2814477.
(279) Actis, P.; Hudson, G.; Bury, A. G.; Vincent, A. E.; Turnbull, D. M. Mitochondrial Isolation: When Size Matters. *Wellcome Open Research*. 2020. https://doi.org/10.12688/wellcomeopenres.16300.2.
(280) O'Brown, Z. K.; Boulias, K.; Wang, J.; Wang, S. Y.; O'Brown, N. M.; Hao, Z.; Shibuya, H.; Fady, P. E.; Shi, Y.; He, C.; Megason, S. G.; Liu, T.; Greer, E. L. Sources of Artifact in Measurements of 6mA and 4mC Abundance in Eukaryotic Genomic DNA. *BMC Genomics* **2019**. https://doi.org/10.1186/s12864-019-5754-6.
(281) Wheeler, D.; Lin, J. -H; Chrambach, A. Distinction between Supercoiled and Linear DNA in Transverse Agarose Pore Gradient Gel Electrophoresis. *Electrophoresis* **1992**. https://doi.org/10.1002/elps.1150130185.
(282) Lightowlers, R. N.; Chinnery, P. F.; Turnbull, D. M.; Howell, N.; Turnbuu, D. M. Mammalian Mitochondrial Genetics: Heredity, Heteroplasmy and Disease. *Trends in Genetics*. 1997. https://doi.org/10.1016/S0168-9525(97)01266-3.
(283) Gómez-Durán, A.; Pacheu-Grau, D.; Martínez-Romero, Í.; López-Gallardo, E.; López-Pérez, M. J.; Montoya, J.; Ruiz-Pesini, E. Oxidative Phosphorylation Differences between Mitochondrial DNA Haplogroups Modify the Risk of Leber's Hereditary Optic Neuropathy. *Biochim. Biophys. Acta - Mol. Basis Dis.* **2012**. https://doi.org/10.1016/j.bbadis.2012.04.014.
(284) Stewart, J. B.; Chinnery, P. F. Extreme Heterogeneity of Human Mitochondrial DNA from Organelles to Populations. *Nature Reviews Genetics*. 2020. https://doi.org/10.1038/s41576-020-00284-x.
(285) DiMauro, S.; Schon, E. A.; Carelli, V.; Hirano, M. The Clinical Maze of Mitochondrial Neurology. *Nature Reviews Neurology*. 2013. https://doi.org/10.1038/nrneurol.2013.126.
(286) Holt, I. J.; Harding, A. E.; Morgan-Hughes, J. A. Deletions of Muscle Mitochondrial DNA in Patients with Mitochondrial Myopathies. *Nature* **1988**. https://doi.org/10.1038/331717a0.
(287) Tengan, C. H.; Moraes, C. T. Detection and Analysis of Mitochondrial DNA Deletions by Whole Genome PCR. *Biochem. Mol. Med.* **1996**. https://doi.org/10.1006/bmme.1996.0040.
(288) Fromenty, B.; Manfredi, G.; Sadlock, J.; Zhang, L.; King, M. P.; Schon, E. A.

Efficient and Specific Amplification of Identified Partial Duplications of Human Mitochondrial DNA by Long PCR. *Biochim. Biophys. Acta - Gene Struct. Expr.* **1996**. https://doi.org/10.1016/0167-4781(96)00110-8.
(289) Van Haute, L.; Spits, C.; Geens, M.; Seneca, S.; Sermon, K. Human Embryonic Stem Cells Commonly Display Large Mitochondrial DNA Deletions. *Nature Biotechnology*. 2013. https://doi.org/10.1038/nbt.2473.
(290) Ronaghi, M.; Uhlén, M.; Nyrén, P. A Sequencing Method Based on Real-Time Pyrophosphate. *Science*. 1998. https://doi.org/10.1126/science.281.5375.363.
(291) Andréasson, H.; Asp, A.; Alderborn, A.; Gyllensten, U.; Allen, M. Mitochondrial Sequence Analysis for Forensic Identification Using Pyrosequencing Technology. *Biotechniques* **2002**. https://doi.org/10.2144/02321rr01.
(292) Belmonte, F. R.; Martin, J. L.; Frescura, K.; Damas, J.; Pereira, F.; Tarnopolsky, M. A.; Kaufman, B. A. Digital PCR Methods Improve Detection Sensitivity and Measurement Precision of Low Abundance MtDNA Deletions. *Sci. Rep.* **2016**. https://doi.org/10.1038/srep25186.
(293) Trifunov, S.; Pyle, A.; Valentino, M. L.; Liguori, R.; Yu-Wai-Man, P.; Burté, F.; Duff, J.; Kleinle, S.; Diebold, I.; Rugolo, M.; Horvath, R.; Carelli, V. Clonal Expansion of MtDNA Deletions: Different Disease Models Assessed by Digital Droplet PCR in Single Muscle Cells. *Sci. Rep.* **2018**. https://doi.org/10.1038/s41598-018-30143-z.
(294) Maitra, A.; Cohen, Y.; Gillespie, S. E. D.; Mambo, E.; Fukushima, N.; Hoque, M. O.; Shah, N.; Goggins, M.; Califano, J.; Sidransky, D.; Chakravarti, A. The Human MitoChip: A High-Throughput Sequencing Microarray for Mitochondrial Mutation Detection. *Genome Res.* **2004**. https://doi.org/10.1101/gr.2228504.
(295) van Dijk, E. L.; Jaszczyszyn, Y.; Naquin, D.; Thermes, C. The Third Revolution in Sequencing Technology. *Trends in Genetics*. 2018. https://doi.org/10.1016/j.tig.2018.05.008.
(296) Goodwin, S.; McPherson, J. D.; McCombie, W. R. Coming of Age: Ten Years of next-Generation Sequencing Technologies. *Nature Reviews Genetics*. 2016. https://doi.org/10.1038/nrg.2016.49.
(297) Ewing, B.; Hillier, L. D.; Wendl, M. C.; Green, P. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Res.* **1998**, *8* (3). https://doi.org/10.1101/gr.8.3.175.
(298) Schirmer, M.; Ijaz, U. Z.; D'Amore, R.; Hall, N.; Sloan, W. T.; Quince, C. Insight into Biases and Sequencing Errors for Amplicon Sequencing with the Illumina MiSeq Platform. *Nucleic Acids Res.* **2015**, *43* (6). https://doi.org/10.1093/nar/gku1341.
(299) Kurelac, I.; Lang, M.; Zuntini, R.; Calabrese, C.; Simone, D.; Vicario, S.; Santamaria, M.; Attimonelli, M.; Romeo, G.; Gasparre, G. Searching for a Needle in the Haystack: Comparing Six Methods to Evaluate Heteroplasmy in Difficult Sequence Context. *Biotechnol. Adv.* **2012**, *30* (1). https://doi.org/10.1016/j.biotechadv.2011.06.001.
(300) Shendure, J.; Ji, H. Next-Generation DNA Se1. Shendure J, Ji H. Next-Generation DNA Sequencing. Nat Biotechnol. 2008;26(10):1135–45. Quencing. *Nat. Biotechnol.* **2008**, *26* (10).
(301) Feng, W.; Zhao, S.; Xue, D.; Song, F.; Li, Z.; Chen, D.; He, B.; Hao, Y.; Wang, Y.; Liu, Y. Improving Alignment Accuracy on Homopolymer Regions for Semiconductor-Based Sequencing Technologies. *BMC Genomics* **2016**, *17*. https://doi.org/10.1186/s12864-016-2894-9.
(302) Li, M.; Schönberg, A.; Schaefer, M.; Schroeder, R.; Nasidze, I.; Stoneking, M.

Detecting Heteroplasmy from High-Throughput Sequencing of Complete Human Mitochondrial DNA Genomes. *Am. J. Hum. Genet.* **2010**, *87* (2). https://doi.org/10.1016/j.ajhg.2010.07.014.
(303) Zhang, P.; Samuels, D. C.; Lehmann, B.; Stricker, T.; Pietenpol, J.; Shyr, Y.; Guo, Y. Mitochondria Sequence Mapping Strategies and Practicability of Mitochondria Variant Detection from Exome and RNA Sequencing Data. *Brief. Bioinform.* **2016**, *17* (2). https://doi.org/10.1093/bib/bbv057.
(304) Santibanez-Koref, M.; Griffin, H.; Turnbull, D. M.; Chinnery, P. F.; Herbert, M.; Hudson, G. Assessing Mitochondrial Heteroplasmy Using next Generation Sequencing: A Note of Caution. *Mitochondrion* **2019**, *46*. https://doi.org/10.1016/j.mito.2018.08.003.
(305) Picardi, E.; Pesole, G. Mitochondrial Genomes Gleaned from Human Whole-Exome Sequencing. *Nature Methods*. 2012. https://doi.org/10.1038/nmeth.2029.
(306) Pyle, A.; Hudson, G.; Wilson, I. J.; Coxhead, J.; Smertenko, T.; Herbert, M.; Santibanez-Koref, M.; Chinnery, P. F. Extreme-Depth Re-Sequencing of Mitochondrial DNA Finds No Evidence of Paternal Transmission in Humans. *PLoS Genet.* **2015**, *11* (5). https://doi.org/10.1371/journal.pgen.1005040.
(307) Just, R. S.; Irwin, J. A.; Parson, W. Questioning the Prevalence and Reliability of Human Mitochondrial DNA Heteroplasmy from Massively Parallel Sequencing Data. *Proceedings of the National Academy of Sciences of the United States of America*. 2014. https://doi.org/10.1073/pnas.1413478111.
(308) Chomyn, A.; Lai, S. T.; Shakeley, R.; Bresolin, N.; Scarlato, G.; Attardi, G. Platelet-Mediated Transformation of MtDNA-Less Human Cells: Analysis of Phenotypic Variability among Clones from Normal Individuals--and Complementation Behavior of the TRNALys Mutation Causing Myoclonic Epilepsy and Ragged Red Fibers. *Am. J. Hum. Genet.* **1994**, *54* (6), 966–974.
(309) Shoffner, J. M.; Lott, M. T.; Lezza, A. M. S.; Seibel, P.; Ballinger, S. W.; Wallace, D. C. Myoclonic Epilepsy and Ragged-Red Fiber Disease (MERRF) Is Associated with a Mitochondrial DNA TRNALys Mutation. *Cell* **1990**. https://doi.org/10.1016/0092-8674(90)90059-N.
(310) Flierl, A.; Reichmann, H.; Seibel, P. Pathophysiology of the MELAS 3243 Transition Mutation. *J. Biol. Chem.* **1997**. https://doi.org/10.1074/jbc.272.43.27189.
(311) Chen, J.; Li, X.; Zhong, H.; Meng, Y.; Du, H. Systematic Comparison of Germline Variant Calling Pipelines Cross Multiple Next-Generation Sequencers. *Sci. Rep.* **2019**. https://doi.org/10.1038/s41598-019-45835-3.
(312) Bris, C.; Goudenege, D.; Desquiret-Dumas, V.; Charif, M.; Colin, E.; Bonneau, D.; Amati-Bonneau, P.; Lenaers, G.; Reynier, P.; Procaccio, V. Bioinformatics Tools and Databases to Assess the Pathogenicity of Mitochondrial DNA Variants in the Field of Next Generation Sequencing. *Frontiers in Genetics*. 2018. https://doi.org/10.3389/fgene.2018.00632.
(313) Sahlin, K.; Sipos, B.; James, P. L.; Medvedev, P. Error Correction Enables Use of Oxford Nanopore Technology for Reference-Free Transcriptome Analysis. *Nat. Commun.* **2021**. https://doi.org/10.1038/s41467-020-20340-8.
(314) Spinelli, J. B.; Haigis, M. C. The Multifaceted Contributions of Mitochondria to Cellular Metabolism. *Nature Cell Biology*. 2018. https://doi.org/10.1038/s41556-018-0124-1.
(315) Hao, Z.; Wu, T.; Cui, X.; Zhu, P.; Tan, C.; Dou, X.; Hsu, K. W.; Lin, Y. Te; Peng, P. H.; Zhang, L. S.; Gao, Y.; Hu, L.; Sun, H. L.; Zhu, A.; Liu, J.; Wu, K. J.; He,

C. N6-Deoxyadenosine Methylation in Mammalian Mitochondrial DNA. *Mol. Cell* **2020**. https://doi.org/10.1016/j.molcel.2020.02.018.
(316) Lee, I.; Razaghi, R.; Gilpatrick, T.; Molnar, M.; Gershman, A.; Sadowski, N.; Sedlazeck, F. J.; Hansen, K. D.; Simpson, J. T.; Timp, W. Simultaneous Profiling of Chromatin Accessibility and Methylation on Human Cell Lines with Nanopore Sequencing. *Nat. Methods* **2020**. https://doi.org/10.1038/s41592-020-01000-7.

Appendices

Appendix 1: List of cell lines and tissues used in this study

Cell lines information			
Cell line description	**Codes**	**Source**	**Publication**
Rho 0 cell line	Rho 0	Laboratory of Prof. Patrick F Chinnery	278
Human cybrid cell line - H haplogroup	613H	Laboratory of Prof. Patrick F Chinnery	283
Human cybrid cell line - J haplogroup	128J	Laboratory of Prof. Patrick F Chinnery	283
Human cybrid cell line - J2 haplogroup	135J2	Laboratory of Prof. Patrick F Chinnery	283
Human primary fibroblast cell line - Control	Control 1	Laboratory of Prof. Patrick F Chinnery	N/A
Human primary fibroblast cell line - Control	Control 2	Laboratory of Prof. Patrick F Chinnery	N/A
Human primary fibroblast cell line - MELAS mutation	m.3243A>G (1)	Laboratory of Prof. Patrick F Chinnery	N/A
Human primary fibroblast cell line - MELAS mutation	m.3243A>G (1)	Laboratory of Prof. Patrick F Chinnery	N/A
Human primary fibroblast cell line - MERRF cell line	m.8344A>G	Laboratory of Prof. Patrick F Chinnery	N/A
Tissues information			
Tissue type	**ID Code**	**Source**	**Gender/Age**
Human Liver	TB15-0139	Addenbrooke's Tissue Bank	Male/36 years
Kidney	TB12-1905	Addenbrooke's Tissue Bank	Male/60 years
Human Kidney	TB15-153	Addenbrooke's Tissue Bank	Male/75 years

Heart	TB12-2860	Addenbrooke's Tissue Bank	Male/28 years
Skeletal Muscle	TB15-2606	Addenbrooke's Tissue Bank	Male/56 years
Skeletal Muscle	TB13-1505	Addenbrooke's Tissue Bank	Male/40 years
Skeletal Muscle	TB05-0578	Addenbrooke's Tissue Bank	Male/82 years

Appendix 2: List of primers and probes used in this study

Primers			
Primer name	Forward 5' - 3'	Reverse 5' - 3'	Used for
2F	TGTAAAACGACGGCCAGTTTAAAACTCAAAGGACCTGGC	-	LR-PCR
D1R	-	CAGGAAACAGCTATGACCAGGGTGATAGACCTGTGATC	LR-PCR
MT-ND1	GGGTTCATAGTAGAAGAGCGATGG	ACGCCATAAAACTCTTCACCAAAG	dPCR
RNASE P	AGATTTGGACCTGCGAGCG	GAGCGGCTGTCTCCACAAGT	dPCR
Illumina_primer_1_Fw	CATCCGTATTACTCGCATCAG	-	
Illumina_primer_1_Rev	-	TTGGCTCTCCTTGCAAAGTT	
Illumina_primer_2_Fw	TATCCGCCATCCCATACATT	-	
Illumina_primer_2_Rev	-	AATGTTGAGCCGTAGATGCC	
Probes			
Probe name	Fluorophore	Sequence 5' - 3'	Quencher
MT-ND1	HEX	ACCCGCCACATCTACCATCACCCTC	BHQ_1
RNASE P	FAM	TTCTGACCTGAAGGCTCTGCGCG	BHQ_1

Appendix 3. False positive positions and methylation values.

MtDNA position	Methylation Frequency In NC	Methylation Frequency in Cell Lines and Tissues (average)	Standard deviation in Cell Lines and Tissues
3034	0.035	0.05	0.07
3405	0.135	0.09	0.07
3494	0.05	0.11	0.11
6241	0.172	0.07	0.04
6688	0.037	0.13	0.06
11590	0.033	0.06	0.04
12052	0.11	0.25	0.13
12123	0.07	0.08	0.08
12190	0.051	0.03	0.03
12455	0.03	0.08	0.06
15146	0.035	0.05	0.07
15274	0.052	0.05	0.04
16410	0.065	0.06	0.06

Appendix 4: List and metrics of WGBS samples that passed quality control.

Appendix 4.1: Bias group

GEO Accession (exp)	Cell Line	Tissue	Total Bases	Ave Mito rd	L-strand Aligned Reads	H-strand Aligned Reads	Tot. Aligned Reads	% L-strand	%. H-strand	Covered bp H-strand	Coverage Perc. H-strand	Covered bp L-strand	Coverage Perc. L-strand
GSM983650	N/A	Left Ventricle	45627599814	81.52	1216	12947	14163	8.59	91.41	15171	91.56255658	4107	24.7872533
GSM983652	N/A	Spleen	41570280536	73.63	255	12127	12382	2.06	97.94	14808	89.37171827	1021	6.16210996
GSM983646	N/A	Small Intestine	42456528367	73.11	267	12036	12303	2.17	97.83	14693	88.67765104	1206	7.27865291
GSM1282350	N/A	Lung	24157867100	130.19	9749	12441	22190	43.93	56.07	15145	91.40563703	14628	88.2853522
GSM1120322	N/A	Thymus	18111125700	80.5	4862	8830	13692	35.51	64.49	15061	90.89866618	14366	86.7040859
GSM675544	H1	N/A	2696663675	67.68	1766	11666	13432	13.15	86.85	14687	88.64143883	7808	47.1241475

GEO Accession (exp)	Cell Line	Tissue	Total Bases	Ave Mito rd	L-strand Aligned Reads	H-strand Aligned Reads	Tot. Aligned Reads	% L-strand	%. H-strand	Covered bp H-strand	Coverage Perc. H-strand	Covered bp L-strand	Coverage Perc. L-strand
GSM675545	H1	N/A	8950212800	83.37	1978	12881	14859	13.31	86.69	15120	91.25475285	6798	41.0284266
GSM675546	H1	N/A	9095955300	82.96	1901	12868	14769	12.87	87.13	15105	91.16422234	6601	39.8394592
GSM675543	H1	N/A	2547041448	64.7	1707	11470	13177	12.95	87.05	14649	88.41209488	7254	43.780554
GSM818006	H1	N/A	8508010500	99.09	4472	12947	17419	25.67	74.33	15142	91.38753093	11590	69.9499065
GSM818003	H1	N/A	9371304897	99.53	5648	13052	18700	30.2	69.8	15050	90.83227714	12528	75.6110809
GSM818004	H1	N/A	9352002787	101.17	5921	13018	18939	31.26	68.74	15085	91.043515	12962	78.2304303
GSM818005	H1	N/A	9369542245	92.55	4416	12942	17358	25.44	74.56	15076	90.98919669	11218	67.7047498
GSM706058	iPS DF 6.9	N/A	9360062890	86.26	2888	13026	15914	18.15	81.85	15105	91.16422234	9766	58.9413966

GEO Accession (exp)	Cell Line	Tissue	Total Bases	Ave Mito rd	L-strand Aligned Reads	H-strand Aligned Reads	Tot. Aligned Reads	% L-strand	%. H-strand	Covered bp H-strand	Coverage Perc. H-strand	Covered bp L-strand	Coverage Perc. L-strand
GSM706054	iPS DF 19.11	N/A	9676571741	89.62	2708	13064	15772	17.17	82.83	15151	91.44184924	9442	56.9859376
GSM706053	iPS DF 19.11	N/A	9468370800	84.85	2559	13004	15563	16.44	83.56	15197	91.71947613	9194	55.4891665
GSM706061	H9	N/A	10053272232	85.65	2570	13096	15666	16.4	83.6	15167	91.53841511	7780	46.9551572
GSM429322	H1	N/A	1557514194	44.1	1001	8621	9622	10.4	89.6	14314	86.39024685	4627	27.9256443
GSM429321	H1	N/A	1428272997	39.79	878	8133	9011	9.74	90.26	14204	85.72635645	4190	25.2881888
GSM432687	IMR90	N/A	1830847836	30.39	1191	5497	6688	17.81	82.19	13847	83.57173034	5861	35.3732875
GSM432689	IMR90	N/A	1188700256	24.89	882	4479	5361	16.45	83.55	13835	83.49930593	4565	27.5514515
GSM432688	IMR90	N/A	1823837724	30.3	1180	5304	6484	18.2	81.8	14004	84.519283	5888	35.5362424

GEO Accession (exp)	Cell Line	Tissue	Total Bases	Ave Mito rd	L-strand Aligned Reads	H-strand Aligned Reads	Tot. Aligned Reads	% L-strand	%. H-strand	Covered bp H-strand	Coverage Perc. H-strand	Covered bp L-strand	Coverage Perc. L-strand
GSM432692	IMR90	N/A	1299424482	32.15	859	6107	6966	12.33	87.67	14046	84.77276842	4094	24.7087935
GSM432690	IMR90	N/A	1950385749	32.95	989	6276	7265	13.61	86.39	13918	84.00024141	4699	28.3601907
GSM432691	IMR90	N/A	1993907586	30.81	896	5821	6717	13.34	86.66	13998	84.48307079	4244	25.6140986
GSM432686	H1	N/A	895485171	46.47	2256	7797	10053	22.44	77.56	14388	86.83686402	8512	51.373046
GSM602252	H1	N/A	3611892212	49.84	1505	9821	11326	13.29	86.71	14239	85.9375943	7390	44.601364
GSM602253	H1	N/A	3577328228	53.93	1515	10504	12019	12.61	87.39	14316	86.40231758	7150	43.1528759
GSM602251	H1	N/A	3468065852	56.34	2014	10448	12462	16.16	83.84	14326	86.46267125	8815	53.2017623
GSM602254	H1	N/A	3549481305	57.56	2056	10491	12547	16.39	83.61	14374	86.75236888	8322	50.2263263

GEO Accession (exp)	Cell Line	Tissue	Total Bases	Ave Mito rd	L-strand Aligned Reads	H-strand Aligned Reads	Tot. Aligned Reads	% L-strand	%. H-strand	Covered bp H-strand	Coverage Perc. H-strand	Covered bp L-strand	Coverage Perc. L-strand
GSM602255	H1	N/A	3577130016	57.93	2107	10523	12630	16.68	83.32	14349	86.6014847	8591	51.8498401
GSM602256	H1	N/A	3579398738	58.56	2000	10799	12799	15.63	84.37	14383	86.80668719	8219	49.6046834

Appendix 4.2: Low bias group

GEO Accession (exp)	Tissue Type	Total Bases	Avg. Mito. rd	L-strand Aligned Reads	H-strand Aligned Reads	Tot. Aligned Reads	% L-strand	% H-strand	Covered bp H-strand	Coverage Perc. H-strand	Covered bp L-strand	Coverage Perc. L-strand
GSM1010978	Left Ventricle	37039943200	150.94	12296	13224	25520	48.18	51.82	15171	91.56255658	14778	89.19065725
GSM1010981	Adrenal Gland	19107066700	151.46	12369	13232	25601	48.31	51.69	15183	91.63498099	14750	89.02166697
GSM1282348	Adipose	18727174700	146.92	11943	13170	25113	47.56	52.44	15144	91.39960167	14687	88.64143883
GSM1282349	Gastric	25708659400	149.65	12268	13213	25481	48.15	51.85	15177	91.59876879	14874	89.77005251
GSM1120320	Left Ventricle	20100786000	150.98	12249	13203	25452	48.13	51.87	15162	91.50823828	14731	88.90699499
GSM1120321	Small Intestine	21307522400	148.35	11862	13169	25031	47.39	52.61	15160	91.49616754	14682	88.611262

GEO Accession (exp)	Tissue Type	Total Bases	Avg. Mito. rd	L-strand Aligned Reads	H-strand Aligned Reads	Tot. Aligned Reads	% L-strand	% H-strand	Covered bp H-strand	Coverage Perc. H-strand	Covered bp L-strand	Coverage Perc. L-strand
GSM1282354	Adrenal Gland	25012202200	150.98	12518	13237	25755	48.6	51.4	15178	91.60480415	14933	90.12613918
GSM1282355	Aorta	24257883800	147.65	12027	13186	25213	47.7	52.3	15172	91.56859195	14913	90.00543183
GSM1282356	Oesophagus	23212572200	148.54	12116	13198	25314	47.86	52.14	15176	91.59273342	14916	90.02353793
GSM1282357	Adipose	19237123400	148.31	12155	13197	25352	47.94	52.06	15172	91.56859195	14913	90.00543183
GSM1282358	Gastric	23600960300	147.61	12396	13222	25618	48.39	51.61	15173	91.57462732	14907	89.96921963
GSM1282359	Pancreas	23176515700	149.79	12322	13201	25523	48.28	51.72	15170	91.55652121	14923	90.0657855
GSM1282360	Psoas Muscle	23086039900	151.84	12516	13240	25756	48.59	51.41	15178	91.60480415	14931	90.11406844
GSM1282361	Small Intestine	23080205700	148.34	12079	13187	25266	47.81	52.19	15172	91.56859195	14915	90.01750257
GSM1282362	Spleen	24832228900	144.01	11479	13043	24522	46.81	53.19	15169	91.55048585	14898	89.91490132
GSM1120325	Adrenal Gland	22328957400	148.05	12403	13247	25650	48.35	51.65	15172	91.56859195	14747	89.00356087
GSM1120326	Aorta	24914636600	148.11	12068	13183	25251	47.79	52.21	15172	91.56859195	14721	88.84664132
GSM1120327	Aorta	25386720300	147.95	12070	13198	25268	47.77	52.23	15179	91.61083952	14729	88.89492426
GSM1120328	Aorta	26181703500	147.5	12039	13178	25217	47.74	52.26	15172	91.56859195	14735	88.93113646

GEO Accession (exp)	Tissue Type	Total Bases	Avg. Mito. rd	L-strand Aligned Reads	H-strand Aligned Reads	Tot. Aligned Reads	% L-strand	% H-strand	Covered bp H-strand	Coverage Perc. H-strand	Covered bp L-strand	Coverage Perc. L-strand
GSM1120329	Aorta	24655689100	145.3	11947	13168	25115	47.57	52.43	15167	91.53841511	14721	88.84664132
GSM1120330	Adipose Tissue	24516017400	147.21	12262	13220	25482	48.12	51.88	15162	91.50823828	14730	88.90095962
GSM1120335	Right Atrium	22649536800	151.33	12379	13236	25615	48.33	51.67	15186	91.65308709	14740	88.9613133
GSM1120334	Right Atrium	23404848500	151.07	12388	13230	25618	48.36	51.64	15177	91.59876879	14754	89.04580844

Appendix 5: List and metrics of samples sequenced with ONS in this study.

Sample Type	Haplogroup	Status	Library Preparation	Replicate No.	Total Bases	Avg. Mito. Read Depth	L-strand Aligned Reads	H-strand Aligned Reads	Tot. Aligned Reads	% L-strand	% H-strand	Coverage % H-strand	Coverage % L-strand
Cybrid cancer cell line	H	WT	g-Tube fragmented	1	1191851646	18.07	39	44	83	46.99	53.01	99.85	99.91
Cybrid cancer cell line	H	WT	BamHI - based	1	1748177382	104.55	102	87	189	53.97	46.03	100.00	100.00
Cybrid cancer cell line	H	WT	g-Tube fragmented	2	554638990	12.75	32	20	52	61.54	38.46	92.26	99.81
Cybrid cancer cell line	H	WT	BamHI - based	2	1221290459	55.05	64	58	122	52.46	47.54	99.96	99.99
Cybrid cancer cell line	H	WT	g-Tube fragmented	3	473399623	5.87	18	12	30	60.00	40.00	48.13	60.34
Cybrid cancer cell line	H	WT	BamHI - based	3	742045001	42.97	43	36	79	54.43	45.57	99.99	99.96
Cybrid cancer cell line	H	WT	g-Tube fragmented	4	1255777617	33.15	90	75	165	54.55	45.45	99.98	99.99
Cybrid cancer cell line	H	WT	BamHI - based	4	2245338595	313.12	228	251	479	47.60	52.40	99.99	100.00
Cybrid cancer cell line	H	WT	g-Tube fragmented	5	1068575050	23.52	54	49	103	52.43	47.57	99.86	99.98
Cybrid cancer cell line	H	WT	BamHI - based	5	1477337742	151.88	126	111	237	53.16	46.84	99.99	100.00
Cybrid cancer cell line	J1	WT	g-Tube fragmented	1	888874016	23.98	56	57	113	49.56	50.44	99.93	99.99

Sample Type	Haplogroup	Status	Library Preparation	Replicate No.	Total Bases	Avg. Mito. Read Depth	L-strand Aligned Reads	H-strand Aligned Reads	Tot. Aligned Reads	% L-strand	% H-strand	Coverage % H-strand	Coverage % L-strand
Cybrid cancer cell line	J1	WT	BamHI - based	1	1388724441	191.25	149	119	268	55.60	44.40	100.00	100.00
Cybrid cancer cell line	J1	WT	g-Tube fragmented	2	434993311	11.41	28	20	48	58.33	41.67	89.00	98.23
Cybrid cancer cell line	J1	WT	BamHI - based	2	1066234226	94.13	80	98	178	44.94	55.06	100.00	99.99
Cybrid cancer cell line	J1	WT	g-Tube fragmented	3	391434319	8.43	28	24	52	53.85	46.15	98.74	99.16
Cybrid cancer cell line	J1	WT	BamHI - based	3	907230105	104.77	80	90	170	47.06	52.94	100.00	99.99
Cybrid cancer cell line	J1	WT	g-Tube fragmented	4	645437367	8.04	21	15	36	58.33	41.67	98.56	99.72
Cybrid cancer cell line	J1	WT	BamHI - based	4	1988784182	72.61	74	57	131	56.49	43.51	99.98	100.00
Cybrid cancer cell line	J1	WT	g-Tube fragmented	5	518581582	10.20	25	27	52	48.08	51.92	99.69	99.47
Cybrid cancer cell line	J1	WT	BamHI - based	5	1186935320	219.32	160	133	293	54.61	45.39	100.00	100.00
Cybrid cancer cell line	J2	WT	g-Tube fragmented	1	947373947	26.51	68	51	119	57.14	42.86	99.93	99.98
Cybrid cancer cell line	J2	WT	BamHI - based	1	988759990	279.20	193	190	383	50.39	49.61	100.00	100.00
Cybrid cancer cell line	J2	WT	g-Tube fragmented	2	452157430	15.49	39	35	74	52.70	47.30	99.79	99.87

Sample Type	Haplogroup	Status	Library Preparation	Replicate No.	Total Bases	Avg. Mito. Read Depth	L-strand Aligned Reads	H-strand Aligned Reads	Tot. Aligned Reads	% L-strand	% H-strand	Coverage % H-strand	Coverage % L-strand
Cybrid cancer cell line	J2	WT	BamHI - based	2	1053592851	103.66	77	95	172	44.77	55.23	99.99	99.99
Cybrid cancer cell line	J2	WT	g-Tube fragmented	3	427184866	7.49	19	23	42	45.24	54.76	88.51	86.62
Cybrid cancer cell line	J2	WT	BamHI - based	3	915506169	135.01	107	101	208	51.44	48.56	100.00	99.99
Cybrid cancer cell line	J2	WT	g-Tube fragmented	4	527209479	13.73	35	24	59	59.32	40.68	96.75	99.93
Cybrid cancer cell line	J2	WT	BamHI - based	4	1187786881	125.89	85	100	185	45.95	54.05	100.00	99.99
Cybrid cancer cell line	J2	WT	g-Tube fragmented	5	353962498	8.59	15	15	30	50.00	50.00	46.48	98.00
Cybrid cancer cell line	J2	WT	BamHI - based	5	419509534	47.99	35	39	74	47.30	52.70	99.96	99.99
Primary Fibroblast Cell Line	T	WT	BamHI - based	1	876468989	84.63	60	51	111	54.05	45.95	94.62	94.62
Primary Fibroblast Cell Line	T	WT	BamHI - based	2	767786286	97.13	65	63	128	50.78	49.22	94.62	94.62
Primary Fibroblast Cell Line	T	WT	BamHI - based	3	583892338	81.11	52	47	99	52.53	47.47	94.62	94.62
Primary Fibroblast Cell Line	T	WT	BamHI - based	1	770948386	90.46	59	60	119	49.58	50.42	94.62	94.65
Primary Fibroblast Cell Line	T	WT	BamHI - based	2	832017755	70.64	47	41	88	53.41	46.59	94.62	94.62

Sample Type	Haplogroup	Status	Library Preparation	Replicate No.	Total Bases	Avg. Mito. Read Depth	L-strand Aligned Reads	H-strand Aligned Reads	Tot. Aligned Reads	% L-strand	% H-strand	Coverage % H-strand	Coverage % L-strand
Primary Fibroblast Cell Line	T	WT	BamHI - based	3	788580455	53.87	35	33	68	51.47	48.53	94.59	94.61
Primary Fibroblast Cell Line	J1	8344 A>G MERRF	BamHI - based	1	797209949	84.63	64	68	132	48.48	51.52	99.99	100.00
Primary Fibroblast Cell Line	J1	8344 A>G MERRF	BamHI - based	2	769587574	97.91	82	44	126	65.08	34.92	99.96	100.00
Primary Fibroblast Cell Line	J1	8344 A>G MERRF	BamHI - based	3	329885349	35.78	33	19	52	63.46	36.54	99.69	99.98
Primary Fibroblast Cell Line	U	32434 A>G MELAS_1	BamHI - based	1	661121759	87.00	62	47	109	56.88	43.12	100.00	100.00
Primary Fibroblast Cell Line	U	32434 A>G MELAS_1	BamHI - based	2	582363472	71.37	51	40	91	56.04	43.96	99.97	100.00
Primary Fibroblast Cell Line	U	32434 A>G MELAS_1	BamHI - based	3	458345590	35.32	25	24	49	51.02	48.98	99.90	99.96
Primary Fibroblast Cell Line	U	32434 A>G MELAS_2	BamHI - based	1	928457028	111.44	77	74	151	50.99	49.01	100.00	100.00
Primary Fibroblast Cell Line	U	32434 A>G MELAS_2	BamHI - based	2	590347810	87.83	50	64	114	43.86	56.14	99.99	100.00
Primary Fibroblast Cell Line	U	32434 A>G MELAS_2	BamHI - based	3	414140512	77.32	55	48	103	53.40	46.60	99.97	99.99
Hum. Don. TB12 - 1905 Kidney	W	WT	BamHI - based	1	N/A	145.69	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Hum. Don. TB12 -	W	WT	BamHI - based	1	N/A	565.98	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Sample Type	Haplogroup	Status	Library Preparation	Replicate No.	Total Bases	Avg. Mito. Read Depth	L-strand Aligned Reads	H-strand Aligned Reads	Tot. Aligned Reads	% L-strand	% H-strand	Coverage % H-strand	Coverage % L-strand
Heart Normal													
Hum. Don. TB15 - 139 Liver	W	WT	BamHI - based	1	N/A	110.81	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Hum. Don. TB15 - 153 Kidney	U	WT	BamHI - based	1	N/A	274.93	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Hum. Don. TB15 - 2606 Muscle	J	WT	BamHI - based	1	N/A	214.72	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Hum. Don. TB13 - 1505 Muscle	J	WT	BamHI - based	1	N/A	637.77	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Hum. Don. TB05 - 578 Muscle	K	WT	BamHI - based	1	N/A	210.32	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Appendix 6: Illumina Miseq and ONS sequencing metrics

Appendix 6.1: general sequencing metrics

Samples			Illumina Miseq				ONS sequencing			
Cell line	**Nanopore replicates**	**Nanopore replicates (Barcodes)**	**Average Read Depth**	**Coverage**	**Haplogroup prediction (MToolBox)**	**Haplogroup prediction (Haplogrep2)**	**Average Read Depth**	**Coverage**	**Haplogroup prediction (MToolBox)**	**Haplogroup prediction (Haplogrep2)**
128J	128J_1	exp_16_barcode10	4260.21x	100%	J1c1g	J1c1g	147.11x	100%	J1c1g	J1c1c
	128J_2	exp_17_barcode04					77.15x	100%	J1c1g	J1c1g
	128J_3	exp_19_barcode04					75.86x	100%	J1c1g	J1c1g
	128J_4	exp_24_barcode04					55.57x	100%	J1c1g	J1c1c
	128J_5	exp_25_barcode10					162.57x	100%	J1c1g	J1c1g

Samples			Illumina Miseq				ONS sequencing			
Cell line	**Nanopore replicates**	**Nanopore replicates (Barcodes)**	**Average Read Depth**	**Coverage**	**Haplogroup prediction (MToolBox)**	**Haplogroup prediction (Haplogrep2)**	**Average Read Depth**	**Coverage**	**Haplogroup prediction (MToolBox)**	**Haplogroup prediction (Haplogrep2)**
135J2	135 J2_1	exp_16_barcode12	4382.19x	100%	J2b1a1	J2b1a1	218.21x	100%	J2b1a1	J2b1a1
	136 J2_1	exp_17_barcode06					78.54x	100%	J2b1a1	J2b1a1
	137 J2_1	exp_19_barcode06					99.55x	100%	J2b1a1	J2b1a1
	138 J2_1	exp_24_barcode06					97.45x	100%	J2b1a1	J2b1a1
	139 J2_1	exp_25_barcode12					35.49x	100%	J2b1a1	H2a2a1
613H	613 H_1	exp_16_barcode08	5558.61x	100%	H1	H1	78.46x	100%	H1	H1
	614 H_2	exp_17_barcode02					40.73x	100%	H1	H2a2a1

Samples			Illumina Miseq				ONS sequencing			
Cell line	**Nanopore replicates**	**Nanopore replicates (Barcodes)**	**Average Read Depth**	**Coverage**	**Haplogroup prediction (MToolBox)**	**Haplogroup prediction (Haplogrep2)**	**Average Read Depth**	**Coverage**	**Haplogroup prediction (MToolBox)**	**Haplogroup prediction (Haplogrep2)**
	615 H_3	exp_19_barcode02					31.13x	100%	H1	H2a2a1
	616 H_4	exp_24_barcode02					240.58x	100%	H1	H1
	617 H_5	exp_25_barcode08					109.65x	100%	H1	H1
Control 1	Control Fibroblast 1_1	exp_26_barcode01	2967.37x	100%	T2b4;T2b_150	T2b4	64.88x	94.6	T2b4;T2b_150	T2b+150
	Control Fibroblast 1_2	exp_27_barcode07					73.08x	94.6	T2b4;T2b_150	T2b4
	Control Fibroblast 1_3	exp_28_barcode01					60.28x	94.6	T2b4;T2b_150	T2b+150
Control 2	Control Fibroblast 2_1	exp_26_barcode02	3707.17x	100%	T2b4;T2b_150	T2b4	67.78x	94.6	T2b4;T2b_150	T2b4

Samples			Illumina Miseq				ONS sequencing			
Cell line	**Nanopore replicates**	**Nanopore replicates (Barcodes)**	**Average Read Depth**	**Coverage**	**Haplogroup prediction (MToolBox)**	**Haplogroup prediction (Haplogrep2)**	**Average Read Depth**	**Coverage**	**Haplogroup prediction (MToolBox)**	**Haplogroup prediction (Haplogrep2)**
	Control Fibroblast 2_2	exp_27_barcode08					53.78x	94.6	T2b4;T2b_150	T2b+150
	Control Fibroblast 2_3	exp_28_barcode02					40.72x	94.6	T2b4;T2b_150	H2a2a1
MERFF	MERFF_1	exp_26_barcode04	321.73x	100%	J1c1a	J1c1a	78.21x	100%	J1c1a	J1c1a
	MERFF_2	exp_27_barcode09					70.36x	100%	J1c1a	J1c1a
	MERFF_3	exp_28_barcode03					26.53x	100%	J1c1a	H2a2a1
MELAS1	MELAS1_1	exp_26_barcode03	638.59x	100%	U5a1f1a1	U5a1f1a1	64.73x	100%	U5a1f1a1	U5a1f1a1
	MELAS1_2	exp_27_barcode10					51.36x	100%	U5a1f1a1	U5a1+@16192

Samples			Illumina Miseq				ONS sequencing			
Cell line	**Nanopore replicates**	**Nanopore replicates (Barcodes)**	**Average Read Depth**	**Coverage**	**Haplogroup prediction (MToolBox)**	**Haplogroup prediction (Haplogrep2)**	**Average Read Depth**	**Coverage**	**Haplogroup prediction (MToolBox)**	**Haplogroup prediction (Haplogrep2)**
	MELAS1_3	exp_28_barcode04					25.84x	100%	U5a1f1a1	H2a2a1
MELAS2	MELAS2_1	exp_26_barcode05	317.95x	100%	U5a1f1a1	U5a1f1a1	79.81x	100%	U5a1f1a1	U5a1f1
	MELAS2_2	exp_27_barcode11					63.25x	100%	U5a1f1a1	U5a1f1
	MELAS2_3	exp_28_barcode05					56.67x	100%	U5a1f1a1	U5a1f1a1

Appendix 7: ONS differential methylation analysis results

Appendix 7.1: Differential methylation on rCRS

Haplogroup results									
H Vs J1									
Position on rCRS	**mu1**	**mu2**	**diff**	**diff.se**	**stat**	**phi1**	**phi2**	**pval**	**fdr**
10400	0.026435	0.25447	-0.2280349	0.036470486	-6.252588281	0.013788412	0.012063583	4.04E-10	8.76041E-08
16128	6.38E-05	0.322185	-0.322121	0.056611601	-5.690018222	0.016607978	0.01246307	1.27E-08	1.83764E-06
16360	0.075261	0.57707	-0.5018093	0.058191544	-8.623406487	0.016923701	0.012531377	6.5E-18	2.82062E-15
H Vs J2									
Position on rCRS	**mu1**	**mu2**	**diff**	**diff.se**	**stat**	**phi1**	**phi2**	**pval**	**fdr**

10400	0.026435	0.251048	-0.2246129	0.03504282	-6.409669451	0.013788412	0.010181706	1.46E-10	3.16463E-08
16128	6.38E-05	0.489763	-0.4896992	0.065747701	-7.448156842	0.016607978	0.012973657	9.47E-14	4.10795E-11
Fibroblasts results									
Ctrl Vs MELAS_1									
Position on rCRS	**mu1**	**mu2**	**diff**	**diff.se**	**stat**	**phi1**	**phi2**	**pval**	**fdr**
4919	0.234492	0.000174	0.2343177	0.039650482	5.909579748	0.017249728	0.024446826	3.43E-09	4.87034E-07
15925	0.662011	0.000265	0.6617455	0.057357352	11.53723961	0.025294671	0.024439598	8.56E-31	3.6478E-28
16128	0.494665	0.013232	0.4814334	0.069263345	6.950767964	0.018512512	0.023168465	3.63E-12	7.73835E-10
Ctrl Vs MELAS80									
Position on rCRS	**mu1**	**mu2**	**diff**	**diff.se**	**stat**	**phi1**	**phi2**	**pval**	**fdr**

4919	0.234492	0.009559	0.2249335	0.04125243	5.452611008	0.017249728	0.017344436	4.96E-08	7.04825E-06
9195	0.05212	0.457422	-0.4053022	0.083713121	-4.841560682	0.016986231	0.02017991	1.29E-06	0.000137197
15925	0.662011	0.015789	0.646222	0.05966286	10.83122792	0.025294671	0.017700563	2.45E-27	1.04306E-24
16128	0.494665	0.026701	0.4679644	0.070009576	6.684290564	0.018512512	0.017039688	2.32E-11	4.94258E-09
Ctrl Vs MELAS_2									
Position on rCRS	**mu1**	**mu2**	**diff**	**diff.se**	**stat**	**phi1**	**phi2**	**pval**	**fdr**
4919	0.234492	0.009171	0.225321	0.041287834	5.457322384	0.017249728	0.015596294	4.83E-08	1.02957E-05
10400	0.021798	0.308006	-0.2862075	0.061239854	-4.673549723	0.015921648	0.015446666	2.96E-06	0.000420374
15925	0.662011	0.000208	0.6618025	0.057331895	11.54335708	0.025294671	0.017256234	7.98E-31	3.39738E-28

Appendix 7.2: Differential methylation on consensus sequences

Haplogroup results									
H Vs J1									
Position on consensus	**mu1**	**mu2**	**diff**	**diff.se**	**stat**	**phi1**	**phi2**	**pval**	**fdr**
10400	7.10854E-05	8.10391E-05	-9.95367E-06	0.003162278	-0.003147627	0.016609169	0.013105445	0.99748856	0.9977468
16128	0.07523248	0.02423998	0.0509925	0.027189258	1.87546505	0.016933241	0.009616188	0.06072877	0.9977468
16360	0.02642609	0.01570084	0.01072525	0.016949165	0.632789498	0.013725968	0.010472008	0.52687112	0.9977468
H Vs J2									
Position on consensus	**mu1**	**mu2**	**diff**	**diff.se**	**stat**	**phi1**	**phi2**	**pval**	**fdr**
1869	7.10854E-05	0.01949721	-0.01942612	0.014811539	-1.311553279	0.016609169	0.011946191	0.18967091	0.9997225
12710	0.02642609	0.007133338	0.01929276	0.015389317	1.253646059	0.013725968	0.010946	0.20997068	0.9997225
Fibroblasts results									
Ctrl Vs MELAS_1									
Position on consensus	**mu1**	**mu2**	**diff**	**diff.se**	**stat**	**phi1**	**phi2**	**pval**	**fdr**

4919	0.00761739	0.000268358	0.007349032	0.008975105	0.818824052	0.019434329	0.024223657	0.412886806	0.983842179
15925	0.000165539	0.013210537	-0.013044997	0.016262329	-0.802160477	0.020084262	0.023015874	0.422460135	0.983842179
16128	0.012458564	0.000176506	0.012282058	0.009560461	1.284672186	0.015883316	0.02422999	0.198906862	0.983842179
Ctrl Vs MELAS_2									
Position on consensus	**mu1**	**mu2**	**diff**	**diff.se**	**stat**	**phi1**	**phi2**	**pval**	**fdr**
4919	0.00761739	0.015760526	-0.008143137	0.019191991	-0.424298693	0.019434329	0.016962779	0.671347992	0.997244709
9195	0.000165539	0.026653968	-0.026488428	0.019277807	-1.374037428	0.020084262	0.016362443	0.169430063	0.997244709
15925	0.012458564	0.009536896	0.002921668	0.015082354	0.193714346	0.015883316	0.016636457	0.846399558	0.997244709
16128	0.052060058	0.124146741	-0.072086683	0.067860766	-1.062273345	0.016915818	0.019606073	0.288111612	0.997244709
Ctrl Vs MERRF									
Position on consensus	**mu1**	**mu2**	**diff**	**diff.se**	**stat**	**phi1**	**phi2**	**pval**	**fdr**
4919	0.00761739	0.000197718	0.007419671	0.008782773	0.844798271	0.019434329	0.018308715	0.39822348	0.999910706
10400	0.012458564	0.009134981	0.003323583	0.015328868	0.21681858	0.015883316	0.016592072	0.828349731	0.999910706
15925	0.0217785	0.02547097	-0.0036925	0.02527861	-0.1460709	0.01586039	0.01568607	0.8838654	0.99991071