

Long-Molecule Assessment of Ribosomal DNA and RNA

A thesis submitted in partial fulfillment of the
requirements of the Degree of Doctor of Philosophy

Zakaryya Ahmad

Primary Supervisor: Professor Vardhman Rakyan

Secondary Supervisor: Dr Matt Silver

Centre for Genomics and Child Health,
The Blizard Institute,
Barts and the London School of Medicine and Dentistry
Queen Mary University of London

Statement of originality

I, Zakaryya Ahmad, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Zakaryya Ahmad

30th June 2022

Publications and Collaborations

Publications

Francisco Rodriguez-Algarra, Robert A. E. Seaborne, Amy F. Danson, Selin Yildizoglu, Harunori Yoshikawa, Pui Pik Law, **Zakaryya Ahmad**, Victoria A. Maudsley, Ama Brew, Nadine Holmes, Mateus Ochôa, Alan Hodgkinson, Sarah J. Marzi, Madapura M. Pradeepa, Matthew Loose, Michelle L. Holland & Vardhman K. Rakyan (2022). Genetic variation at mouse and human ribosomal DNA influences associated epigenetic states. *Genome Biol.* 23, 54

Collaborations

Dr Francisco Rodriguez-Algarra (QMUL) wrote the scripts for analysing nanopore RNA sequencing data. The culture of mouse embryonic stem cells, the generation of embryoid bodies and the validation of embryoid differentiation, as well as metaphase chromosome preparation and staining was conducted together with Selin Yildizoglu (QMUL). Dr Michelle Holland (King's College London) derived the mouse embryonic fibroblasts used for nanopore sequencing and molecular combing. Dr Robert A. E. Seaborne (QMUL) assisted with the culture of human lymphoblastoid cell lines. Prof. Vardhman Rakyan (Blizard Institute) and Dr Amy Danson (QMUL) sourced and provided the mouse tissue. All other work in this thesis was conducted by Zakaryya Ahmad.

Abstract

The genes encoding ribosomal RNA and their transcriptional products are essential for life, however, remain poorly understood. Even with the advent of long-range sequencing methodologies, rDNA loci are difficult to study and remain obscure, prompting the consideration of alternative methods to probing this critical region of the genome. The research outlined in this thesis utilises molecular combing, a fibre stretching technique, to isolate DNA molecules measuring more than 5 Mbp in length. The capture of DNA molecules of this size should assist in exploring the architecture of entire rDNA clusters at the single-molecule level. Combining molecular combing with SNP targeting probes, this study aims to distinguish and assess the arrangement of rDNA promoter variants which have been shown to exhibit dramatically different environmental sensitivity. Additionally, through the application of Oxford Nanopore Technologies direct RNA sequencing, the work here has demonstrated the capture of near full-length rRNA primary transcripts, which will allow for assessing post-transcriptional modification across the length of multiple coding subunits within a single molecule, for the first time. Furthermore, an exploration of RNA modification profiles across sample types representative of different developmental stages has been conducted. This study predicts many sites to be differentially modified across these different developmental conditions, several of which are known to be important for, if not crucial in ribosome biogenesis and function. The work outlined in this thesis provides a framework for future studies to conduct long-molecule, genetic, and epitranscriptome profiling of this vital region of the genome, and its dynamic response to a changing environment.

Table of Content

Statement of originality	2
Publications and Collaborations	3
Abstract	4
Table of Contents	5
List of Figures	11
List of Tables	13
List of Abbreviations	14
Acknowledgements	17
1 Introduction	18
1.1 Gene-environment interactions	18
1.1.1 Developmental programming and the predictive adaptive response	18
1.1.2 The thrifty phenotype hypothesis	19
1.1.3 Studying Gene-environment interactions in controlled settings	20
1.2 The ribosome - A potential substrate for gene-environment interactions	21
1.2.1 Structure of the ribosome	21
1.2.2 The rDNA loci	22
1.3 Ribosomal DNA epigenetics and expression	25
1.3.1 Epigenetic regulation of rDNA expression	25
1.3.2 rDNA loci methylation	25
1.3.3 rDNA promoter methylation	26
1.3.4 Ribosomal RNA processing	26
1.4 Genome instability and position effect	28
1.4.1 The rDNA landscape	30

1.4.2	Sequencing ribosomal DNA arrays	30
1.5	The epitranscriptome and rRNA modifications	31
1.5.1	RNA modifications	31
1.5.2	Ribosomal RNA modifications	32
1.5.3	rRNA modifications and ribosome structural integrity	33
1.5.4	rRNA modifications on ribosome function	33
1.5.5	The implication of rRNA modifications in disease states	34
1.6	Ribosome heterogeneity	35
1.6.1	Ribosome heterogeneity and the ribosome filter hypothesis	35
1.6.2	Ribosomal proteins- a source of ribosome heterogeneity	35
1.6.3	Ribosomal DNA- a source of ribosome heterogeneity	36
1.6.4	Ribosomal RNA modifications- a source of ribosome heterogeneity	36
1.7	Exploring ribosomal heterogeneity in C57BL/6J mice	37
1.7.1	rDNA promoter variants	37
1.7.2	Variant-specific methylation dynamics	38
1.7.3	Distinct rDNA haplotypes	40
1.8	Sequencing methodologies	41
1.8.1	Short read sequencing methods	41
1.8.2	Nanopore sequencing	42
1.9	Gaps in the knowledge	44
1.9.1	Large-scale arrangement of rDNA and genetic variation within clusters	45
1.9.2	Epitranscriptomic profiles of rRNA and haplotype-specific modifications	45
1.10	Thesis aims	46
1.11	Thesis structure	46
2	Materials and methods	48
2.1	Cell culture techniques	48
2.1.1	Mouse Embryonic Fibroblasts (MEFs)	48
2.1.2	Mouse Embryonic Stem Cells (MESCs)	48

2.1.3	Human Lymphoblastoid Cell Line (LCLs)	49
2.1.4	Human Embryonic Kidney Cells (HEK-293)	49
2.2	DNA and RNA techniques	49
2.2.1	Agilent Bioanalyzer	49
2.2.2	Gel electrophoresis	49
2.2.3	Assessing nucleic acid purity and concentration	50
2.2.4	Quantitative polymerase chain reaction	50
2.3	Generation of control cell lines for SNP-specific probe testing	51
2.3.1	C57BL/6J promoter variant isolation	51
2.3.2	Lentivirus generation	52
2.3.3	Lentiviral transduction	53
2.4	SNP-specific localisation of C57BL/6J rDNA promoter variants	54
2.4.1	Lentivirus generation	54
2.4.2	Lentiviral transduction	55
2.4.3	FACS sorting and affinity purification of SpdCas9-3xGFP	55
2.5	Molecular combing	55
2.5.1	Molecular combing machine assembly	55
2.5.2	Silane slide preparation	57
2.5.3	High Molecular Weight (HMW) DNA extraction	58
2.5.4	Ultralong DNA preparation in agarose plugs	58
2.5.5	Molecular combing	59
2.6	DNA labelling techniques	60
2.6.1	FISH Probe Synthesis	60
2.6.2	Metaphase chromosome spread preparation	61
2.6.3	DNA Fluorescence <i>in situ</i> Hybridisation	62
2.7	Mouse Embryoid Body formation	63
2.7.1	Mouse Embryonic Stem Cell differentiation	63
2.7.2	Embryoid Body germ layer validation with qPCR and immunofluorescence	63
2.8	Ribosomal RNA processing inhibition	64

2.8.1	Drug treatment	64
2.8.2	Propidium Iodide staining and FACS cell cycle analysis	65
2.8.3	qPCR validation of rRNA precursor processing	65
2.9	Oxford Nanopore Sequencing methods	66
2.9.1	Oxford Nanopore Sequencing Technology	66
2.9.2	Total and nuclear RNA extraction	66
2.9.3	Pre-processing of RNA for Nanopore library preparation	67
2.9.4	Nanopore cDNA and Direct RNA library preparation	67
2.9.5	Nanopore cDNA and Direct RNA Sequencing data analysis	68
2.9.6	Ribosomal RNA modification calling with Nanocompare	68
3	Single-Molecule Analysis of rDNA Promoter Variants	69
3.1	Introduction	69
3.1.1	Aims	69
3.1.2	Molecular Combing	70
3.1.3	Evolution of DNA combing methods	71
3.1.4	Combing Ultra-long DNA molecules	73
3.1.5	Factors influencing DNA combing	74
3.1.6	Probing genomic loci	76
3.2	Materials and Methods	79
3.3	Results	80
3.3.1	Manufactured silanised slides outperform silanised slides produced in-house	80
3.3.2	Effect of pH on DNA combing	82
3.3.3	Kaykov et al. and Genomic Vision DNA extraction protocols result in the inadequate combing of MEF DNA	86
3.3.4	Adapting Kaykov et al. protocol for combing MEF DNA	87
3.3.5	Adapted combing protocol allows for isolation of DNA molecules measuring >5 Megabases in length	90
3.3.6	Visualising genomic loci	93
3.3.7	Generation of control cell lines for testing specificity of SNP-CLING probes	96
3.3.8	Generation of SNP-CLING probes for distinguishing rDNA promoter variants	100

3.4 Discussion	102
3.4.1 DNA combing optimisation	102
3.4.2 Probing genomic loci	105
3.4.3 Conclusions	107
4 Long-Read Sequencing Analysis of Ribosomal RNA Modifications	108
4.1 Introduction	108
4.1.1 Aims	108
4.1.2 Detecting and mapping RNA modifications	108
4.1.3 Nanopore RNA sequencing	110
4.1.4 Mapping RNA modifications using Nanopore data	111
4.1.5 Nanocompore	112
4.1.6 Predicting ribosomal RNA modification on full-length transcripts	112
4.2 Methods and Materials	114
4.3 Results	115
4.3.1 Nanopore cDNA sequencing of <i>in vitro</i> poly-adenylated ribosomal RNA	115
4.3.2 Nuclear RNA extracts permit increased capture of rRNA processing intermediates compared to cellular extracts	121
4.3.3 Size selection-based enrichment of rRNA processing intermediates	123
4.3.4 5-Fluorouracil exposure of MEF cells hinders rRNA processing	126
4.3.5 ONT Direct RNA sequencing of ribosomal RNA	133
4.3.6 <i>In vitro</i> 5' capping of rRNA increases the 5' coverage of Nanopore DRS reads	137
4.3.7 ONT DRS data set generation	139
4.3.8 ONT DRS allows for the capture of near full-length rRNA primary transcripts	142
4.3.9 Expression of rRNA haplotypes is cell type-specific	144
4.3.10 Analysing rRNA modifications with Nanocompore	147
4.3.11 Differential modification profiles are observed between MEF and MESC rRNA	148
4.4 Discussion	155
4.4.1 Nanopore RNA sequencing of ribosomal RNA	155
4.4.2 Cell-specific expression of ribosomal RNA alleles	157

4.4.3	The cell-specific differential modification of rRNA	158
4.5	Conclusions	160
5	Discussion and Conclusions	161
5.1	Research summaries	161
5.1.1	Aim 1	161
5.1.2	Aim 2	162
5.2	Research Challenges	163
5.3	Future experiments and directions	164
5.4	Conclusion	165

List of Figures

Figure 1.1 The structure of eukaryotic ribosome

Figure 1.2 Location of rDNA loci on human and mouse chromosomes

Figure 1.3 Arrangement of individual units within a rDNA cluster

Figure 1.4 Pre-ribosomal rRNA processing in mouse cells

Figure 1.5 C57BL/6J rDNA promoter variants

Figure 1.6 Key results from Holland et al., 2016

Figure 1.7 Long-range haplotype characterization of 45S rDNA in the C57BL/6J strain

Figure 1.8 Oxford Nanopore Technologies Nanopore sequencing principles

Figure 1.9 Nanopore flow cell comparison

Figure 1.10 MinION flow cell sequencing capacity progression

Figure 1.11 ONT ultra long DNA sequencing kit (SQK-ULK001) read length distribution histogram

Figure 2.1 Motorised molecular combing machine

Figure 3.1 Schematic outlining the phases of dynamic DNA combing

Figure 3.2 Schematic outlining the agarose plug method to obtaining ultra-long combed DNA

Figure 3.3 Common molecular combing issues and possible causes.

Figure 3.4 Structural analysis of human rDNA gene array with DNA combing

Figure 3.5 Allele specific labelling with SNP-CLING

Figure 3.6 DNA combing efficiency on inhouse and manufactured silanised slides

Figure 3.7 Effect of pH on DNA combing

Figure 3.8 DNA combing using pre-prepared MES pH 5.5 buffer.

Figure 3.9 Comparing Kaykov et al. and Genomic Vision's DNA extraction protocols

Figure 3.10 Optimising Kaykov et al. protocol parameters

Figure 3.11 Optimising cell number and DNA density

Figure 3.12 Optimised combing protocol allows for the isolation of ultra-long DNA fibres.

Figure 3.13 Histogram displaying size distribution of combed DNA fibres

Figure 3.14 Composite image of longest combed DNA fibre

Figure 3.15 Synthesis of DNA FISH probes

Figure 3.16 Visualisation of rDNA in fixed MEF nuclei and chromosome spreads

Figure 3.17 Isolation of rDNA genetic variants.

Figure 3.18 Generation of stable HEK-293T cell lines with integrated rDNA promoter sequences

Figure 3.19 SNP-CLING probes generation

Figure 4.1 Current genome wide detection methods used to identify RNA modifications

Figure 4.2 ONT cDNA Sequencing pilot study sample preparation

List of Figures

Figure 4.3 Comparison of ONT cDNA sequencing pilot runs

Figure 4.4 Nanopore cDNA sequencing of C57BL/6J muscle rRNA

Figure 4.5 Nanopore cDNA sequencing of MEF total RNA, read analysis

Figure 4.6 Nanopore cDNA sequencing of MEF nuclear RNA, read analysis.

Figure 4.7 Size analysis of size-selected MEF nuclear cDNA using TapeStation 2200

Figure 4.8 Nanopore cDNA sequencing of MEF, size selected, nuclear RNA

Figure 4.9 Cell cycle analysis of 5-FU treated MEF cells

Figure 4.10 Validation of 5-FU inhibition of rRNA processing in MEF cells with qPCR

Figure 4.11 Nanopore cDNA sequencing of 5FU treated, MEF, size selected, nuclear RNA.

Figure 4.12 Comparison of ONT cDNA sequencing runs

Figure 4.13 Nanopore DRS of MEF total RNA

Figure 4.14 Nanopore DRS of MEF nuclear RNA

Figure 4.15 Nanopore DRS of MEF, size selected, nuclear RNA

Figure 4.16 Nanopore DRS of MEF, *in vitro* 5' capped nuclear RNA

Figure 4.17 Optimised sample pre-processing protocol for nanopore sequencing of ribosomal RNA

Figure 4.18 EB formation and differentiation validation

Figure 4.19 Longest pre-rRNA reads captured with ONT DRS

Figure 4.20 Coverage of haplotype specific SNPs is unequal

Figure 4.21 rDNA haplotype-identifying alleles are differentially expressed

Figure 4.21 Establishing baseline Nanocompore modification-calling 'noise'

Figure 4.22 Nanocompore differential modification-calling in MESC and EB DRS datasets

Figure 4.23 Nanocompore differential modification-calling in MEF and MESC DRS datasets

Figure 4.24 Nanocompore differential modification calling in MEF and EB DRS datasets

Figure 4.25 Nanocompore differential modification calling in MESC and liver DRS datasets

List of tables

Table 1.1 Primers used for PCR amplification of C57BL/6J promoter sequence

Table 2.1 Primers used for PCR amplification of C57BL/6J rDNA templates for FISH probe synthesis

Table 2.2 Embryoid Body germ layer marker qPCR validation primer sequences

Table 2.3 qPCR primers used for rRNA precursor processing assessment

Table 3.1 Comparison of 3rd generation sequencing and optical mapping platforms

Table 4.1 ONT DRS data sets generated on MinION and PromethION devices

List of Abbreviations

5-FU	5-Fluorouracil
AFM	Atomic Force Microscopy
BAC	Bacterial artificial chromosome
BML	Bloom Syndrome Protein
C57BL/6J	C57 black 6 inbred strain
CAST	Wild inbred strain of mouse model
cDNA	Complementary DNA
CHIR	Highly selective inhibitor of glycogen synthase kinase 3
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DAPI	4',6-diamidino-2-phenylindole
DC	Decoding centre
dCas9	Catalytically inactive Cas9 nuclease
DJ	Distal junction
DNA	Deoxyribonucleic Acid
DNMT1	DNA (cytosine-5)-methyltransferase 1
DRS	Direct RNA Sequencing
dUTP	2'-Deoxyuridine, 5'-Triphosphate
EB	Embryoid body
ERK	Extracellular signal-regulated kinase
ETS	External transcribed spacer
FACS	Fluorescence activated cell sorting
FDR	False discovery rate
FISH	Fluorescence <i>in situ</i> hybridisation
GATA4	GATA Binding Protein 4
GFP	Green fluorescent protein
GMM	Gaussian mixture model
GRCm38	Mouse genome reference

gRNA	guideRNA
HEK	Human embryonic kidney cells
HMW	High molecular weight
IF	Immunofluorescence
ITS	Internal transcribed spacer
LC-MS/MS	Liquid Chromatography with tandem mass spectrometry
LCL	Lymphoblastoid cell line
Lenti-A	Lentiviral particles used for the integration of A variant rDNA promoter sequence
Lenti-C	Lentiviral particles used for the integration of C variant rDNA promoter sequence
LIF	Leukemia inhibitory factor
LMP	Low melting point
m5C	5-methylcytosine
m62A	N6,N6-dimethyladenosine
m6A	Methyladenosine
m7G	7-Methylguanosine
Mbp	Mega base pairs
MEF	Mouse embryonic fibroblasts
MESC	Mouse embryonic stem cells
NGS	Next generation sequencing
ONT	Oxford nanopore technologies
PAM	Protospacer adjacent motif
PEG	Polyethylene glycol
PI	Propidium iodide
PJ	Proximal junction
pLenti-puro-A	pLenti-puro vector with A variant sequence
pLJM1-EGFP-C	pLJM1-EGFP vector with C variant sequence
POL1	RNA polymerase 1
PTC	Peptidyl transfer centre
rDNA	Ribosomal DNA
RFP	Red fluorescent protein

RNA	Ribonucleic acid
RRBS	Reduced representation bisulphite sequencing
rRNA	Ribosomal RNA
snoRNA	Small nucleolar RNA
SNP	Single nucleotide polymorphism
SNP-CLING	Spatiotemporal allele organization by allele-specific CRISPR live-cell imaging
SNV	Single nucleotide variation
SPRI	Solid-phase reversible immobilization
tRNA	Transfer RNA
UBF	Upstream binding factor
UCE	Upstream control element
WGS	Whole genome sequencing
WT	Wild type

Acknowledgements

Firstly, I would like to thank my supervisor Vardhman Rakyen for his guidance and support in all aspects of this thesis. I am endlessly grateful to Francisco Rodriguez-Algarra for his invaluable part in all things computational and for his unwavering patience. Thank you to the rest of the Rakyen Lab for their contributions, particularly Selin Yildizoglu and Robert Seaborn. The members of the QMUL Epigenetic Hub have been a constant source of knowledge and expertise. A special thanks to Andrea Cerase, Hemanth Tummala and Giuseppe Trigiante for selflessly dedicating their time to provide invaluable guidance and Pradeepa Madapura and Miguel Branco for their suggestions and advice.

1 Introduction

1.1 Gene-environment interactions

An increasing amount of evidence supports the idea that both nature (i.e. genetics) and nurture (i.e. the environment), interact to shape an organism's development. It has become a fact that an individual's underlying genetics dictate phenotypic outcomes, depending on the environmental influences to which they are exposed (Ottman, 1996). This phenomenon termed the gene-environment interaction has become a field of increasing interest in recent years, owing to its relevance to many human diseases (Ober and Vercelli, 2011). These interactions are considered to be critical in 'complex' diseases such as cardiovascular disease, diabetes, ageing, certain cancers, and even susceptibility to various infections. Additionally, disease 'triggering' environmental stimuli are often clustered within families, prompting the exploration of how complex diseases can be inherited (Mcgrath *et al.*, 2013). Substantial efforts are being made to dissect the interplay between disease-associated alleles and environmental cues, in a bid to predict an individual's disease predisposition and serve the interest of public health.

1.1.1 Developmental programming and the predictive adaptive response

Gene-environment interactions occurring during prenatal or the early developmental window can cause permanent changes to the anatomy, physiology, and behaviour of an organism, with critical impacts on health, welfare, and development (Barker *et al.*, 1993; Woodall *et al.*, 1996; Sutton, Centanni and Butler, 2010). These interactions can also have transgenerational effects, altering phenotypes in not only the individual but also their future offspring (Carone *et al.*, 2010; Zimmer *et al.*, 2017). The mechanisms underlying this 'developmental programming' are thought to have evolved to prime an organism for survival by assessing the environment during gestation, predicting the postnatal conditions, and expressing as a predictive adaptive response (Gluckman and Hanson, 2004). Many examples of this are seen in the animal world, for instance, in response to maternal exposure to predatory pheromones, *Daphnia cucullate*, a type of water flea, develops greater protective 'helmets' which protect them from predation (Weiss, Leimann and Tollrian, 2015). Another example is the coat thickness determination in vole pups, determined by the maternal experience of photoperiod length (Lee and Zucker, 1988). These predictive adaptations do not confer an immediate advantage to the developing organism; however, if the postnatal environment is correctly predicted then the adaptations are designed to be advantageous in later life. If, however, the environment during later life is incorrectly predicted, this results in a mismatch (Godfrey *et al.*, 2007). In such cases, the predictive response becomes ineffectual and may even pose a threat to the organism's health

(Fortier, Ponton and Gilbert, 1995; Nederhof and Schmidt, 2012). In the context of human health, maternal exposure to nutrient insults, teratogens such as pollutants, drugs, and alcohol, as well as altered hormonal balance resulting from maternal health conditions, can lead to increased susceptibility to disease in offspring later in life (Rice *et al.*, 2010). This is evident in several human epidemiological studies, some of which underpin the thrifty phenotype hypothesis.

1.1.2 The thrifty phenotype hypothesis

The thrifty phenotype, proposed by Barker and colleagues posits that low birth weight is strongly associated with chronic conditions such as coronary heart disease, stroke, diabetes, and hypertension (Barker and Osmond, 1986; Barker *et al.*, 1989, 1993, 2009). The increased susceptibility is said to result from maternal undernutrition, which prompts certain adaptations by the developing foetus as it grows in an environment limited in nutrients. It is thought that metabolic adaptations and altered resource management lead to reduced birth weight, which acts to assist in its survival (Barker *et al.*, 1993). Evolutionarily, such adaptations are considered to aid in the development of an unborn child, such that it will be prepared for survival in an environment in which resources are likely to be short. However, if exposed to markedly improved nutrition in post-natal life, the individual over-compensates, leading to rapid weight gain and an increased risk of the associated pathologies (Remacle, Bieswal and Reusens, 2004; Barker and Thornburg, 2013; Lynch, Chan and Drake, 2017).

A well-documented example that supports the thrifty phenotype hypothesis is the Dutch famine of 1944-45, during which Nazi troops blocked the provision of food to the West Netherlands. During this period of food restriction, the average calorie consumption for an individual was limited to 400-800 per day (Schulz, 2010; Ekamper *et al.*, 2017). Longitudinal studies found that children of women who were pregnant during the famine had a significantly increased incidence of metabolic diseases such as obesity, diabetes, and cardiovascular disease when compared to the rest of the population (Painter, (Roseboom *et al.*, 2000; Roseboom, de Rooij and Painter, 2006). These findings were unexpected considering that the children were born after food restrictions had ended and therefore, themselves had access to “good” nutrition throughout their life. Additionally, it was discovered that the time frame of exposure impacted disease outcomes (Schulz, 2010). For instance, individuals subjected to the famine during early gestation were more likely to suffer from obesity and breast cancer than expected, whilst these outcomes were not observed if the exposure was during late gestation. Exposure to famine at any stage of gestation, however, still resulted in an increased risk of later life glucose intolerance (Roseboom *et al.*, 2006). Such historical incidents have helped illuminate the possible disease outcomes of various environmental stressors. However, understanding the underlying mechanisms remains a challenge. This is in part due to the high variability and

unpredictability of such events, as well as the fact that the precise timing and nature of these exposures cannot be controlled.

1.1.3 Studying Gene-environment interactions in controlled settings

There are however opportunities to better understand the mechanism governing gene-environment interactions in human cohorts, in more controlled environments. A study by Erikson et al. (2017), explored the influence of intergenerational in utero parental energy and nutrient restriction on offspring growth in the rural Gambia (Eriksen *et al.*, 2017). Though some parts of the country head toward urbanisation, many groups, such as the Mandinka people live largely detached, relying mainly on subsistence farming to survive. This west African country undergoes drastic, cyclic seasonal weather fluctuations experienced as either a prolonged hot and dry season, or a short, wet season. Due to this, the food supply is inconsistent throughout the year, meaning that there is a nutrient-restricted, 'hungry' season once the majority of the harvested crop is exhausted. The seasonal nutritional restriction experienced naturally in a repeating annual pattern provides a unique opportunity to explore the consequences of nutritional restriction *in utero* within a large population. From the analysis of comprehensive antenatal and child growth data collected over several decades, it emerged that Infants born during the hungry season, a time marked by weight loss, increased labour, and risk of infection, had lower birth weights compared with infants born in the harvest season. This was observed alongside higher mortality from infectious diseases in young adulthood. Additionally, it was found that mothers exposed to nutrient restriction in the latter part of their fetal development gave birth to smaller babies than unexposed mothers, even if the child itself was not exposed, suggesting a transgenerational impact.

Even so, the heterogeneity of human populations is vast, making it exceedingly difficult to identify gene-environment interactions that dictate long-term responses to early life exposure in such a diverse background (Ober and Vercelli, 2011). For this reason, inbred strains of animal models, given the known degree of genetic variation, play a vital role in understanding the molecular mechanisms governing phenotypic changes in response to the early life environment (Reynolds *et al.*, 2010). The experimental setup can be further simplified using cell types derived from these organisms. The ease of collection, rapid growth kinetics, and large-scale expansion to perform multiple, high-throughput experiments, permits the reconstitution of *in vivo/in utero* assessments in an *in vitro* format (Hirsch and Schildknecht, 2019). Such approaches provide the possibility to yield new and important fundamental insights into how gene-environment interactions shape the epigenome and phenotypic outcomes and identify key regulators of this process

1.2 The ribosome - A potential substrate for gene-environment interactions

The ribosome is a vital molecular machine, responsible for the synthesis of proteins in all living cells. With 2,000-10,000 ribosomes produced every minute in eukaryotic cells, ribosome biogenesis stands as the most energy consumptive process in an actively proliferating cell (Warner, 1999). Considering this, it is not difficult to imagine that ribosome biosynthesis and associated processes may be targeted for gene-environment interactions. Ribosomes have long been overlooked in gene-environment studies, often being assumed to lack functional specificity in their role in protein manufacture. However, increasing evidence suggests that ribosome biogenesis may act as a key molecular regulator in determining phenotypes in response to early life insults (Moss *et al.*, 2007; Holland *et al.*, 2016; Berres *et al.*, 2017). In comparison to many thoroughly characterised regions of the genome, some loci involved in ribosome biogenesis remain obscure and poorly understood. This is largely due to the repetitive and long-spanning structure of certain ribosomal loci which are incompatible with current sequencing and computational technologies (Treangen and Salzberg, 2012). To fully understand how ribosome biogenesis is implicated in phenotypic determination in response to a changing environment, it is critical to dissect ribosomal genomic architecture and the many points of regulation that modulate its biosynthesis.

1.2.1 Structure of the ribosome

The ribosome itself is a complex assembly of proteins and ribosomal RNA (rRNA) which coalesce to form the distinct small and large subunits (Moss *et al.*, 2007; Babler and Hurt, 2019). Eukaryotic ribosomes (80S ribosomes) are composed of a small 40S subunit and a large 60S subunit (**Figure 1.1**). Here, (S) refers to the Svedberg unit used to measure the sedimentation coefficient, denoting the rate at which particles sediment when centrifuged, a reflection of particle size. The 40S subunit contains the 18S rRNA and 33 proteins (Yusupov *et al.*, 2001). The 60S subunits differ between species, being made up of around 46-50 proteins and three rRNAs: 5S, 5.8S, and 25S (Moss *et al.*, 2007; Klinge and Woolford, 2019). The translational activity of the ribosome occurs within two main functional sites, each responsible for a different phase of protein synthesis. These sites are the decoding centre (DC), and the peptidyl transferase centre (PTC). The DC is the site at which an mRNA codon is matched with the incoming aminoacyl-tRNA anticodon and is located within the small subunit (Ogle *et al.*, 2001; Terenin *et al.*, 2005). The PTC is located within the large subunit, specifically in a cleft within the subunit interface, and serves as the ribosomes' primary catalytic centre. It is the site at which peptide bond formation occurs between amino acids in a growing peptide chain, as well as the site at which hydrolysis of peptidyl-tRNA occurs, leading to the release of the newly synthesized

Introduction peptide (Polacek and Mankin, 2005; Beringer, 2008). Additionally, the large subunit contains three distinct tRNA binding sites, termed the “A”, “P” and “E” sites. The A-site (aminoacyl), is the first binding site in the ribosome, the P-site (for peptidyl), is the second, whilst the E-site (exit), is the third (Schmeing, Moore and Steitz, 2003). Though the protein subunits act as vital scaffolds, orientating mRNA transcripts and tRNA, the essential catalytic abilities of the ribosome are conferred by the RNA components (Lafontaine and Tollervey, 2001).

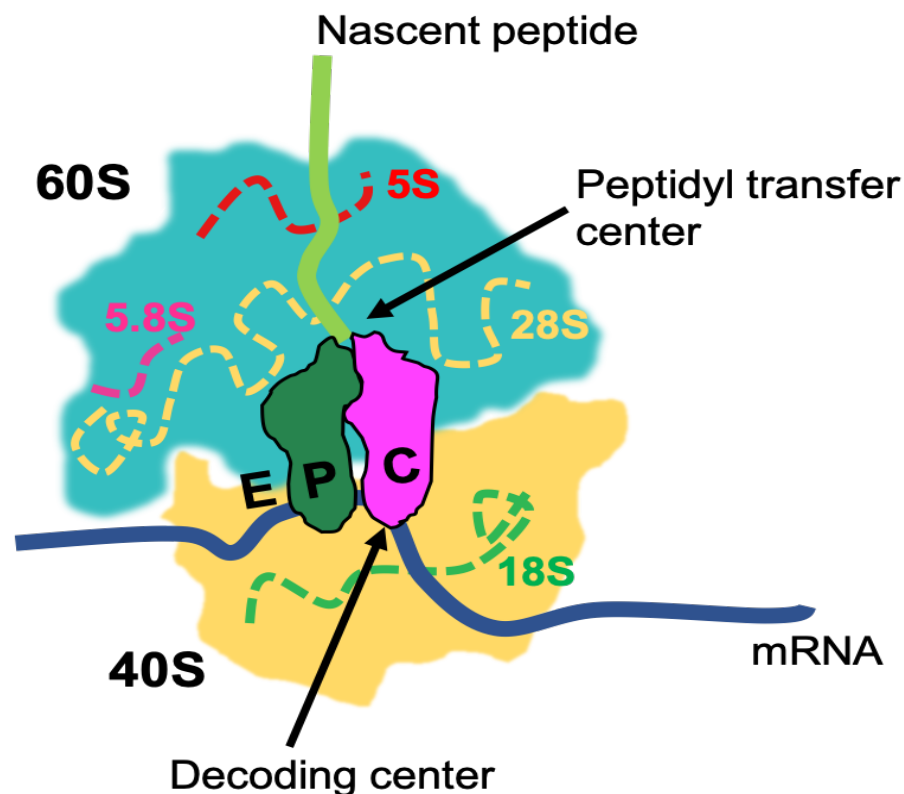


Figure 1.1 The structure of eukaryotic ribosome. A diagram displaying a eukaryotic ribosome during protein synthesis, with key elements labelled. The large subunit (60S) is shown in blue, containing rRNAs 5S (red), 5.8S (pink), 28S (yellow), whilst the small subunit (40S) is shown in yellow containing the 18S rRNA (green). Key functional sites, the peptidyl transfer centre (PTC) and decoding centre (DC) are labelled along with the 3 tRNA binding sites, A, P and E. Sites P and A are specifically shown with a bound tRNA at each site.

1.2.2 The rDNA loci

Ribosomal RNAs are encoded by ribosomal DNA (rDNA), the most ubiquitously transcribed genes in the eukaryotic genome, the transcription of which accounts for 90% of the total cellular RNA content (Warner, 1999). Whilst the gene encoding the 5S rRNA is genetically isolated (Steffensen, Duffey and Prenskey, 1974), the 18S, 5.8S, and 28S rRNAs are all encoded by a single transcriptional unit, the 45S rDNA (Srivastava and Schlessinger, 1991). In eukaryotes, 45S rDNA is dispersed across multiple chromosomes found as clusters composed of tandemly arranged repeating units, with hundreds of 45S rDNA copies found within a single mouse and human genome (Gibbons *et al.*, 2015). **Figure 1.2**

shows the arrangement of 45S rDNA on both human and mouse chromosomes. In humans, the 45S rDNA is positioned on the p arm of the acrocentric chromosomes, closely above the centromere, and makes up a large proportion of the p-arm, with clusters flanked on either side by heterochromatic proximal and distal junctions (Eickbush and Eickbush, 2007). In humans, the 45S rDNA occupies chromosomes 13, 14, 15, 21, and 22 with chromosome 1 carrying 5S rDNA (Henderson *et al.*, 1972; Worton *et al.*, 1988). In inbred mouse strain, 45S rDNA chromosomes are generally thought to be chromosomes 12, 15, 18, and 19 with chromosome 8 carrying 5S rDNA (Kurihara *et al.*, 1994; Lebofsky and Bensimon, 2003) however, these loci are known to differ between specific strains.

The 45S rDNA genes occur in clusters, composed of many repeating units, arranged in tandem (Wellauer and Dawid, 1977) (**Figure 1.3**). Each unit consists of a coding region that encodes the 3 rRNAs (18S, 5.8S, and 28S), as well as an intergenic spacer that separates units within a tandem array (Richard, Kerrest and Dujon, 2008). Within the coding unit, the 3 rRNAs are separated by internal transcribed spacer (ITS) sequences, with the concatenated gene array flanked by 5' and 3' external transcribed spacer (ETS) sequences (Wellauer and Dawid, 1977). The rRNA components appear to be highly conserved in evolution, whilst significant divergence is observed in both transcribed and non-transcribed spacer regions (Richard *et al.*, 2008).

The size of rDNA clusters can greatly vary with some clusters containing 100s of copies. Considering a single copy of the 45S rDNA unit measures ~43 kbp and ~45 kbp in humans and mice respectively, entire rDNA clusters may often span upwards of a few Megabases (Gonzalez and Sylvester, 1995; Grozdanov, Georgiev and Karagyzov, 2003). Whilst rDNA repeats within a cluster are generally arranged in a head-to-tail fashion, units are also observed in a variety of unconventional conformations such as palindromic or inverted arrays (Caburet *et al.*, 2005). The average rDNA content of a human diploid cell is thought to be between 300- 600 copies (Stults *et al.*, 2008), however, significant copy number variation is observed amongst members of the same species and even between different tissues of the same organism (Wang *et al.*, 2017). Furthermore, the copy number is not static, instead, clusters can shrink and expand with copy number loss and amplification being associated with physiological aberrances such as tumor growth and disease (Xu *et al.*, 2017). Altogether, rDNA is an extremely unpredictable and dynamic region of the genome that displays high levels of inter-and intra-individual variability and remains to be fully understood.

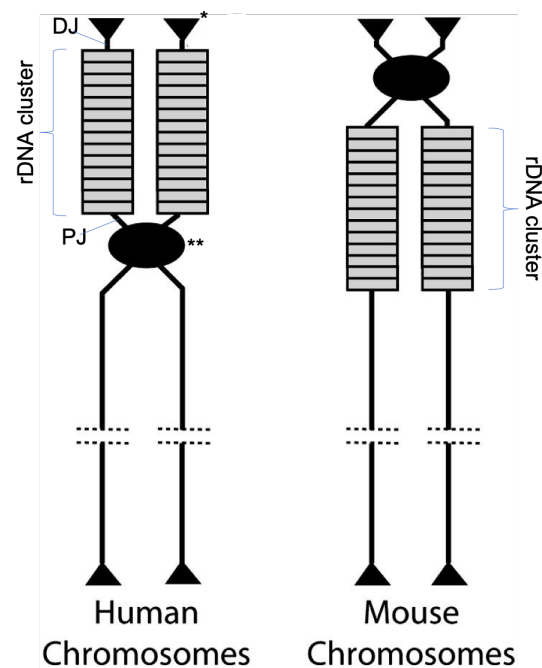


Figure 1.2 Location of rDNA loci on human and mouse chromosomes. Chromosome depicted are representative of all non-homologous chromosomes containing rDNA loci for both human and mouse. Individual units within the labelled cluster represent individual rDNA units. Each chromosome is depicted post-replication to show two sister chromatids still attached at the centromere. The position of both the distal junction (DJ) and proximal Junction (PJ) are shown. *Black triangles represent telomeres. **Black ovals represent centromeres
Image adapted from Eickbush & Eickbush, 2007 Figure 3.

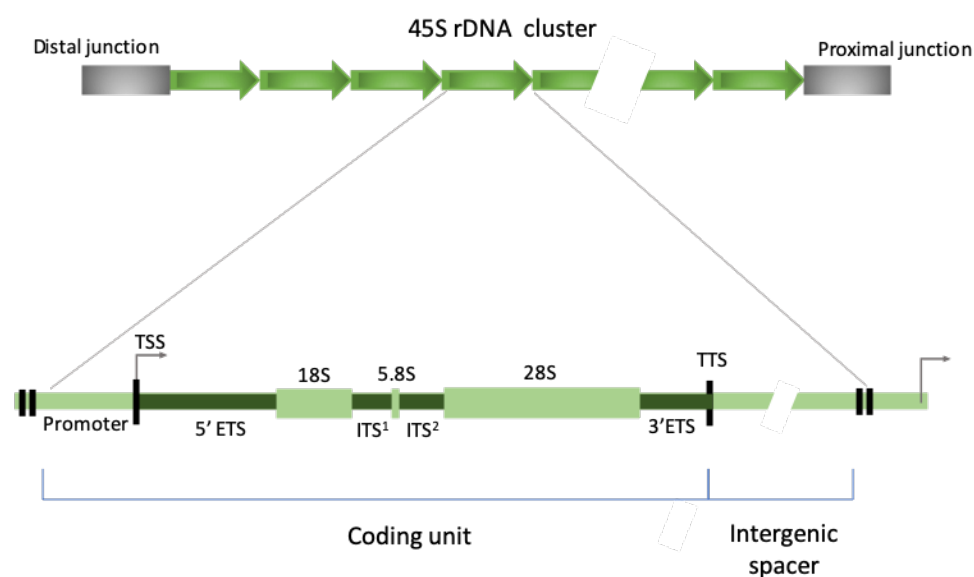


Figure 1.3 Arrangement of individual units within a rDNA cluster. A representative rDNA cluster containing tandemly arranged rDNA repeat units. A single rDNA unit is composed of a coding unit and an adjacent non-transcribed intergenic spacer. **Each** coding unit contains the genes encoding the 18S, 5.8S, 28S rRNAs, separated and flanked by transcribed spacers (5'ETS, ITS¹, ITS², 3'ETS) and is under the control of a single promoter.

1.3 Ribosomal DNA epigenetics and expression

1.3.1 Epigenetic regulation of rDNA expression

Considering the extraordinary abundance of ribosomes and the potential energy expense arising from the unrestrained expression, it is unsurprising that ribosome biogenesis is tightly regulated. Over 200 genes are involved in ribosome biosynthesis regulation in yeast (Hall, Wade and Struhl, 2006), with rRNA transcription being considered a key regulatory target (Warner *et al.*, 1999). Various studies have shown that not all rDNA copies are available for active transcription at any given time (Santoro and Grummt, 2001; Grummt and Ladurner, 2008). Rather, a study by Conconi *et al.*, (1989), assessing the accessibility of rDNA through psoralen crosslinking, found that no more than 50% of the copies are active at any one time (Conconi *et al.*, 1989). Meanwhile, other studies have demonstrated that rDNA copy activity positively correlates with total copy number (Rodriguez-Algarra *et al.*, 2022). Active copies of rDNA are characterised by an 'open', accessible euchromatin, featuring epigenetic modifications such as dimethylation of histone H3 at lysine 4 (H3K4me2), acetylation of histone H4 and DNA hypomethylation. Contrastingly, silenced copies of rDNA are characterised by 'closed' heterochromatin which features repressive epigenetic modification such as trimethylation of H3K9, H4K20, and H3K27, histone H4 hypoacetylation, and DNA hypermethylation (Moss *et al.*, 2007; Grummt and Pikaard, 2003).

1.3.2 rDNA loci methylation

Several experiments have shown the methylation of rDNA to be vital for normal cellular (Sinclair and Guarente, 1997; Gagnon-Kugler *et al.*, 2009) showed that the knockout of DNA methyltransferase 1 (DNMT1), an enzyme responsible for maintaining DNA methylation patterns, resulted in the loss of rDNA gene silencing and a significant increase in 45 rRNA synthesis. This was observed alongside a disruption to rRNA processing and nucleolar morphology and the accumulation of episomal rDNA which is specifically linked to ageing in yeast (Sinclair and Guarente, 1997). Additionally, the genomic instability of rDNA is a key feature of chromosomal aberration in tumours (Agrawal and Ganley, 2018). For instance, Individuals with Bloom syndrome, a rare autosomal disorder resulting from a mutation in the Bloom syndrome protein (BML), a RecQ helicase involved in the suppression of homologues recombination in the nucleolus (Blasiak *et al.*, 2020), have increased rDNA cluster instability leading to increased incidence of cancer (Schawalder *et al.*, 2003). This suggests that rDNA stability and its transcriptional regulation via DNA methylation is critical for the normal functioning of the cell and plays an important role in metabolic homeostasis (Grummt and Längst, 2013).

1.3.3 rDNA promoter methylation

Whilst rDNA can be methylated across the entire locus (Holland *et al.*, 2016), transcription can effectively be regulated by the methylation of select CpG's (Shiao *et al.*, 2011). Concerning CpG sites in the locality of the Pol 1 promoter, humans have at least 25 which are methylated in a mosaic pattern (Ghoshal *et al.*, 2004). In mice, however, the methylation of a single CpG site at position -133 in the upstream control element (UCE) of the rDNA promoter, has been shown to successfully hinder the binding of the POL 1 basal transcription factor, upstream binding factor (UBF) (Santoro and Grummt, 2001). The role of UBF is critical in defining the expression of rDNA gene copies, with Sanij *et al.* (2008) demonstrating that UBF1, a subtype of UBF, prevents linker histone H1-induced assembly of heterochromatin and regulates the open chromatin structure of active rDNA genes (Sanij *et al.*, 2008). Furthermore, directing UBF1 to heterochromatin results in extensive chromatin de-condensation whilst a decrease in UBF1 levels has been correlated with a diminished pool of active rDNA. A study by (Stefanovsky *et al.*, 2001) challenged the established view that rRNA transcription responds to changes in cellular metabolic demand in an indirect manner. Rather, the study was able to demonstrate the existence of a direct link between growth factor signalling and ribosome biogenesis, implicating the phosphorylation of UBF via the ERK pathway (Upstream binding factor (UBF) as a positive regulator of rDNA transcription. Additionally, UBF-mediated dysregulation of rDNA transcription is also implicated in specific pathologies. Treacher Collins syndrome, a disease resulting in severe craniofacial disfigurement arises from mutations in the *TCOF1* gene encoding the protein Treacle, which directly interacts with UBF to promote rDNA transcription (Valdez *et al.*, 2004). Hence, the evidence suggests that Methylation at CpG -133 in the mouse rDNA is enough to effectively prevent the transcription of the affected rDNA gene unit, complications that may result in adverse phenotypic outcomes.

1.3.4 Ribosomal RNA processing

Ribosomal RNAs undergo both co- and post-transcriptional processing, involving a combination of cleavage, folding, and nucleotide modification steps before the formation of a mature, functional ribosome. In eukaryotes, the 18S, 5.8S, and 28S rRNAs are synthesised as one large precursor molecule, the primary transcript (47S), which is processed to liberate the constituent rRNAs. The processing of the primary transcript is a concerted effort by a range of nuclear and cytoplasmic, endo- and exonucleases that can occur via multiple pathways to yield many short-lived intermediate pre-rRNAs. The process is thoroughly described in eukaryotic cells, with studies in yeast, murine and human cells providing valuable insights into the multiple processing pathways, the major

intermediates formed as well as the identification of key enzymes (Eichler and Craig, 1994; Ansel *et al.*, 2008; Preti *et al.*, 2013; Henras *et al.*, 2015).

Based on the abundance of certain processing intermediates, a major pathway for post-transcriptional processing of rRNA precursor molecule 47S has been proposed for murine cells. The process is described in **Figure 1.4** alongside an alternative pathway (reviewed in detail by Henras *et al.*, (2015)). In murine cells, maturation of the primary rRNA transcript initiates in the nucleolus with partial cleavage of the 47S 5'ETS and complete removal of the 3'ETS to yield the 45S pre-rRNA (Bowman *et al.* 1983; Eichler and Craig, 1994). From here, processing pathways can diverge with the main pathway involving cleavage within the ITS¹ to yield the 34S rRNA (containing the 18S rRNA) and 32S rRNA (containing the 5.8S-28S rRNAs) (Wang and Pestov, 2011). Maturation of the 18S rRNA from the 34S rRNA is achieved by multiple endonucleolytic cleavages of the 5'ETS and the sequential removal of the remaining ITS¹ by both endo- and exonucleases (Kent, Lapik and Pestov, 2009). An immature 18S is exported out into the cytoplasm where the 3'-5' exonucleolytic removal of ITS yields its mature form (Preti *et al.*, 2013). Alongside this, 32S processing is initiated with the 5'-3' exonucleolytic removal of ITS¹ followed by the endonucleolytic cleavage within ITS², yielding the 12S and 28.5S pre-rRNAs containing the 5.8S and 28S respectively. The remnants of ITS¹ are removed via 3'-5' exonucleolytic digestion of 12S pre-rRNA and 5'-3' exonucleolytic digestion of 28.5S pre-rRNA, yielding an immature 5.8S and mature 28S (Wang *et al.*, 2014). Both rRNAs are transported out into the cytoplasm where a final 3'-5' exonucleolytic removal of ITS² yields the mature 5.8S rRNA. The alternative pathway diverges from the major pathway after the generation of the 45S pre-rRNA and continues with the endonucleolytic removal of the 5' ETS and subsequent cleavage of ITS¹. A series of endo- and exonucleolytic steps then lead to the formation of the mature 18S, 5.8S, and 28S rRNAs (Carron *et al.*, 2011).

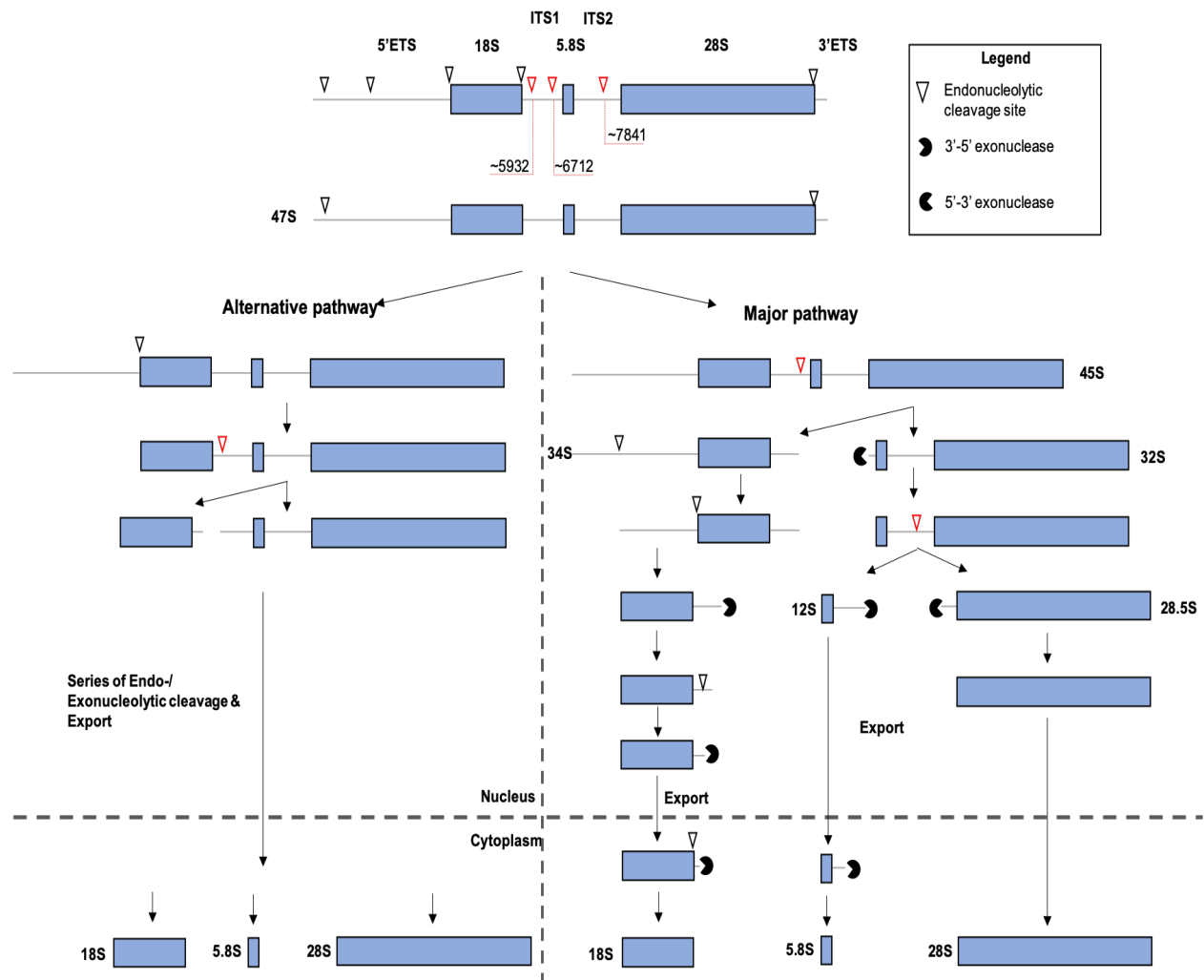


Figure 1.4 Pre-ribosomal rRNA processing in mouse cells. Schematic of mouse rRNA primary transcript containing 3 rRNA molecules 18S, 5.8S and 28S, flanked by 5'- and 3'- external transcribed spacers (ETS) and separated by 2 internal transcribed spacers (ITS¹ & ITS²). Position of 8 endonucleolytic cleavage sites are marked with arrow heads with the 3 sites occurring within ITS sequences marked red and their approximated sequence positions stated. Two processing pathways are depicted, the major pathway is shown with all endo- and exonucleolytic cleavage steps leading to the maturation of rRNAs 18S, 5.8S and 28S as well as with all major intermediate pre-rRNAs formed. The alternative pathway is presented partially with focus on the initial endonucleolytic cleavage step within ITS¹. The mapping of cleavage sites presented here are reviewed by Mullineux and Lafontaine (2012). The numbering of the nucleotides refers to GenBank sequence BK000964.3. Schematic is adapted from Henras et al., 2015 Figure 2.

1.4 Genome instability and position effect

It is widely documented that the spatial positioning of a gene and its local chromosomal environment can impact its expression (Kleinjan and Van Heyningen, 1998; Chen and Zhang, 2016). Correct gene expression is broadly determined by 3 factors: (i) the promoter element (iii) enhancer/silencer elements and (iii) the local chromatin environment. Chromosomal rearrangement events, including deletions, duplications, inversions, and translocations can cause detrimental spatial re-organisation of

genes leading to transcriptional and regulatory failures (Harewood and Fraser, 2014; Chen and Zhang, 2016; Spielmann, Lupiáñez and Mundlos, 2018). A deleterious change in the expression of a gene resulting from its repositioning relative to its normal chromosomal environment is a phenomenon termed the 'position effect' (Sturtevant, 1925). Events that alter the local chromatin environment are of key importance as promoters and enhancers can only function in permissive chromatin environments. Chromatin organisation can be crudely divided into the open and accessible transcriptionally active euchromatic state or a tightly condensed transcriptionally inactive heterochromatic state. Furthermore, heterochromatin has the unique ability to spread and serve as a substrate for the recruitment of a variety of regulatory proteins and in turn affect the expression of neighbouring genes in both a sequence and region-specific manner (Talbert and Henikoff, 2006; Grewal and Elgin, 2007). The process is dynamic and heterochromatin domains have been documented to change their stability in response to environmental cues (reviewed by Wang et al 2016).

The phenotypic impact of position effect has been well documented in *Drosophila* and is demonstrated in the classic example of position effect variegation (PEV) concerning abnormal eye colour (Muller, 1930). In such cases, the *white* gene which is normally expressed uniformly in each cell of the adult *Drosophila* eye resulting in the red-eye phenotype is abnormally translocated next to pericentromeric heterochromatin resulting in its repression (Wakimoto and Hearn, 1990; Henikoff, Jackson and Talbert, 1995). As a result, the *white* gene is only expressed in some cells of the eye leading to a red-white mosaic-coloured eye phenotype (Hermann J. Muller, 1930). Such dramatic phenotypic variegation is also observed in mice. Female mice with translocation of an autosomal coat colour gene into the X chromosome experience its repression due to heterochromatinisation and display abnormal coat colouring when compared to wild-type individuals, (Russel and Bangham, 1961; Cattanaach, 1966)

Some loci, such as rDNA arrays are inherently unstable and prone to genomic rearrangement owing to the tandem arrangement of repeats, the high rates of transcription, and the difficulty in replication that is associated with repetitive sequences. Whilst many obstacles are encountered during the replication of repetitive sequences, transcription remains one of the most mutagenic processes, particularly in highly transcribed genes (Kim and Jinks-Robertson, 2012). Transcription/translation conflicts can often lead to double-stranded breaks that are repaired by one of many homologous recombination-dependent repair pathways (Kobayashi *et al.*, 2004). Due to the presence of many near-identical rDNA repeats within an array that can serve as a template for these repair processes,

rDNA is highly prone to unequal recombination-mediated spatial changes (Kobayashi, 2011). With the highly variable and dynamic spatial arrangement of rDNA, it is important to fully dissect the rDNA genetic landscape, specifically concerning chromosome location, the size of arrays, and their respective variants compositions, but also the local environment contributing to the differential environmental sensitivity of genetic variants.

1.4.1 The rDNA landscape

Owing to the instability of rDNA, the number of rDNA clusters, loci, and copy number is known to differ among species, populations, and even individuals (Jhanwar, Prenskey and Chaganti, 1981; Kopp, Mayr and Schleger, 1988). Though individual units are largely thought to be ordered as tandem repeats in a head-to-tail fashion, various other orientations such as palindromic repeats have also been observed in humans (Caburet *et al.* 2005). Though most studies concur that in most inbred strains of mice including C57BL/6J, 45S rDNA containing chromosomes are likely 12, 15, 16, 18, 19 (Kurihara *et al.*, 1994; Matsuda *et al.*, 1994), some studies cite less whilst others proclaim more. Using silver staining of active NORs, Dev *et al.* recognised 4 45S carrying chromosomes, identifying them as chromosomes 12, 15, 16, and 18 (Dev *et al.*, 1977), whilst chromosome 11 was touted as an additional carrier by (Gibbons *et al.* 2015). Additionally, genetic mapping of C57BL/6J using southern blot analysis has also shown 18S ribosomal RNA-related loci on chromosomes 5, 6, 9, 12, 17, 18, 19, and X (Rowe *et al.* 1996). These varied and somewhat conflicting observations may be true biological differences or simply due to the experimental approach used. For instance, silver staining is commonly used to identify rDNA chromosomes, however, this approach only identifies actively transcribing NORs potentially leading to underrepresentation (Goodpasture and Bloom, 1975). On the other hand, FISH allows for direct visualisation of rDNA sequences but probes may be prone to non-specific binding leading to false positives (Cui, Shu and Li, 2016). Considering the high variability and lack of consensus between studies it is important to establish the characteristics of rDNA within the cell line used in this study and gain a more thorough understanding of the rDNA landscape in C57BL/6J.

1.5 The epitranscriptome and rRNA modifications

1.5.1 RNA modifications

DNA modifications are just one level of gene regulation control; this is a multi-layered process that continues beyond the point of transcription. The processing of nascent RNA through a multitude of post-transcriptional pathways can give rise to a variety of distinct molecules expanding the biological diversity, function, and impact of a single gene (Sloan *et al.*, 2017). Alternative splicing events can give

rise to unique RNA isoforms from a single mRNA precursor molecule to expand an organism's protein repertoire (Black, 2003). Eukaryotic mRNA modifications like the 5'-cap modulate RNA export (Lewis and Izaurralde, 1997) and translation, promotes mRNA stability (Guhaniyogi and Brewer, 2001). Whilst these have long been considered the only relevant post-transcriptional changes, every RNA nucleotide can be chemically modified or even completely interchanged (Reviewed in detail by Li *et al.*, 2014; Roundtree *et al.*, 2017). Recent transcriptome-wide mapping approaches show that all major classes of RNA are modified in some form or another, with increasing evidence suggesting that RNA modification changes play a vital role in fine-tuning gene expression during development and stress responses as well as being critical for RNA metabolism.

Nucleotide modifications are found in all 3 domains of life and, currently, over 170 different modifications have been discovered, collectively referred to as the 'epitranscriptome' (Cantara *et al.*, 2011; Machnicka *et al.*, 2013). Internal modifications, such as methyl-6-adenosine (m^6A), methyl-5-cytosine (m^5C), ribose-methylation (2'-O-Me), and pseudouridine (Ψ), are commonly found in coding RNAs (cRNA), have been known to exist for over 50 years, and, whilst their sites and functions are slowly uncovered, modifications on the vast majority of RNAs remain unmapped and their functional significance unknown (Kumar and Mohapatra, 2021).

1.5.2 Ribosomal RNA modifications

Ribosomal RNAs are the second most modified type of RNA after transfer RNAs (tRNAs), with ~2% of nucleotides modified, equivalent to over 200 sites in a single human and mouse pre-rRNA molecule. Though a great variety of RNA modifications are seen in nature, only a few types are noted in rRNA, with human rRNA only known to contain 14 distinct types of rRNA modification at ~228 sites (Taoka *et al.*, 2018). The most common modifications found in eukaryote rRNAs are 2'-O-methylations, which is methylation of the ribose of any nucleotide, and Pseudouridylation (Ψ), which is the isomerisation of uracil to pseudouridine. These types of modifications are carried out by RNA-dependent nuclear mechanisms which rely on small nucleolar RNAs (snoRNA) to guide enzymes to the modification sites via sequence specific base pairing. Small nuclear RNAs can be categorised into two types, with box C/D type snoRNAs involved in 2'-O-Methylation and box H/ACA type involved in pseudouridylation (Kiss-László *et al.*, 1996; Ganot, Bortolin and Kiss, 1997). Of each of these commonly occurring modifications, ~50 are reported in yeast rRNA and ~100 of each in human rRNA (Birkedal *et al.*, 2015; Sharma and Lafontaine, 2015; Sloan *et al.*, 2017; Taoka *et al.*, 2018). Base-specific modifications, like m^6A , m^5C (methylation of adenine and cytosine respectively), are also found throughout rRNA, however to a much lesser degree, and are outnumbered by 2'-O-Me and Ψ sites by almost 10-fold in yeast and human (Sloan *et al.*, 2017) and are installed by stand-alone enzymes via RNA-independent mechanisms (Yang *et al.*, 2016).

Modifications to rRNA are considered to be introduced at different stages during the maturation of the ribosomal subunits with findings suggesting that snoRNP-mediated modifications largely introduced during the early stages of ribosome biogenesis, when the pre-ribosomal complexes are thought to have a more open structure. Kinetic labeling in yeast has revealed that the vast majority of 2'-O-methylations in the 18S rRNA are introduced co-transcriptionally, while such methylations are introduced both co- and post-transcriptionally into the 25S rRNA (Kos *et al.*, 2010). In the 18S rRNA only one modification, Am100, occurs after release of the nascent transcript from the rDNA, while the extent of 2'-O-methylation of sites A817, G867, A867, A2256, U2421 and A2640 has been shown to be significantly higher in the mature 25S than in the chromatin associated rRNA (Lapeyre *et al.*, 2004). In the thermophilic filamentous fungus *Chaetomium thermophilum*, it was observed that the recent structure of a 90S pre-ribosomal complex did not contain any rRNA modifying snoRNPs implying the dissociation of most snoRNPs from the pre-rRNA transcript upon its release from the rDNA (Kornprobst *et al.*, 2016). Similarly, In human cells, majority of snoRNA-guided modifications are thought to likely occur at early pre-ribosomal complexes, however, some snoRNAs have been shown to associate with later pre-SU particles (Sloan *et al.*, 2015). Additionally, snoRNAs are known to form extensive and often overlapping base-pairing interactions with their target rRNA sequences, implying that in many cases individual modifications must be introduced in a stepwise manner (Birkdel *et al.*, 2015). However, whether the association of particular snoRNAs with their pre-rRNA base-pairing sites occurs stochastically or if there is a defined hierarchy for snoRNA recruitment to pre-ribosomal complexes currently remains unclear.

In contrast to the 2'-O-methylations and pseudouridylations that are largely introduced during the early stages of ribosomal subunit maturation, base modifications are generally thought to occur later. For instance, in the case of the N³-acp modification of 18S- m¹Ψ1191, the exclusively cytoplasmic localization of the Tsr3 enzyme that installs this modification clearly identifies this as a late event in yeast (Meyer *et al.*, 2016). Additionally, whilst early cytoplasmic pre-40S complexes show low levels of N³-acp modification of 18S- m¹Ψ1191, later particles show modification levels similar to that of mature 18S rRNA (Fatica *et al.*, 2003). Whilst some base modifications such as this are well studied, the precise timing of most remains yet to be determined.

1.5.3 rRNA modifications and ribosome structural integrity

Ribosomal RNA modifications fundamentally expand the topological potential of specific nucleotides and contribute to the stabilisation of the secondary and tertiary structures, ultimately impacting ribosome function (Helm, 2006; Yang *et al.*, 2016). For instance, pseudouridylation, improves the rigidity of the sugar-phosphate backbone by conferring greater hydrogen bonding potential than uridine (Davis, 1995). Similarly, 2'-O-Methylation stabilises helices and improves base stacking (Kawai *et al.*, 1992). Base modifications confer similar stabilising advantages but also have modification-

specific benefits. For instance, N³-Methylation of uridine promotes hairpin formation (Micura *et al.*, 2001), whilst N⁷-Methylation of guanine increases the positive charge of the nucleotide and promotes ionic interactions between proteins and RNA (Agris, Sierzputowska-Gracz and Smith, 1986). Analysis of yeast ribosomes lacking methylation of 25S-C2278 (cytosine at position 2278 in 25S, yeast equivalent of the 28S) and ribose methylation of 25S-G2288, show decreased ribosomal stability and a loss of proteins from the large subunit (Gigova *et al.*, 2014). Besides the modulation of the local ribosomal environment, rRNA modifications are also suggested to facilitate communication between distant regions to control ribosomal structure, either through the alterations in ribosome folding and assembly or the interactions between rRNA modification and ribosomal proteins (Birkedal *et al.*, 2015; Sharma and Lafontaine, 2015). Overall, it is clear that rRNA modifications play a critical role in maintaining the stability of the ribosomes' overall structure.

1.5.4 rRNA modifications on ribosome function

The sites of modification have been extensively mapped in yeast and human rRNA, with sites found throughout all 4 rRNA species (reviewed in Sloan *et al.* 2016). The distribution of both snoRNA-guided and RNA-independent base modifications is not random, rather they are found clustered in functionally significant regions including the tRNA binding sites, DC and PTC, as well as the subunit interface (Decatur and Fournier, 2002; Ben-Shem *et al.*, 2011). The essential role of rRNA modifications for ribosome activity is demonstrated by a variety of functional studies in yeast involving catalytically inactive mutants of Nop1 and Cbf5, enzymes vital for 2'-O-Methylation and Pseudouridylation respectively (Tollervey *et al.*, 1993; Zebarjadian *et al.*, 1999). Whilst the resulting global loss of these modifications has a catastrophic impact on ribosome function and organismal health, the loss of only a few individual modifications has been shown to significantly impact cell viability and ribosome methylation of 25S-G2922 greatly impacts ribosome structure and cell growth (Baxter-Roshek, Petrov and Dinman, 2007). Similarly, deletion of snR35, which initiates the modification of 18S-U1191 results in significant impairment in the ribosomal small subunit biogenesis (Baudin-baillieu *et al.*, 2009). Though specific modifications are vital for ribosome function and biogenesis, studies involving the deletion of clusters of modification, specifically from functional regions show that significant phenotypes are often only seen upon the loss of multiple modifications (King *et al.*, 2003; Baudin-baillieu *et al.*, 2009). Strains with deletion of snoRNAs resulting in loss of 2'-O-Methylation and pseudouridylation at tRNA binding sites P and A are accompanied by diminished translation efficiency (Baudin-baillieu *et al.*, 2009). A similar effect is observed upon removal of 6 Ψ 's in the peptidyltransferase site or 3-4 modifications within a helical interunit bridge (Liang *et al.*, 2009; King *et al.*, 2003). These studies demonstrate that modifications largely act cumulatively to enable proper ribosome function.

Besides impacting global protein synthesis, certain modifications act to regulate the translation of specific mRNA subsets. For instance, changes in rRNA pseudouridylation are shown to alter the affinity of specific mRNAs containing internal ribosome entry site (IRES), impacting transcriptional initiation (Yoon *et al.*, 2006; Basu *et al.*, 2011). Additionally, loss of methylation at 25S-C2278 has been shown to promote the polysomal recruitment of a subset of mRNA involved in the oxidative stress response, modulating their translation (Schosserer *et al.*, 2015). Overall, the evidence suggests that RNA modifications impact ribosomal functioning, biogenesis and regulate both global and specific protein expression.

1.5.5 The implication of rRNA modifications in disease states

Considering the vital role rRNA modifications play in ribosomal stability and function, there is growing evidence linking defects in rRNA modification machinery to both developmental aberrations and disease. For instance, a point mutation in the gene *EMG1*, which encodes an rRNA methyltransferase, causes Bowen-Conradi syndrome, a rare but highly lethal developmental defect. Another example is the deletion of a chromosomal segment that encompasses genes *WBSR22* and *WBSR20*, methyltransferases involved in rRNA base modifications. Deletion of this loci is implicated in Williams-Beuren syndrome, rare genetic disorder characterized by prenatal and postnatal growth retardation (Doll and Grzeschik, 2001; Armistead *et al.*, 2009). Additionally, changes in rRNA 2'-O-methylation patterns at various rRNA sites are linked to cancer development, with varying levels of the modification observed for different cancer types (Krogh *et al.*, 2020). Whether directly or indirectly, aberrations in rRNA pathogenesis. The knock-on disturbance to ribosome function may likely lead to changes in translation and the cellular proteome, contributing to disease phenotypes.

1.6 Ribosome heterogeneity

Beyond the aberrations in health arising from disturbances to ribosome composition, 'healthy' ribosomes are far from identical. Once considered to be uniform, indiscriminate, protein production machines, ribosomes are now thought to display a great deal of heterogeneity, having specialized roles in the cell. Heterogeneity may arise at any level of ribosome biosynthesis and may serve to expand the modulation of genes, and fine-tune protein expression in a dynamic and environmentally contextual way (Genuth and Barna, 2018).

1.6.1 Ribosome heterogeneity and the ribosome filter hypothesis

The ribosome filter hypothesis proposes that the interactions between ribosomal proteins, rRNAs, and mRNAs play important roles in the fine-tuning and control of gene translation. Besides the established mechanisms of translation and regulation, accumulating evidence suggests that ribosomes

themselves may be regarded as gene regulatory elements with studies suggesting that ribosomes can selectively influence the translation rate of specific mRNAs. This process is thought to be dependent on the interaction of specific sequences found in sub-sets of mRNA, which compete for binding sites on ribosomal subunits (Mauro and Edelman, 2002). The extent of these interactions may be altered by differences in ribosome composition which in turn may impact the ribosomes' affinity for specific mRNAs (Xue and Barna, 2012). This heterogeneity may arise from differences in the rRNA and protein composition, or post-translational modification of ribosomal proteins as well as variations and modifications of rRNA. As a result, structurally distinct populations of ribosomes may differ in their ability to translate specific subsets of mRNAs with heterogeneity leading to differential rates of mRNA translation in different cells in a condition-specific manner, with heterogeneity giving rise to ribosomes that 'specialise' in the translation of specific mRNAs (Xue *et al.*, 2012; Parks *et al.*, 2018).

1.6.2 Ribosomal proteins- a source of ribosome heterogeneity

In mammalian cells, the majority of ribosomal proteins are encoded by a single gene. However, the small subunit protein eS4 (S4) is encoded by three genes (*RPS4Y1*, *RPS4X*, *RPS4Y2*), located on the X and Y chromosomes. In males, *RPS4Y1* and *RPS4X* are expressed in nearly all cells, whilst *RPS4Y2* is expressed only in the testis and prostate (Xue and Barna, 2012), suggesting a role for tissue-specific developmental context, with developmental stage dependant differences in ribosomal protein composition and modification observed. During the vegetative stage of amoebae, *Dictyostelium discoideum*, the ribosomal protein eS19 (S19) is phosphorylated, uS10 (S20) protein is dephosphorylated and uL2 (L2) protein is methylated. When amoebae aggregates to form a fruiting body (a more advanced developmental stage) the eS19 protein is dephosphorylated, uS10 (S20) protein is phosphorylated and uL2 undergoes demethylation. Additionally, during aggregation, ribosomes are depleted of eL18 (L18) protein, indicating that it may not be necessary in certain growth stages (Ramagopal, 1990).

A striking example of ribosome heterogeneity can be seen in mice, where the loss of a particular ribosomal protein can alter the translation of certain mRNAs and dramatically alter the individual's anatomical structure. In this example, mutant mice, depleted of the ribosomal protein eL38, experience inhibition of homeobox (*Hox*) mRNA translation, without significant effect on global protein synthesis. *Hox* genes are involved in morphogenesis and the loss of eL38 changes mice rib cage patterns with mutant mice having an extra pair of ribs and unusually kinked tails compared to the wild type. The involvement of eL38 in tRNA movement during translation and positioning of rRNA is thought to contribute to its translational control of specific mRNAs (Gopanenko *et al.*, 2021).

1.6.3 Ribosomal DNA- a source of ribosome heterogeneity

Ribosomal heterogeneity is not only based on varying protein combinations and modifications but also differences in ribosomal RNA. The genome of the halophilic red archaeon *Haloarcula marismortui*, encodes three types of paralogous rRNA operons (*rrnA*, *rrnB*, *rrnC*), which serve to facilitate its survival at both high and low temperatures (Baliga *et al.*, 2004; Sato, Fujiwara and Kimura, 2017). Operons *rrnA* and *rrnC* are identical and are expressed at low temperatures. Operon *rrnB*, however, is highly divergent, and is repressed at low temperatures but expressed at high temperatures. Comparing operon gene expression, at 50°C, operon *rrnB* expression was seen to be four times higher than *rrnA* and *rrnC*, whilst its expression was 3 times lower at 15°C. The *rrnB* operon contains ~135 SNPs across the three 16S, 23S, and 5S rRNA genes and a much greater percentage of GC base pairs, which are thought to increase the structural stability at higher temperatures via the increased potential for hydrogen bonding (Lopez *et al.*, 2007).

1.6.4 Ribosomal RNA modifications- a source of ribosome heterogeneity

Besides differences in the core genetic sequence, post-translational modification of rRNA serves to further extend ribosome heterogeneity. Recent advances in RNA modification detection and mapping have uncovered that rRNA modifications are not constitutive as once thought, instead some sites present with partial modification ((Birkedal *et al.*, 2015)). Under normal growth conditions, base modifications appear to be constitutive however, sites of pseudouridylation and 2'-O-methylation appear at substoichiometric levels. Out of the 112 modification sites identified in yeast, 18 sites are modified in less than 85% of ribosomes (Taoka *et al.*, 2016), whilst studies in human cell lines approximate one-third of 2'-O-methylation to be at substoichiometric levels (Krogh *et al.*, 2016). For the most part, the cause of partial modification remains unknown, however, it appears that cell-specific abundance of certain snoRNAs is a likely contributing factor. This is supported by the observation that low levels of cellular snR51, a snoRNA that guides 2'-O-methylation of 18S-A100, correlate with substoichiometric modification at this position, which can subsequently be reversed by snR51 overexpression (Buchhaupt *et al.*, 2014). Besides the partial modifications observed under normal growth conditions, varying degrees of modification at specific positions are also observed in response to environmental changes. Some striking examples of this include the changes in pseudouridylation levels detected at positions 25S-2314 and 5S-50 in response to diauxic shift and heat shock respectively. Changes in modification patterns in a developmental context have also been observed, with differential methylation of a subset of rRNA sites reported between developing organs and their adult counterparts (Hebras *et al.*, 2020).

Together, these observations support the idea that ribosome heterogeneity greatly expands the functional significance of ribosomes and may be fundamental for facilitating gene-environment

interactions and modulating gene expression in a dynamic environment.

1.7 Exploring ribosomal heterogeneity in C57BL/6J mice

1.7.1 rDNA promoter variants

Although rDNA units are considered ‘copies’, units within a gene cluster are not completely identical (Tseng *et al.*, 2008). Sequence variation is seen in rDNA genes for both mice and humans, with single nucleotide polymorphisms (SNPs) within both the coding and promoter regions (Qu, Nicoloso and Bachellerie, 1991; Shiao *et al.*, 2005; Tseng *et al.*, 2008). Several studies have shown that genomic context is a critical determinant of DNA methylation patterns at certain loci, with allele-specific methylation profiles shown to be dictated by SNPs (Kerkel *et al.*, 2008; Schilling, El Chartouni and Rehli, 2009; Docherty *et al.*, 2012). Work from our lab has shown that inbred C57BL/6J mice exhibit two distinct rDNA promoter variants, distinguished by a promoter SNP at position -104 upstream of the transcriptional start site (Holland *et al.*, 2016). Specifically, the two variants termed the “A” and “C” variants are respectively defined by either adenine or cytosine at position -104 and lie close to a functionally significant CpG site at position -133 (**Figure 1.5**).

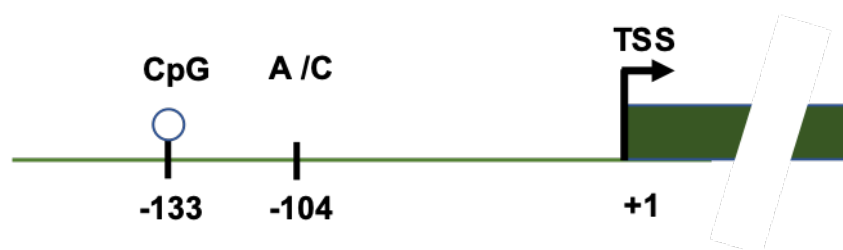


Figure 1.5 C57BL/6J rDNA promoter variants. Schematic of the C57BL/6J rDNA promoter showing the positions of A / C SNPs at position -104, in relation to the functional CpG site at -133 and transcriptional start site (TSS).

1.7.2 Variant-specific methylation dynamics

An investigation into the impact of early life nutritional stress on the epigenetic regulation of rDNA revealed differential methylation of the “A” and “C” rDNA promoter variants, the key findings of which are presented in **Figure 1.6**. In the context of *in utero* protein restriction, Holland *et al.* (2016) observed that *in utero* protein restriction correlated with weaning weight (**Figure 1.6A**). Exploring the potential role of rDNA in this observed phenotype, it was found that rDNA copies with an “A” at position -104 were preferentially methylated at CpG -133, in comparison to rDNA copies with a “C” at position -104, in both control and protein-restricted offspring sperm (**Figure 1.6B**). The “A” variants are associated with 30–80% methylation of promoter -133 CpG site, in contrast to “C” variants which display < 25% methylation (**Figure 1.6C**). Additionally, a positive correlation between the proportion of A variant rDNA copies and methylation of CpG -133 emerged but only in response to protein restriction (**Figure**

1.6D). Furthermore, the study revealed a negative correlation between the weaning weight and the proportion of rDNA “A” variant units that were methylated at CpG-133, this was only observed in the PR offspring (**Figure 1.6E**). Interestingly, the genetic variation made little difference to the degree of rDNA methylation in control mice, suggesting context-specific environmental sensitivity of rDNA A/C promoter variants. These observations together suggest that rDNA genetic variation may dictate the epigenetic response, and may predetermine, or prime an individual’s sensitivity and susceptibility to environmental insults.

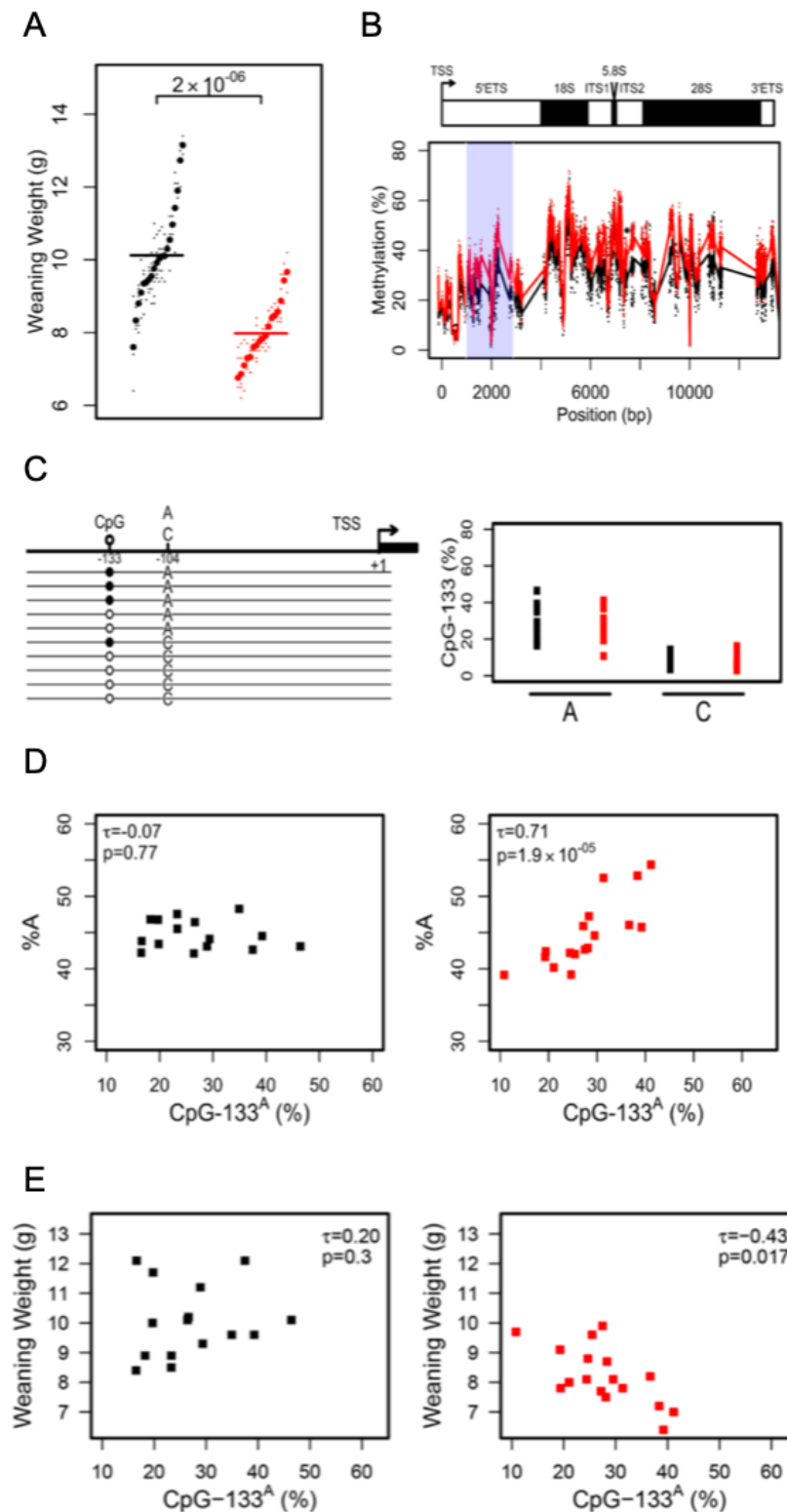


Figure 1.6 Key results from Holland et al., 2016. (A) Weaning weight comparison of Control males (black) and PR males (red). Horizontal line = litter means, coloured circles = individuals. (B) RRBS determination of rDNA promoter hypermethylated in PR sperm (red, $n=8$) compared to controls (black, $n=7$). (C) Schematic showing Bis-PCR amplicons spanning both the CpG-133 (black circle = methylated) and the variant at position -104 (A/C) (left panel). The percentage of CpG-133 methylation was greater for A variants compared to C variants in both control (black, $n=12$) and PR (red, $n=15$) sperm (right panel). (D) Methylation of A-variant rDNA copies at CpG-133 positively correlates with the percentage of total “A” rDNA copies in PR sperm (red, $n=15$; $\tau=0.71$, $P=1.9 \times 10^{-5}$) (right), but do not correlate in control sperm (black, $n=12$; $\tau=-0.08$, $P=0.77$) (left). (E) Methylation of A variant copies at CpG-133 correlates negatively with weaning weight but only in the PR group (red, $n=17$; $\tau=-0.43$, $P=0.017$) and not in the control group (black, $n=15$; $\tau=0.2$, $P=0.3$).

1.7.3 Distinct rDNA haplotypes

A more recent study conducted by our lab has shown that units of 45S rDNA in the C57BL/6J mouse strain exist as distinct genetic haplotypes. Using a combination of short-read whole-genome sequencing (WGS) and whole-genome long-read Nanopore sequencing, the study by Rodriguez-Algarra et al., (2022) identified 88 different coding unit single-nucleotide variants (SNVs). Using the previously identified -104 promoter SNP as an anchoring point, the study revealed 4 different rDNA haplotypes found in approximately equal proportions (Rodriguez-Algarra *et al.*, 2022). The haplotypes, termed “ATA,” “ATG,” “CCA,” and “CTA,” can be distinguished by SNVs at specific haplotype defining positions -104, 8063, and 12736, with position 12736 distinguishing the 2 haplotypes with “A” at -104, and position 8063 distinguishing the 2 haplotypes with “C” at -104 (**Figure 1.7**). Using a combination of Nanopore methylation analysis and reduced representation bisulfite sequencing (RRBS) it was shown that the ATA haplotype displayed significant DNA methylation ($\geq 60\%$) across the length of the coding unit, whilst the CCA haplotype showed comparatively low levels of methylation ($\leq 20\%$), and the other two haplotypes (ATG & CTA) were largely unmethylated. Additionally, the analysis of individual reads revealed that the individual coding units are either unmethylated or almost completely methylated. Ribosomal RNA-seq analysis shed light on the functional outcomes of DNA methylation showing that gene methylation levels negatively correlated with haplotype expression, showing that the epigenetic state influences the transcriptional output of a unit. These findings present an exciting basis from which to build our understanding of the interplay between the genetic code and our environment.

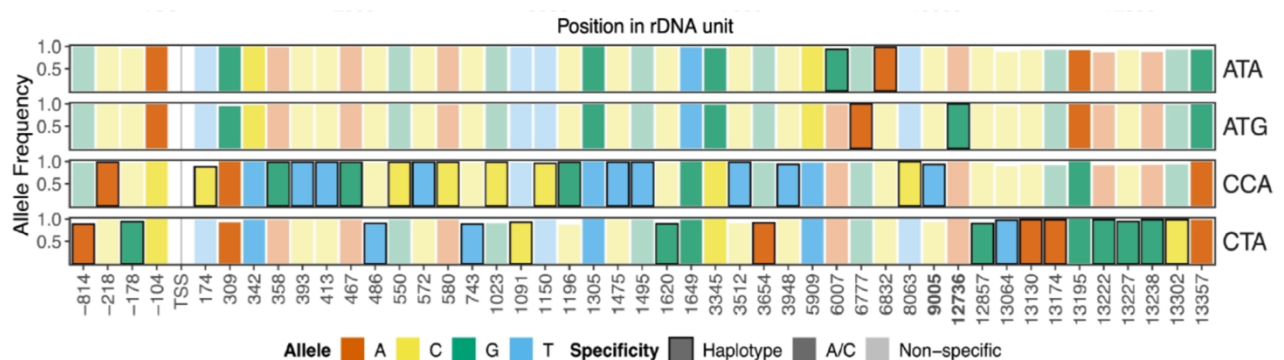


Figure 1.7 Long-range haplotype characterization of 45S rDNA in the C57BL/6J strain. Only SNVs that distinguish rDNA haplotypes are shown in each track. Bars with bold outline represent SNVs unique to the specific haplotype. Bars with non-muted colours and no outline indicate positions associated with the A/C haplogroups defined by the variant at position -104. Bars with muted colours and no outline indicate non-specific nucleotides. Labelled nucleotide positions x-axis that are in bold (9005, 12736) are variants found within the 28S rRNA and incorporated into mature ribosomes. Figure is taken directly from Algarra et al., 2022 Figure 1B.

The great degree of genetic diversity found across 45S rDNA and the differential methylation observed between distinct haplotypes is a likely contributor to ribosomal heterogeneity. Genetic differences in rDNA may serve to alter rRNA modification profiles, refining the function of the ribosomes into which distinct rRNAs are incorporated. It remains to be studied, if rDNA haplotype expression and the modification profiles of rRNA display cell-or tissue-specificity, and if differential alterations to rRNA variants are observed within a developmental context. To answer these questions, it is important to further dissect the genetic, and epigenetic landscape of this obscure region of the genome.

1.8 Sequencing methodologies

Over the last fifty years, tremendous effort has gone into deciphering the genetic code governing life and disease. Advancements in the fields of genetics now permit the sequencing of whole genomes within a matter of hours, a feat considered incomprehensible at the dawn of the field of genetic research. This incredible progress is a result of innovative sequencing methods, from the inception of first-generation Sanger sequencing in 1977 to the current day third-generation long-read sequencing methods developed by Oxford Nanopore Technologies. During the past few decades, the field has seen significant improvements in read length, and accuracy as well as a cost reduction, opening up avenues to pursue novel scientific curiosities (Bansal and Boucher, 2019).

1.8.1 Short read sequencing methods

Sequencing technologies can be broadly divided into short-read and long-read approaches distinguished depending on their read length, i.e. the size of the nucleotide sequence inferred from a single continuous molecule. Short read sequencing technologies include first-generation Sanger sequencing and next-generation sequencing technologies [NGS] like Illumina sequencing. With a read length of a few 100 bp, these methods remain the gold standard and are well suited for most genetic profiling purposes (Slatko, Gardner and Ausubel, 2018). Short read sequencing methods generally rely on sequence by synthesis and are dependent on the artificial fragmentation of nucleotide molecules, both of which limit the size of molecules that can be assessed. This is a concern, especially for deciphering over two-thirds of a eukaryote genome, which is thought to be made up of highly repetitive regions (Gemmell, 2021). Long, repetitive stretches of the genome such as rDNA clusters are incompatible with short-read sequencing methods since it is entirely impossible to accurately reassemble them from a pool of short reads (Treangen and Salzberg, 2012). As a result, understanding the structural arrangement of entire units of repetitive sequences as well as the relationship between distally positioned SNVs cannot be gained by employing these methods.

Conventional RNA sequencing and modification mapping methods suffer from similar issues as short-read DNA sequencing. They largely rely on second-generation sequencing methods that necessitate the reverse transcription of RNA into cDNA (Schwartz and Motorin, 2017). As a result, the RNA molecule is not directly sequenced in its native state and RNA modifications have to be called through convoluted and indirect methods. Additionally, due to the reliance on NGS platforms, molecule fragmentation is a prerequisite preventing the study of larger transcripts and the relationship between distally located SNVs.

1.8.2 Oxford Nanopore Technology sequencing

In contrast, Nanopore sequencing, developed by Oxford Nanopore Technologies (ONT) employs a unique sequencing approach not constrained by the limitations of sequencing by synthesis, or molecule fragmentation. The general principle of nanopore sequencing is depicted in **Figure 1.8**. Briefly, the technology is based on the use of an insulated membrane into which specialised nanopores are embedded and through which an ionic current is generated. A motor protein is used to thread single-stranded nucleotide molecules through the nanopore, with the bases that occupy the nanopore, effectively blocking the flow of ions (Wang *et al.*, 2021). Owing to their distinct chemical and structural properties, different nucleotide bases disrupt the flow of ions and the resulting current recordings in characteristic ways (Stephenson *et al.*, 2022). Depending on the sequencing approach, at any one time, 3-6 nucleotides occupy the pore, with this string of bases referred to as a 'k-mer'. By employing pre-trained machine learning algorithms, overlapping current disruptions from sequential k-mers are used to infer the nucleotide sequence and effectively, 'base-call' the molecule (Furlan *et al.*, 2021).

ONT offers a range of sequencing platforms for many sequencing needs, several library preparation kits for a variety of sample types, as well as an array of flow cells with different sequencing capacities and kit compatibilities (**Figure 1.9**). Amongst others, there are three main ONT sequencing platforms, the Flongle, MinION, and PromethION, compatible with unique flow cells, each intended for increasing sequencing output and genome coverage. The Flongle flow cell is designed with up to 126 nanopore channels, with a maximum theoretical 1D sequencing output of 3.3 Gigabases (Gbp). It is designed for sequencing small genomes such as those of viruses and bacteria, or for the quality control of library preparations before larger sequencing experiments. The MinION flow cell is designed with up to 512 nanopore channels, with a maximum theoretical sequencing output of 40 Gb, and is designed for the low-pass sequencing of larger genomes. The PromethION flow cell has the highest output of any nanopore flow cell and is designed to have up to 2,675 sequencing channels and a maximum theoretical 1D sequencing output of 290 Gb, permitting the sequencing of large genomes to high

coverage. Here, Sequencing output specifically refers to the number of bases sequenced and is determined by factors including the number of available sequencing pores, the speed of molecule translocation through the pore as well as the length of the sequencing run, and directly relates to genome coverage output (Wang *et al.*, 2021). The availability of sequencing pores differs significantly between flow cell types, and whilst a certain capacity can be expected, owing to the biological nature of pores, differences between flow cells of the same type are often observed. Additionally, with increasing sequencing run time, a reduction in sequencing capacity is observed, negatively impacting output yield (**Figure 1.10**). This is often due to deterioration of the translocation machinery, and the non-covalent blockage of sequencing pores with nucleotide molecules (Li *et al.*, 2021). Whilst a fraction of blocked pores can be rescued by digesting the nucleotide sample through a 'Flow cell wash kit' protocol to increase overall sequencing output as shown in **Figure 1.10**, an overall reduction in pore capacity is unavoidable across time and persist post-washing (Kubota *et al.*, 2019). As a result, the read output from secondary sequencing runs conducted post wash on used flow cells are likely to fall far below the theoretical maximum yields.

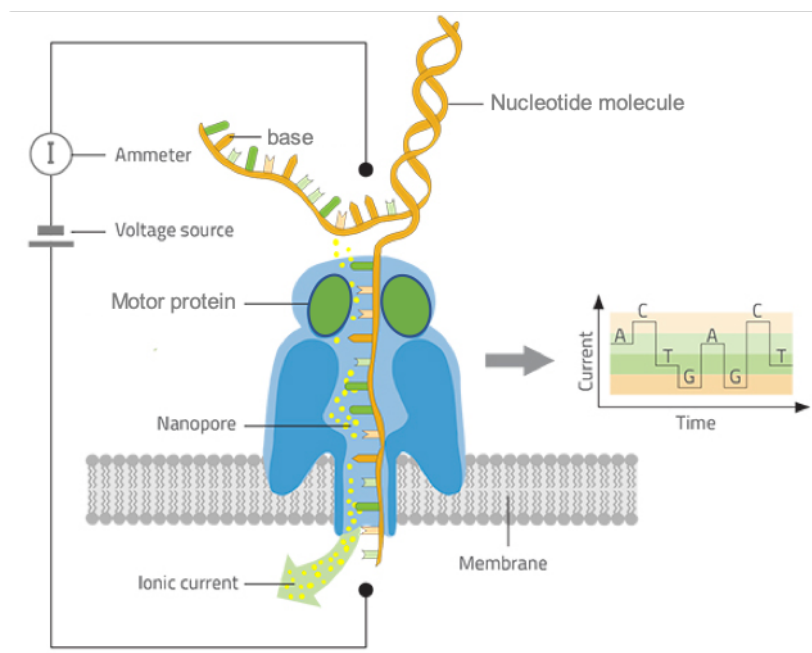


Figure 1.8 Oxford Nanopore Technologies Nanopore sequencing principles. A schematic presenting the principles behind ONT nanopore sequencing. A single representative pore is shown, embedded into an insulated membrane through which an ionic current is passed. A motor protein threads a single strand of a nucleotide molecule through the nanopore, and the nucleotides occupying the pore disrupt the flow of ions. Ionic current disruptions are detected by an ammeter and as due to their specificity to different nucleotides, are used to infer the nucleotide sequence using machine learning algorithms. Image adapted from Oxford Nanopore Technologies.


Flow cell	Flongle (FLO-FLG001)	MinION (FLO-MIN106D)	PromethION (FLO-PRO112)
			
Cost per flow cell*	£60	£380-£720	£480-£1,120
Available sequencing channels	Up to 126	Up to 512	Up to 2,675
Theoretical 1D max output	Up to 2.8 Gb	Up to 50 Gb	Up to 290 Gb
Direct RNA sequencing kit (SQK-RNA002) compatible **	No	Yes	Yes
Direct RNA sequencing kit (SQK-RNA002) expected yield	-	1 million reads per full flow cell	3 million reads per full flow cell

Figure 1.9 Nanopore flow cell comparison. * Cost per flow cell varies depending on promotional offers and bulk orders. Cost does not include start-up costs or the price of technical services associated with the use of some flow cell types and ONT devices. ** SQK-RNA002 kit compatibility is presented based on ONT certified protocols only. Flongle, MinION and PromethION node images are reproduced from Oxford Nanopore Technologies. All information is accurate as of 2022.

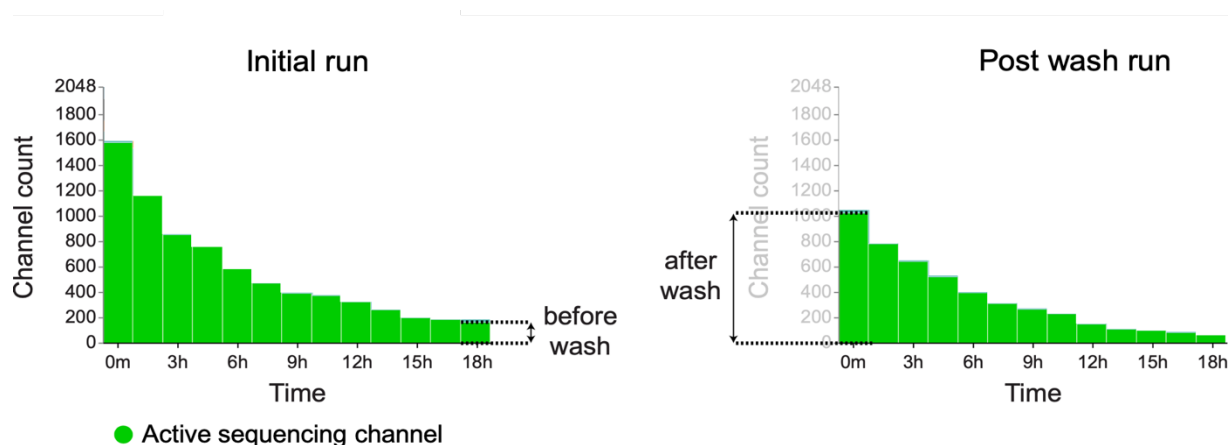


Figure 1.10 MinION flow cell sequencing capacity progression. A MinION flow cell loaded with a sequencing library results in a decrease in active sequencing channels, leading to a decrease in the rate of data acquisition: after 18 hours, the flow cell has <200 single pores available for sequencing from a starting point of ~1600. Washing the flow cell with ONT flow cell wash kit (EXP-WSH004) rescues a proportion of blocked unavailable pores with the number of available single pores increasing to ~1000. Images reproduced from Oxford Nanopore Technologies.

Nanopore sequencing, directly interprets the sequence of an individual nucleotide molecule, without the need for prior fragmentation and in theory imposes no size limitation to read length (Goodwin, *et al.* 2016). Rather, read length is dependent on the length of nucleotide molecules successfully isolated for library preparations, with ONT DNA sequencing routinely yielding reads 50-100 Kbp in length (Branton *et al.*, 2009; Amarasinghe *et al.*, 2020). Read length can be further increased with the ONT Ultra-Long DNA Sequencing Kit (SQK-ULK001) which offers a means of preparing ultra-high molecular weight (uHMW) DNA for sequencing by prioritizing the minimisation of mechanical sheering of DNA to preserve molecule length. Using this method, the largest recorded read measures 2.3+ Mbp in published data sets (Payne *et al.*, 2019) and 4+ Mbp in ONT data sets. However, whilst the capture of multi Mbp reads is possible, the reality is that reads of such size are far and few in between. **Figure 1.11** is a representative read length distribution histogram for ONT uHMW DNA Seq published by ONT to demonstrate the capabilities of this kit. A standard sequencing run can be expected to provide an output with an N50 >50 KB and reads routinely measuring in excess of 100 Mbp. However, read length tapers off substantially with read length negatively correlating with fraction of total reads and read length generally expected to cap out at ~400 kbp. Therefore, whilst the capture of multi Mbp reads is possible, it is not the norm, and even when captured, such reads are likely to occur in microscopic quantities

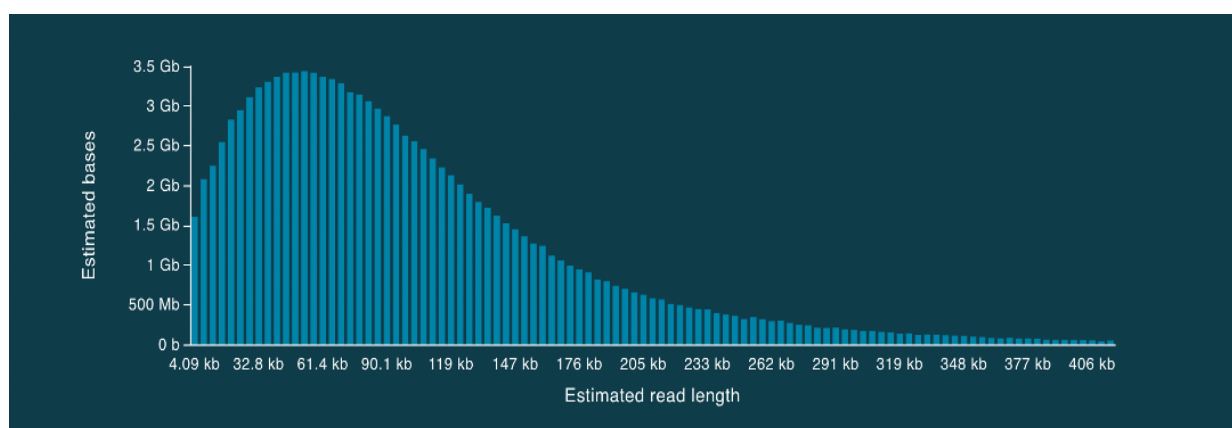


Figure 1.11 ONT ultra long DNA sequencing kit (SQK-ULK001) read length distribution histogram. Images reproduced from Oxford Nanopore Technologies.

1.8.3 Sequencing ribosomal DNA arrays

Currently, rDNA arrays remain poorly characterised in reference genomes and are often under represented as single rDNA reference copies or unassembled and completely detached from chromosomal context (Zentner *et al.*, 2011). Long-read sequencing methodologies present the possibility for understanding the chromosomal location of rDNA as well as its structural, genetic, and

copy number variation. Although efforts to fully assemble rDNA arrays remain underway, the limitation currently associated with long-read methods, specifically read length and base calling accuracy, means this remains a challenge. A recent study showcased the assembly of the *C. elegans* 5S rDNA array demonstrating the potential of long-read sequencing methods, however, this same study failed in assembling across the 45S array as a result of insufficiently long reads and the little variation that exists between 45S rDNA copies (Ding, *et al.* 2021). The difficulty of this feat is further emphasised by the results of a separate study aiming to recomplete the *C. elegans* genome in which a combination of PacBio and Nanopore sequencing also failed to obtain reads spanning the entire 45S array (Yoshimura *et al.*, 2019). A concerted effort combining BAC cloning, short-read, and long-read sequencing was however successful in assembling one of two rDNA arrays in the *A. thaliana* genome (Sims, Schlögelhofer and Kurzbauer, 2021).

Decoding the rDNA arrays of complex mammalian genomes poses a much greater challenge than that of simpler model organism due to the greater multitude of loci and larger cluster and repeat unit sizes. Despite this, an impressive effort to characterise rDNA arrays in a functionally haploid human cell line has led to notable success. The study by the Telomere-to-Telomer consortium, has published the first ‘complete’ 3 Bbp human genome of cell line T2T-CHM13, providing novel insights into previously unmapped region of the genome including rDNA arrays. An international effort by a 100 researchers across a number of academic institutes, and an immense sequencing effort utilising multiple technologies, including 30× PacBio circular consensus sequencing (HiFi) and 120× Oxford Nanopore ultralong-read sequencing, the study reports successfully assembling the smallest two of five human rDNA arrays in their entirety (Nurk *et al.*, 2022). In an attempt to explore the architecture of rDNA loci, the study utilised Hi-Fi (PacBio) based sparse de Bruijn graphs for each of the five rDNA arrays. ONT reads were subsequently aligned to the graphs to identify a set of walks which were converted to sequence, segmented into individual rDNA units, and clustered into “morphs” according to sequence similarity. The copy number of each morph was estimated from the number of supporting ONT reads with ONT reads spanning two or more rDNA units used to build a morph graph representing the internal structure of each array. The study found that the shorter arrays found on chromosomes 14 and 22 consist of a single sequence morph type arranged in a head-to-tail fashion, whereas the longer arrays on chromosomes 13, 15, and 21 exhibit a more mosaic structure involving multiple, interspersed morphs. Owing to read length limitations the ONT reads were not long enough to fully resolve the ordering for longer rDNA arrays on chromosomes 13, 15 and 21, and the primary morphs were artificially arranged to reflect the estimated copy number in the model sequences presented.

Similarly, Nanopore sequencing has been employed to explore the large scale structure of rDNA loci in human cells lines (Hori *et al.*, 2021). By utilising both publicly available Oxford Nanopore whole-

genome sequencing (WGS) data from the Human Pangenomics Project (HPGP) and study specific Cas9-enriched rDNA reads from Epstein-Barr virus (EBV) transformed B cells and primary fibroblast cells, the study found that whilst each human rDNA copy has some variations in its noncoding region (IGS), contiguous copies of rDNA display similar variation patterns. By specifically analyzing the differences in lengths of the R and Butterfly/Long repeat regions (repeat sequences within the IGS portion of rDNA units), it was shown that the distributions of length difference between contiguous copies were clearly shorter than the randomized simulated control in both regions. These observations indicate that contiguous copies are more similar than non-contiguous ones, suggesting that homogenisation through gene conversion frequently occurs between copies. Additionally, analyses of the large scale structural features of rDNA indicated that rDNA in human cells is regularly arranged. In contrast to previous studies suggesting that human rDNA contains many noncanonical irregular copies, such as palindromic structures (Caburet *et al.*, 2005), analyses of reads containing multiple rDNA copies showed that such arrangements were in fact extremely rare and most of the rDNA copies were beautifully tandemly aligned on the chromosomes. Even so, nanopore read length limitations prevented the true large scale structure of rDNA arrays to be explored in this study, with observations made based on reads containing only a handful of rDNA repeats (<10), rather than the capture of entire rDNA clusters. Additionally, whilst readily detectable sequence variations within the IGS were exploited in this study to evaluate sequence similarity of contiguous copies, the arrangement of rDNA alleles differentiated by subtle SNP's remain unexplored.

Whilst such studies have proven successful in providing novel insights into rDNA arrays, their internal structures and architectural variations, accurately dissecting larger rDNA arrays of greater complexity is a goal which remains unattainable with the read length limitations of current sequencing methods. Whilst rDNA arrays vary greatly in size, if we consider a large mouse rDNA array containing a 100 copies (~45 kb/copy), this would measure in at ~4.5 Mbp. Considering the extreme rarity of reads of such length and the substantial efforts required to capture them, the study of rDNA arrays in their entirety with long-read sequencing methods such as ONT Seq necessitates dramatic improvements in read length. Whilst the accuracy and read length capabilities of such technologies continues to slowly improve and until sequencing costs fall, it is critical to turn our attention to alternative approaches to address this epic challenge.

1.9 Gaps in the knowledge

In this literature review, the existing knowledge regarding rDNA and its transcriptional output has been thoroughly examined. Specifically, the role of rDNA as a regulatory target in a changing environment has been discussed, with a focus on the mechanisms which govern its expression. Secondly, the variation and repetition exhibited by this region of the genome have been considered and the challenges associated with thoroughly dissecting its genetic landscape, evaluated. Finally, ribosomal heterogeneity and the contribution of rRNA to this phenomenon have been reviewed, with particular focus on post-translation modifications which may gain ribosome's environmentally specific functioning. Even so, there remain considerable gaps in the knowledge, which this study will aim to address.

1.9.1 Large-scale arrangement of rDNA and genetic variation within clusters

Considering the sheer size and repetitive nature of rDNA clusters, these loci remain poorly understood and reference genomes remain largely devoid of entire rDNA cluster sequences (Gemmell, 2021). Due to this, there is a severe lack of understanding regarding how large-scale genetic and epigenetic processes occur at the cluster level. There remains debate as to which chromosomes house rDNA clusters, the copy number of respective loci, and the genetic variation which exists within each. Additionally, it has been shown that genetic variations between different rDNA units are not ‘silent’, rather they can act to dictate and fine-tune an organism's epigenetic response, and are differentially expressed in different cell types and under varying conditions (Holland *et al.*, 2016; Rodriguez-Algarra *et al.*, 2022). It remains unknown how these environmentally sensitive variants are arranged in rDNA loci and the interactions between them. Whilst rDNA analysis is considered incompatible with short-read sequencing methods, even with the advent of ultra-long imaging and sequencing methods rDNA remains obscure. A commendable effort by the Telomere-to-Telomer consortium has allowed for the sequencing of 2 of the smallest rDNA clusters in humans (Nurk *et al.*, 2022). However, such global and costly efforts are unfeasible for most studies. The work in this study intends to circumvent the size limitations imposed by ultra-long-read sequencing technologies and the staggering costs associated with them, by proposing an effective yet economical alternative approach.

1.9.2 Epitranscriptomic profiles of rRNA and haplotype-specific modifications

Ribosomal RNA is heavily decorated with chemical modifications, and these post-translation alterations are known to dictate its maturation and function, and directly impact protein translation. Studies concerned with rRNA modifications have exclusively assessed the collective modification profiles across transcript ensembles (Taoka *et al.*, 2018) and largely focus on mature transcripts (Stephenson *et al.*, 2022). Though the sequence of mature coding rRNAs is largely evolutionary conserved, the evidence shows that rRNA in fact exhibits high levels of sequence variation within transcribed spacers (Rodriguez-Algarra *et al.*, 2022). These elements, though not incorporated into mature ribosomes, are heavily involved in the pre-processing of rRNA transcripts, and serve to facilitate the post-translational process. There is a lack of understanding of how sequence variation within these unsuspecting RNA regions contributes to rRNA modification profiles. Considering the transcriptional outcomes of genetic variation, it is important to assess what post-transcriptional outcomes it may also dictate.

1.10 Thesis aims

Given the gaps in the existing knowledge discussed above, this thesis had the following aims which will be addressed over two research chapters:

1. To establish a methodology in which molecular DNA combing could be combined with the use of SNP-specific probes, to allow for:
 1. The isolation of entire rDNA clusters and their analysis at the single-molecule level
 2. Deducing the arrangement of rDNA promoter variants across the length of entire rDNA clusters
 3. Probing the epigenetic response of rDNA clusters in response to nutrient stress and other environmental insults
2. To establish the use of Nanopore direct RNA sequencing methods in the study of rRNA, with a specific focus on:
 1. Capturing near full-length pre-rRNA transcripts
 2. Studying the modification profiles across coding subunits within a single transcript at the single-molecule level
 3. Deduce any haplotype-specific sites of differential modification.

1.11 Thesis structure

Chapter 2 – Materials and Methods

Chapter 2 is a general materials and methods section, outlining the techniques used across all of the research chapters. Within each chapter, there is a table of materials and methods that were used in that work and the table references the relevant section and page number for the method. Much of the work in chapter 4 was the development of methods to establish ONT DRS for the capture of near full-length pre-rRNA primary transcripts, therefore optimisation steps have been largely excluded from the general materials and methods chapter but are discussed alongside the results in chapter 4.

Chapter 3 – Single-Molecule Analysis of rRNA Promoter Variants

Chapter 3 outlines the work conducted to achieve Aim 1 of this study. Specifically, the work was conducted to establish the molecular combing methodology to isolate individual DNA molecules spanning multi Mbp in length. The methodology was based on a study by Kaykov *et al.*, (2016), however substantial changes were made to the published protocol to achieve a uniform spread of consistently long DNA fibres (Kaykov *et al.*, 2016). The chapter also outlines the generation of dCas9-based SNP-specific probes intended for use on combed DNA, to visualise the arrangement of rDNA

promoter variants across entire rDNA clusters, at the single-molecule level. Additionally, the generation of chimeric cell lines in which to assess the specificity of SNP-specific probes is discussed.

Chapter 4 – Long Read Sequencing Analysis of Ribosomal RNA Modifications

Chapter 4 goes through the steps taken to apply ONT DRS to the study of ribosomal RNA. This chapter firstly focuses on the optimisation of pre-sequencing sample preparation to maximise the capture of large pre-rRNA molecules, in which multiple coding subunits occur within a single transcript. Additionally, DRS data sets from samples representative of different developmental stages, are used to assess the differential expression of rRNA haplotypes. Alongside this, the RNA modification tool ‘Nanocompore’ (Leger *et al.*, 2021) is applied to predict potential sites of differential modification within this developmental context.

Chapter 5 - Discussion and conclusions

Chapter 5 summarises the main discussion points from each of the chapters and highlights the common strengths, weaknesses, and opportunities presented by the work in this thesis. Additionally, there is a discussion of the future experiments that could be conducted to fulfill, and expand on the research goals outlined.

2 Materials and methods

2.1 Cell culture techniques

2.1.1 Mouse Embryonic Fibroblasts (MEFs)

Mouse embryonic fibroblasts (MEFs) were isolated from C57BL/6J E14.5 embryos and immortalised by Dr. Michelle Holland at the Blizard Institute, QMUL. The MEF cell lines were then initially cultured by Pui Pik law at King's College London. For culture maintenance, the cells were grown in flasks in Dulbecco's Modified Eagles Media (Gibco, Cat. 11965-092) supplemented with 3% HEPES buffer, 1% penicillin-streptomycin, and 10% fetal bovine serum. Cells were incubated at 33°C in 5% CO₂ and split every 3-4 days or when they reached ~80% confluency. For harvesting, cells were washed twice with sterile PBS and detached from the flasks via incubation with 1% trypsin-EDTA for 5 minutes. Cells were centrifuged for 5 mins at 1000 x g and cell pellets were washed with sterile PBS. The supernatant was aspirated and cell pellets were stored at -80 °C for later use. The MEFs used in this study were generally between passages 10-20.

2.1.2 Mouse Embryonic Stem Cells (MESC)s

Mouse embryonic stem cells were isolated from C57BL/6J embryos by Dr. Michelle Holland at the Blizard Institute, QMUL. The MESC)s were grown in 2i media composed of 50% neurobasal™ media (Gibco, Cat. 21103-049, and 50% DMEM/F12 GlutaMax (Gibco, Cat. 10565018) supplemented with 5 ml 50x B-27® (Gibco, Cat. 17504-044), 2.5 ml 100X N-2 (Gibco, Cat. 17502-048), 0.5 ml BSA (25mg/ml) (Sigma, A3311-10G), 2.5 ml GlutaMAX™ (35050-038), 0.25 ml Insulin 20mg/ml (Sigma, I1882-100MG), 5 ml penicillin-streptomycin, 10 ml fetal calf serum, 14 µl 1-Thioglycerol (Sigma, Cat. M6145-25ML), 50 µl PD 0325901 10mM (Axon, Cat. 1408), 150 µl CHIR 99021 10mM (Axon Cat.1386) and 15 µl LIF (Esgro, Cat. ESG1106). Cells were grown on culture dishes gelatinised with 0.1% porcine gelatine in PBS and incubated at 37°C in 5% CO₂. Media was changed daily at the same time, to ensure pluripotency and cells were cultured for 2-3 days before splitting. For harvesting, cells were washed in sterile PBS and incubated with 0.5% trypsin-EDTA for 5 minutes. Cells were centrifuged at 250 x g and split at a 1:3 ratio or stored at -80 °C for later use. Once cultures were established, cells were harvested for RNA extraction and pluripotency was confirmed via qPCR gene expression analysis in which expression of pluripotency markers NANOG, OCT4, and SOX2 was tested against relative expression in MEFs.

2.1.3 Human Lymphoblastoid Cell Line (LCLs)

Human Lymphoblastoid Cell Lines (LCL) were cultured in RPMI 1640 Medium, GlutaMAX™ (Gibco, Cat. 61870036) supplemented with 10 % FBS and 1% penicillin-streptomycin and incubated at 37°C in 5% CO₂. Cells were seeded at ~1 x10⁶ cells and grown in 5 ml suspensions and split every 3 days. To split cells, a 3rd of the cell suspension was centrifuged at 1000 x g for 5 minutes and cell pellets were resuspended in 5 ml of complete media. For harvesting, cell pellets were washed in PBS and taken forward for DNA/RNA extraction or stored at -20°C for later use.

2.1.4 Human Embryonic Kidney Cells (HEK-293)

Human Embryonic Kidney Cells (HEK-293) (Merk, Cat. c12022001) were cultured in DMEM supplemented with 2mM Glutamine, 10% FBS, and 1% penicillin-streptomycin at 37°C in 5% CO₂. Cells were seeded at an initial density of ~1 x10⁶ cells in T75 flasks and split every 3 days or until 80-90% confluent. For harvesting, spent media was removed and cells were washed with sterile PBS and detached from culture flasks via incubation with 1% trypsin-EDTA for 5 minutes. Trypsin was neutralised by the addition of complete media and cells were centrifuged for 5 mins at 1000 x g. The supernatant was aspirated and cell pellets were stored at -80 °C for later use. For splitting, cells were harvested as outlined above and split using a 1:3 ratio.

2.2 DNA and RNA techniques

2.2.1 Agilent Bioanalyzer

The integrity and size of DNA and RNA samples were assessed using the 2100 Bioanalyzer Instrument (Agilent, Cat. G2939BA) alongside either the DNA High Sensitivity Kit (Agilent, Cat. 5067-4627) or RNA 6000 Nano kit (Agilent, Cat 5067-1511) according to the manufacturer's instructions.

2.2.2 Gel electrophoresis

Unless otherwise stated, all mono-directional agarose gel electrophoresis runs were with 2% agarose gels made using Ultra-Pure agarose (Invitrogen, Cat. 16500100) in 1X TBE (10X TBE: 1 M Boric Acid, 1 M Tris, 0.02 M EDTA pH 8.0). Agarose gels were run between 100-150 V for ~30-90 mins depending on the band resolution required. Gels were pre-stained with GelRed dye (Cam Bio, Cat. 41003-BT). The DNA ladders used are stated in specific figure legends and were either 1kb Hyperladder (Bioline, Cat. BIO-33026) or 100 bp ladder (NEB, Cat. N3231S).

2.2.3 Assessing nucleic acid purity and concentration

The ND-1000 NanoDrop spectrophotometer was employed to assess the purity of DNA preparations and quantify sample concentrations. The apparatus was blanked with the appropriate buffer, and 1.5 µl of the sample was added to the instrument's detection port. Samples were considered of high purity when having A260/280 ratios of ~1.8 for DNA and ~2.0 for RNA (indicative of equal ratios of all four nucleotides and an absence of protein which absorbs at 280 nm) and a A260/230 ratio of between 2.0-2.2 (indicative of the absence of contaminants that absorb at 230 nm such as EDTA, phenol, EDTA OR EtOH).

The Qubit™ assay was employed for more accurate and sensitive measurements of DNA and RNA concentration. For measurements of DNA, either the Broad Range DNA Kit (Life Technologies, Cat. Q32850) or the High Sensitivity DNA Kit (Life Technologies, Cat. Q32851) was used according to the manufacturer's instructions. For measurements of RNA, either the Broad Range RNA Kit (Life Technologies, Cat. Q10210) or the High Sensitivity RNA Kit (Life Technologies, Cat. Q32852) was used according to the manufacturer's instructions.

2.2.4 Quantitative polymerase chain reaction

SuperScript™ VILO™ MasterMix (Invitrogen, Cat. 11755-050) was used for cDNA generation with 1 µg of DNase I treated RNA used as input, and the reaction was carried out according to the manufacturer's recommendation. The cDNA was diluted to 1/10th and used as input for qPCR reactions, prepared using PowerUp™ SYBR™ Green Master Mix (Applied Biosystem, Cat. A25741) with no more than 10% of the total reaction volume made up of diluted cDNA. Reactions were set up as per the manufacturers' guidelines and run using default parameters using Applied Biosystem StepOnePlus™ Real-Time PCR System. Primer design and specificity assessment was carried out using the online tool PrimerBlast (Ye J et al, 2012). Before use with experimental samples, primers were validated using genomic DNA by assessing 1) amplicon size via gel electrophoresis, 2) annealing specificity via melt curve analysis, and 3) primer efficiency through the generation of a standard curve using serial dilutions of template DNA. Reverse transcriptase negative reactions were carried out in parallel to control for any DNA contamination. Relative expression levels were determined using the $2^{\Delta C_t}$ formula, with all data presented after normalisation and averaging across all control genes.

2.3 Generation of control cell lines for SNP-specific probe testing

Cell lines with stably integrated A and C promoter variant sequences were generated to create an environment in which to test the specificity of SNP-specific probes individually against each variant. To this end:

2.3.1 C57BL/6J promoter variant isolation

Primers (**Table 1.1**) were designed for the PCR amplification of a 216 bp sequence of the C57BL/6J rDNA promoter. The sequence spanning positions -48 to -264 upstream of the TSS encompassed SNPs -178 and -104 and CpG -133 and was amplified with flanking restriction sites for EcoNI (5') and MfeI (3') as well as additional 3' A overhangs. The DNA used from promoter amplification was extracted from C57BL/6J liver tissue by Dr. Amy Danson at the Blizzard Institute, QMUL. The PCR reactions were prepared using PCR Master Mix (2X) (Thermo Scientific, Cat. K0171), with 50 ng of template DNA used and 0.2 μ M of each primer. The reaction was set up according to the manufacturer's instructions and run on a thermocycler using a combination of recommended reaction parameters and primer-specific annealing temperatures, run as a gradient \pm 5°C. An aliquot of each PCR reaction was separated and visualised via agarose gel electrophoresis, with only reactions bearing a single band of the desired fragment size taken forward. Successful PCR amplification reactions were purified using QIAquick PCR Purification Kit (Qiagen, Cat. 28104) as per the manufacturer's instructions, and DNA was assessed for purity and concentration using the nanodrop and Qubit DNA assay.

Target	Forward primer (5'-3')	Reverse primer (5'-3')	Amplicon size (bp)
rDNA promoter (-48 to -264 TSS)	CCTATAAAGGCAATTGAG CCCTCTCTGTCCCTGT	CAATTGCATATGACAG GCCACAGAGAATAC	235

Table 1.1- Primers used for PCR amplification of C57BL/6J promoter sequence. Forward primer is designed with additional 5' site for EcoNI digest, whilst Reverse primer is designed with additional 3' site for MfeI digest.

Promoter sequences were cloned using TOPO™ TA Cloning™ Kit (Invitrogen, Cat. K457502), with which amplicons were first ligated into vector pCR™4-TOPO™ and recombinant vectors cloned using

TOP10 Chemically Competent *E. coli* according to manufacturer's instructions. Transformed cells were cultured on LB-carbenicillin agar plates prepared with 10 mM IPTG and 2% X-gal for blue-white colour selection and incubated at 37°C overnight. 12 successfully transformed bacterial colonies were selected and grown in 5 ml of LB broth containing 100 mg/mL Carbenicillin and incubated at 37°C overnight with shaking at 225-250 rpm. An aliquot was taken from each culture to produce a glycerol stock for later use. Cultures were then processed using QIAprep Spin Miniprep Kit (Qiagen, Cat. 27106X4) according to manufactured instructions to obtain purified plasmids. Plasmid inserts were sequenced by utilising m13 amplification sites flanking the insertion site, with plasmids deposited for Sanger sequencing to Genome centre sequencing facility, Blizard institute, QMUL. Sequencing results were analysed to identify the colonies transformed with specific promoter variants A and C, and respective cultures were expanded from glycerol stocks. Cultures were expanded in 50 ml of LB + carbenicillin at 37°C overnight with shaking at 225-250 rpm. Cultures were processed using QIAprep Spin Miniprep Kit (Qiagen, Cat. 27106X4) according to manufactured instructions to obtain purified plasmids containing inserted sequences for either A or C promoter variants.

2.3.2 Lentivirus generation

Transfer plasmids pLenti-puro (Addgene, Cat. 39481) and pLJM1-EGFP (Addgene, Cat. 19319) were selected for lentiviral particle generation based on selection markers and reporter genes as well as restriction digest compatibility. 1 µg of each recombined promoter plasmid and transfer vector was individually double digested with EcoNI (NEB, Cat. R0521S) and MfeI (NEB, Cat. R3589S) in rCutSmart buffer according to manufactures instructions. Linearised plasmids were separated via agarose gel electrophoresis and transfer vector and promoter insert bands were excised and purified using QIAquick Gel Extraction Kit (Qiagen, Cat. 28706X4). DNA samples were assessed for purity and concentration using the nanodrop and Qubit DNA assay. Isolated promoter fragments A and C were assembled into vectors pLenti-puro and pLJM1-EGFP respectively using T4 DNA Ligase (NEB, Cat. M0202S) using a molar ratio of 1:3 vector to insert according to the manufacturer's instructions to yield recombined transfer vectors pLenti-puro-A var and pLJM1-EGFP-C var. For lentiviral particle generation a plasmid mix was made up separately for each promoter variant, composed of 7 µg recombined transfer vector combined with 3 µg packaging plasmid psPAX2 (Addgene, Cat. 12260) and 1 µg envelope plasmids pMD2.G (Addgene, Cat. 12259). A transfection mix was made up of each plasmid mix combined with 33 ng of polyethyleneimine (PEI) to achieve a 1:3 ratio of DNA: PEI, which was then combined with 4 ml of DMEM (-) FBS (-) Pen/Strep to produce the transfection media.

Packaging Hek-293 cells (passage 3-10), were seeded at 5×10^5 in T75 flasks a day prior and left to grow overnight, before being incubated with transduction media at 37°C in 5% CO₂ overnight. Transfection media was removed and replaced with 10 ml of complete DMEM media and cells were incubated at 37°C in 5% CO₂ overnight. Media enriched with lentiviral particles was extracted at 48 and 72 hours post-transfection, and media was replaced each time with 10 ml of complete DMEM. The collected media was centrifuged at 500 x g for 5 minutes to remove cells and passed through a 0.45 µm filter. The lentiviral physical titre of each supernatant fraction was quantitatively assessed using Lenti-X GoStix Plus (Clontech, Cat. 631280) p24-based detection, with each fraction yielding $> 5 \times 10^5$ TU/mL. The greatest detected titre was for supernatant collected at 48 hours, which was used for subsequent transductions. Two lentiviral particles termed Lenti-A var and Lenti-C var were produced and used for subsequent transductions.

2.3.3 Lentiviral transduction

HEK 293-T cells were seeded at 2.5×10^5 cells in T25 culture flasks and cultured in growth media (DMEM +10% FBS +1% Pen/Strep) and incubated at 37°C with 5% CO₂ overnight. Growth media was removed and replaced with transduction media composed of 1 ml of fresh growth media + 1 ml of respective lentivirus supernatant (supplemented with polybrene (8 µg/ml) and cells incubated at 37°C in 5% CO₂ overnight. Transduction media was removed and replaced with 5 ml of growth media and cells were incubated at 37°C in 5% CO₂ overnight. Cells were split 1:3 and grown in complete growth media for 48 hours at 37°C in 5% CO₂. After 48 hours of growth, cells were subjected to selection with the addition of 2.5 µg/ ml puromycin to growth media. The selection was maintained for a minimum of 10 days during which cells were split upon reaching 80% confluency. During this period cell growth was closely monitored alongside control cultures of non-transduced cells cultured in selection media. Puromycin resistance and GFP expression were confirmed as indicators of successful transduction with Lenti- C var and Lenti- A var respectively. Cells were grown in selection media for an additional 5 days after the complete death of control cultures after which cells were harvested for DNA extraction or stored at -80°C in complete media +10% DMSO for later culturing. DNA was extracted using DNeasy Blood & Tissue Kit (Qiagen, Cat. 69504) and promoter sequence integration was additionally confirmed via PCR amplification using promoter sequence-specific primers. All in all, HEK-293 cells were transduced with Lenti-A var, Lenti-C var and a combination of both to generate 3 distinct cells lines: 1) HEK-293 with integrated C57BL/6J A variant promoter, 2) HEK-293 with integrated C57BL/6J C variant promoter, and 3) HEK-293 with integrated C57BL/6J A and C variant promoters.

2.4 SNP-specific localisation of C57BL/6J rDNA promoter variants

Spatiotemporal allele organization by allele-specific CRISPR live-cell imaging (SNP-CLING) is a method used to detect SNP allele-specific localisation and dynamics within live cells. The method relies on probes composed of SpdCas9 nucleases bound indirectly to different fluorescent molecules complexed with guide RNAs which are used to target specific genomic loci. The method exploits the PAM specificity of dCas9 by targeting probes to SNPs that either create or disturb the PAM sequence, hence permitting or preventing probe binding and the resulting allele-specific localisation. SNP-CLING has served as inspiration for the generation dCas9 SNP-specific probes which may be used to differentiate between C57BL/6J rDNA promoter variants and visualise the arrangement of promoter variants on combed DNA.

2.4.1 Lentivirus generation

For lentiviral particle generation, a plasmid mix was made up of 7 µg pHAGE-TO-dCas9-3xGFP (Addgene, Cat. 64107), transfer vector combined with 3 µg packaging plasmid psPAX2 (Addgene, Cat. 12260) and 1 µg envelope plasmids pMD2.G (Addgene, Cat. 12259). A transfection mix was made up of each plasmid mix combined with 33 ng of **polyethylenimine (PEI)** to achieve a 1:3 ratio of DNA: PEI, which was then combined with 4 ml of DMEM (-) FBS (-) Pen/Strep to produce the transfection media.

Packaging Hek-293 cells (passage 3-10), were seeded at 5×10^5 in T75 flasks a day prior and left to grow overnight, before being incubated with transduction media at 37°C in 5% CO₂ overnight. Transfection media was removed and replaced with 10 ml of complete DMEM media and cells were incubated at 37°C in 5% CO₂ overnight. Media enriched with lentiviral particles was extracted at 48, 72, and 96 hours post-transfection, and media was replaced each time with 10 ml of complete DMEM. The collected media was centrifuged at 500 x g for 5 minutes to remove cells and passed through a 0.45 µm filter. The lentiviral physical titre of each supernatant fraction was quantitatively assessed using Lenti-X GoStix Plus (Clontech, Cat. 631280) p24-based detection, with each fraction yielding $> 5 \times 10^5$ TU/mL. The greatest detected titre was for supernatant collected at 48 hours, which was used for subsequent transductions

2.4.2 Lentiviral transduction

HEK 293-T cells were seeded at 2.5×10^5 cells in T25 culture flasks and cultured in growth media (DMEM +10% FBS +1% Pen/Strep) and incubated at 37°C with 5% CO₂ overnight. Growth media was removed and replaced with transduction media composed of 1 ml of fresh growth media + 1 ml of lentivirus supernatant (supplemented with polybrene (8 µg/ml) and cells incubated at 37°C in 5% CO₂ overnight. Transduction media was removed and replaced with 5 ml of growth media and cells were incubated at 37°C in 5% CO₂ overnight. Cells were split 1:3 and grown in complete growth media for 7 days under standard conditions during which GFP expression was confirmed as an indicator of successful transduction.

2.4.3 FACS sorting and affinity purification of SpdCas9-3xGFP

Cell cultures were harvested as previously described and subjected to FACS to select GFP-positive cells. Flow cytometry was carried out by Dr. Gary Warnes at the Blizzard Flow Cytometry Core Facility, QMUL. Sorted cells were cultured under standard conditions as previously described and expanded for 5-7 days during which GFP expression was monitored across the culture populations. Cells (10^7) were harvested as described previously and subjected to total protein extraction and affinity purification using the anti-GFP ChromoTek GFP-Trap[®] Agarose kit (Chromotek, Cat. Gta) as per the manufacturer's instructions. Purified SpdCas9-3xGFP was eluted in 200 mM glycine pH 2.5 and immediately neutralised with 1 M Tris pH 10.4.

2.5 Molecular combing

Molecular combing is a highly effective fibre stretching technique used to physically align DNA molecules onto a treated surface and is commonly used in conjunction with DNA labelling methods to visualise large-scale genomic architecture. The method can allow for the stretching of DNA measuring Megabases in length and the isolation of hundreds of copies of an organism's genome onto a single treated surface.

2.5.1 Molecular combing machine assembly

The molecular combing machine used here was assembled following directions provided by Kaykov et al., (2016). The machine setup displayed in **Figure 2.1** was assembled using precision mechanical modules sources from ThorLabs, including a 50mm motorized linear translational stage (MTS50-M-Z8) mounted on a right-angle bracket (MTS50C-Z8) and connected to a T-cube DC servo motor controller (KDC101). A removable setscrew (TR2) was screwed onto the translational stage

perpendicular to the movement axis and a post holder thumbscrew (PH2) was fitted onto the removable screw. A compact dual holder (DH1) was screwed parallel to the movement axis on the post holder thumbscrew and was used to clip glass slides. The movement was either controlled manually using the T-cube motor controller set to the appropriate speed or via the apt[™] software operated on a PC connected to the motor controller.

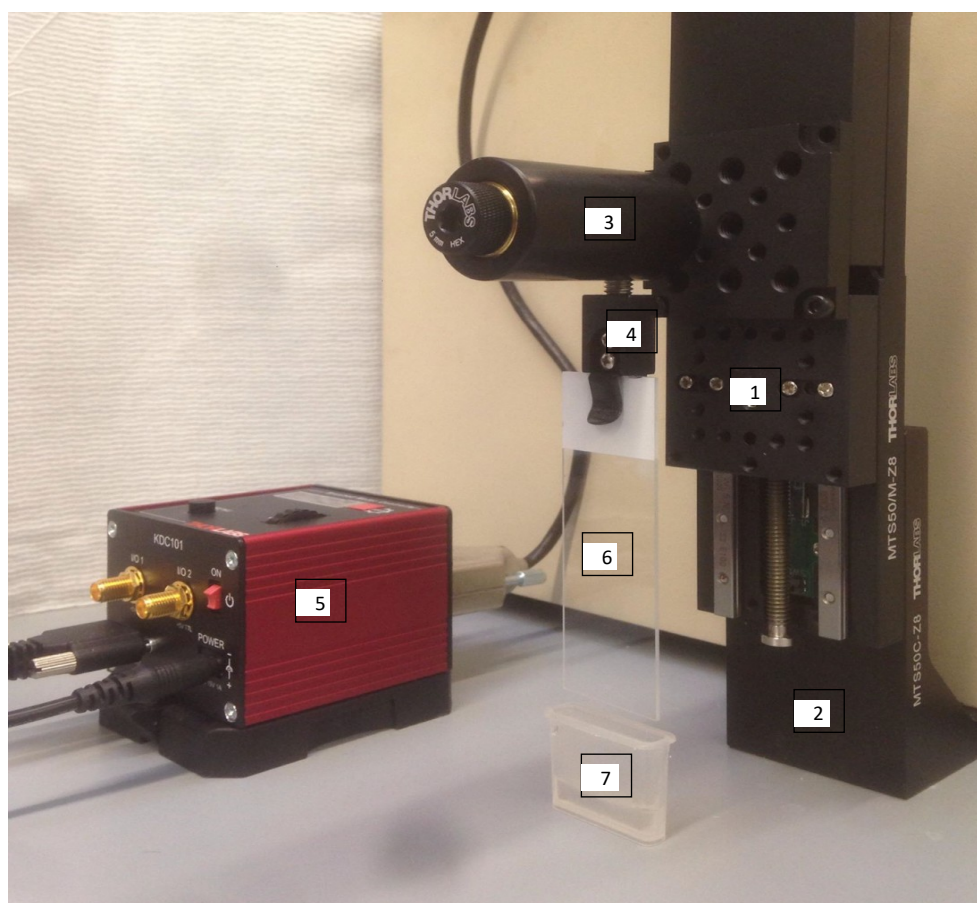


Figure 2.1 Motorised molecular combing machine. Assembled from parts purchased from ThorLabs assembled using 1) 50mm motorized linear translational stage (MTS50-M-Z8), 2) right-angle bracket (MTS50C-Z8), 3) removable setscrew (TR2) + post holder thumbscrew (PH2), 4) compact dual holder (DH1), 5) T-cube DC servo motor controller (KDC101). Combing machine was used in conjunction with 6) Silane treated slides and 7) disposable DNA combing reservoirs (Genomic Vision, Cat. RES-001)

2.5.2 Silane slide preparation

Besides manufactured Silane prep slides (Merk, Cat. S4651) used for the majority of DNA combing experiments, attempts were made to produce silanised glass slides in-house. For gas-phase silanisation with 7-Octenyltrichlorosilane the slide preparation protocol outlined in Kaykov et al. 2016 was followed. Glass coverslips (25x25 mm) (ThermoFisherSci, Cat.15522802) were placed in a ceramic holder and sonicated in chloroform (Sigma, Cat. 34854) using an ultrasonic water bath cleaner and dried thoroughly with nitrogen gas. Glass slides were activated with a benchtop plasma treater (Henniker Plasma, Model HPT-200) for 30 seconds each side and immediately placed in a vacuum desiccator (ThermoFisherSci, Cat. 5310-0250) containing 250 μ L 7-Octenyltrichlorosilane (ThermoFisherSci, Cat. H53434). A vacuum pump was used to remove air from the vacuum chamber to create an atmosphere saturated with 7-Octenyltrichlorosilane vapour. The slides were incubated in this environment for 2 hours for effective surface functionalisation and then stored in an airtight container for later use.

For liquid phase silanisation with Trimethoxy-octenylsilane, the protocol published in Labit et al. 2018 was followed. Glass coverslips (25x25 mm) (ThermoFisherSci, Cat.15522802) were rinsed in acetone and sonicated using an ultrasonic water bath cleaner, for 20 min in 50% methanol/water, and then 20 min in chloroform. Slides were dried completely with nitrogen gas and activated with a benchtop plasma treater (Henniker Plasma, Model HPT-200) for 1 minute for each side. Coverslips were placed in a sterile, dust-free slide holder and completely dehydrated in an oven at $>100^{\circ}\text{C}$ for 1 h. For surface salinisation, 100 μ L of (7-octen- 1-yl) trimethoxysilane (Sigma, Cat. 376221) was diluted in 100 mL n-heptane (Merck, Cat. 104379) and dried coverslips were rapidly submerged into silane solution. Slides were incubated in silane solution overnight whilst placed in a vacuum desiccator (ThermoFisherSci, Cat. 5310-0250) and briefly attached to a vacuum pump to remove excess air. Slides were submerged in n-heptane and sonicated for 5 minutes after which they were transferred individually into fresh distilled water and sonicated for 5 minutes each. Slides were dried with nitrogen gas and sonicated for 5 minutes in anhydrous chloroform (Merck, Cat 288306) after which they were left to dry under a vacuum. Slides were stored in an airtight container for later use.

As a measure of effective surface functionalisation, surface hydrophobicity was assessed. A drop of water was placed on the silanised slides and the configuration was observed. A water droplet forms a characteristic round drop on treated surfaces, compared to a flatter, spread-out configuration on an untreated surface.

2.5.3 High Molecular Weight (HMW) DNA extraction

The protocol for HMW DNA extraction was adapted by Amy Francis from a protocol developed by Josh Quick for the RAD004 library preparation kit (Jain et al., 2018b). A frozen pellet of MEFs containing six T75 flasks was used for each extraction ($\sim 30 \times 10^6$ cells). The pellet of frozen cells was thawed at RT and resuspended in 100 μ l of PBS and the sample briefly vortexed. Resuspended cells were mixed with 5 ml of cell lysis buffer TLB (100 mM NaCl, 25 mM EDTA pH 8.0, 0.5% [w/v] SDS, 10 mM Tris-Cl pH 8.0) was + RNase A (20 μ g/ml) and the solution was incubated at 37 °C for 30 mins. Once the solution was clear, proteinase K was added at a concentration of 200 μ g/ml and the solution was gently mixed via slow and controlled inversions. The solution was further incubated at 50 °C for 90 minutes with regular inversions every 30 mins. 15 ml falcon tubes were pre-prepared with light phase-lock gel (5PRIME, Cat. 2302820), with 3 aliquots of light phase-lock gel centrifuged into each 15 ml falcon tube, one at a time. Upon complete incubation, the sample solution was visually confirmed to be 'gloopy' and was decanted into the pre-prepared 15 ml tube containing the phase lock gel. This was combined with 5 ml of TE-saturated phenol (Sigma Aldrich, Cat. 77607) and the tube was mixed gently on a HulaMixer set at 20 rpm for 10 mins to obtain a fine emulsion. The mixture was then centrifuged for 10 mins at 4500 rpm. The phase lock gel promoted the separation of the phases and the clear aqueous phase settled at the top was poured off into a second pre-prepared 15 ml tube containing light phase-lock gel. This was combined with both 2.5 ml of TE-saturated phenol and 2.5 ml of chloroform-isoamyl alcohol (Sigma Aldrich, Cat. 25666). The mixture was rotated on a HulaMixer for 10 mins at 20 rpm and then centrifuged for 10 mins at 4500 rpm. The aqueous phase settled at the top and was poured into a 50 ml falcon tube. To this, 15 ml of ice-cold 100% EtOH and 2 ml of ammonium acetate (VWR, Cat. 437453A) were added. The tube was incubated at RT for several minutes to allow for the precipitation of DNA. For retrieval of DNA from the solution, a glass pipette was melted at the end using a Bunsen burner to create a hook and was used to 'fish' out the DNA. The DNA aggregate, still attached to the glass hook was submerged in 80% ethanol to promote its condensation into a white clump, this was then left to air dry before being submerged in 50 μ l of 10 mM Tris to dissolve. The extracted DNA was heated at 65 °C for 10 mins before being allowed to slowly dissolve at 4 °C for a minimum of 2 days. A 1 μ l aliquot of the sample was then diluted 1:10 for measurement of concentration using the Nanodrop and QuBit BR dsDNA assay. Once the quality of the DNA and the concentration were established, the DNA was diluted to $\sim 1 \mu$ g/ μ l and stored at 4 °C for later use. For size assessment, a 1 μ g/ μ l aliquot of HMW DNA was diluted with gel loading dye and added to a 1% gel made with 0.5X TBE for PFGE. The gel was run for 48 hours at 4 V/cm with switch times between 5-120 seconds and an included angle of 120°. A lambda phage ladder (NEB, Cat. N0341S) was run along the gel as a size marker (Figure 2.2).

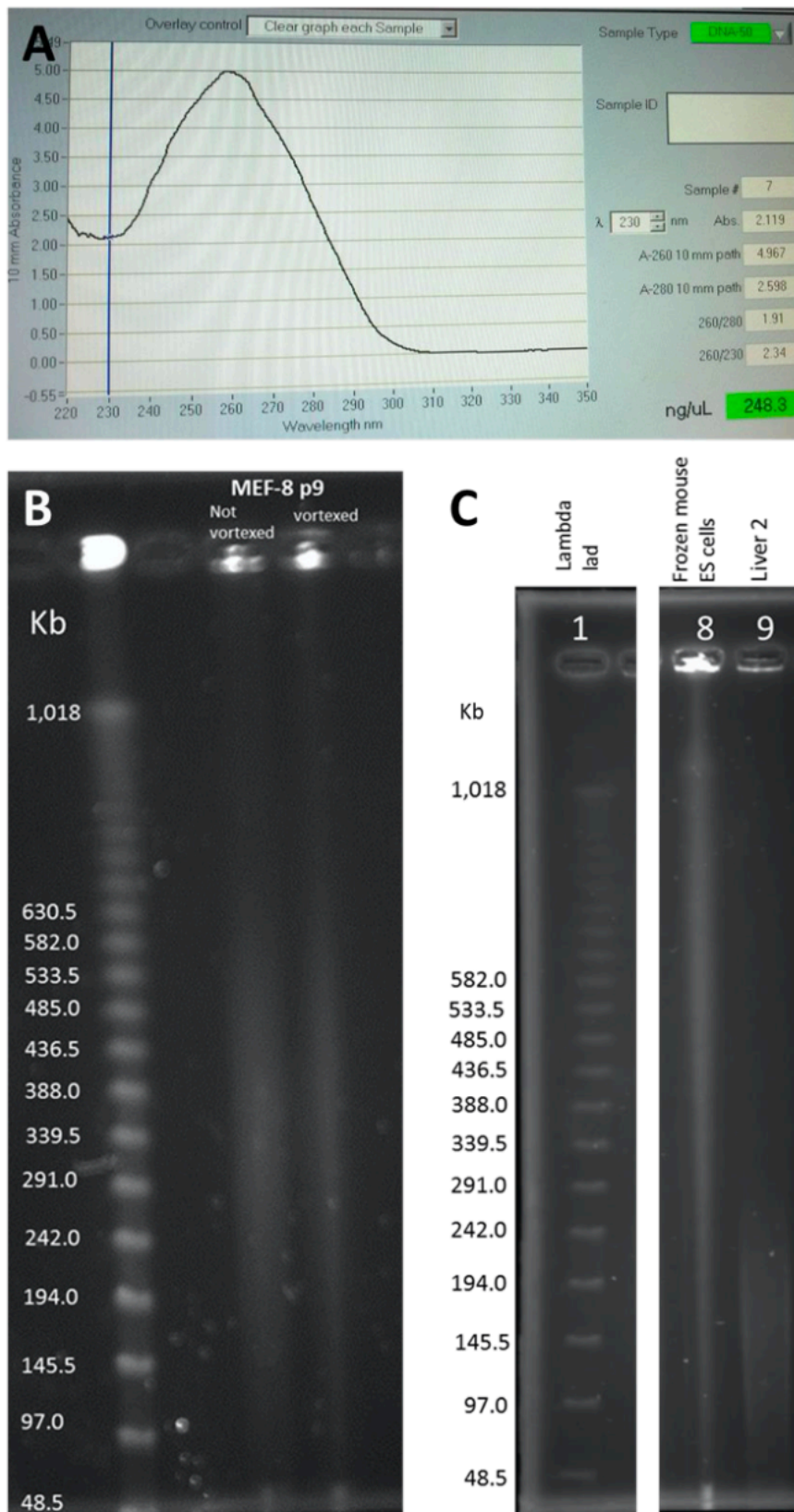


Figure 2.2 Quality control of high molecular weight DNA extractions. A) A representative Nanodrop reading for high molecular weight DNA extraction from passage 9 of the MEF-8 cell line (MEF-8 p9) diluted 1:10. B) Pulsed-field gel of the same sample, also showing minimal difference in the size distribution between the sample before and after vortexing. C) An example of a successful HMW DNA extraction from frozen cells. Fragments larger than 1 Mb gather at the resolution limit at the top of the gel. Lane 9 shows a poor extraction with the median length below 100 Kb.

2.5.4 Ultralong DNA preparation in agarose plugs

A frozen pellet of MEFs containing $\sim 5 \times 10^6$ cells was thawed at RT and resuspended in 1 ml of sterile PBS. The cell solution was centrifuged at 1000 x g for 3 minutes and the supernatant was discarded. The cell pellet was resuspended in 1 ml of CSB (10mM Tris-HCl pH=7.5, 20mM NaCl, 50mM EDTA) and centrifuged at 1000 x g for 3 minutes and the supernatant was discarded. The cell pellet was resuspended in 1 ml of CSB and filtered using a 70 μ m separation filter (MiltenyiBiotec, Cat. 130-095-823) to remove cell aggregates and ensure a single cell suspension. The cell suspension was diluted for the various cell densities tested, with each separate suspension containing 2 times the required cell density. An equal volume of 2% solution of low melting point agarose Mbp grade (BioRad, Cat. 1613108) dissolved in CSB with added 0.2% NaN_3 was added to each cell suspension and equilibrated at 45 °C. The agarose-cell solution was gently but thoroughly mixed with a wide bore pipette tip and 80 μ l added to each well in a strip of agarose plug molds (BioRad, Cat. 1703713) and then incubated at 4°C for 10 minutes to solidify. Agarose plugs were ejected into DB (1mg/ml Proteinase K, 1% *N*-Laurouylsarcosine, 0.2% Na Deoxycholate, 100mM EDTA, 10mM Tris-HCl pH 7.5) and incubated at 50°C for 1 hour. The DB was changed, and the plugs were incubated at 50°C for a total of 48 hours with the DB refreshed 4 times throughout. For iterations where agitation was introduced during the digestion step plugs in DB were agitated on a HulaMixer set at 10 rpm. The plugs were washed for 48 hours in WB (TE 1X pH 7.5 + 100mM NaCl) with WB changed 4 times throughout and samples agitated on a HulaMixer set at 10 rpm. The plug was melted in 2 ml of MES 50mM (pH 5, 5.5, 6 and 6.5) +100mM NaCl, directly in disposable DNA combing reservoirs (Genomic Vision, Cat. RES-001) for 15 minutes at 70°C. For iterations where plug melting time was increased, plugs were melted for 30 mins at 70°C. Samples in combing reservoirs were cooled to 42°C before the addition of 2 μ l β -Agarase I (New England BioLabs, Cat. M0392) and incubated overnight or for 24 hours at 42°C without mixing. Once agarose plugs were melted and DNA liberated, measures were taken to limit any mechanical disturbance to samples to preserve DNA molecule length.

2.5.5 Molecular combing

Combing reservoirs containing DNA solutions were allowed to cool down to RT and DNA was combed onto silanised glass surfaces with the assembled combing machine at a speed of $\sim 300 \mu\text{m}/\text{sec}$. Measures were taken to limit any mechanical disturbance to samples to preserve DNA molecule length. The combed DNA was dehydrated by incubating slides at 65°C for 2 hours and either stained immediately or stored at -20°C for later use. To assess the quality and size of combed DNA, a handful of slides were chosen per batch and stained with YOYO™-1 Iodide (Invitrogen, Cat Y3601) prepared in antifade mounting media (Invitrogen, Cat P10144) at a 1:10,000 ratio. Slides were visualised with Leica

DM4000 Epi-Fluorescence Microscope using appropriate excitation and detection parameters.

2.5.6 Fibre length assessment

A Leica DM4000 Epi-Fluorescence Microscope was used along with the appropriate excitation (491 nm) and detection parameters for preliminary visualisation of combed DNA fibres during protocol development. Fibre lengths were crudely quantified using imageJ measurement toolbox. For a more accurate quantification of full length DNA fibres, combing slides were imaged using the TissueFAXS PLUS upright scanning fluorescence and brightfield system. A 25x25 mm field of view was assigned and images acquired using 63x Oil objective. Prior to imaging, DNA fibres were manually located and z-axis assigned based on fluorescence detection value of individual fibre. The xy-axis scanning strategy was set to 'snake' and tiled images captured with default overlap. Tiled images were stitched using internal stitch settings or an overlap of 15% to minimise adjacent image infringement. Post-acquisition, fibre lengths were measured using TissueFAX measurement toolkit.

2.6 DNA labelling techniques

2.6.1 FISH Probe Synthesis

To visualise major ribosomal DNA elements, 18s, 5.8, and 28s sequences were selected as appropriate hybridisation targets and DNA templates for probe generation were produced via PCR amplification of these target sites. Primers were designed to allow for amplification of a single amplicon encompassing the majority of the 18S sequence (~1.7 kb) whilst another amplicon was generated spanning the majority of the 5.8- 28S sequences (~5.8kb) (**Table 2.1**). For each amplification reaction, 10 ng of template DNA was used, combined with PCR Master Mix (2X) (ThermoFisherSci, Cat. K0171) and 1 μ M of each forward and reverse primer as well as the addition of DMSO (5% of total reaction volume). The amplification reaction was set up according to the manufacturer's protocol with alterations made according to the primer-specific annealing temperatures and amplicon size appropriate extension times.

Target amplification was confirmed via gel electrophoresis, and PCR amplicons were purified using the QiAquick PCR purification kit (Qiagen, Cat 28104). DNA purity was assessed via the nanodrop assay before being used for probe generation. Probes were generated using a nick translation kit (Abbot, Cat Cat: 07J00-001) according to the manufacturer's protocol. Approximately, 1 μ g of purified PCR amplicon was used as a template for probe synthesis with 18S probes labelled with SpectrumGreen dUTP and 5.8S-28S probes labelled with SpectrumRed dUTP. The third probe against mouse chromosome 12 was similarly synthesised using 1 μ g of mouse chromosome 12 BAC clone 23-225M6

(Empire Genomics, Cat. 23-225M6) labelled with SpectrumGreen dUTP.

Target	Forward primer (5'-3')	Reverse primer (5'-3')	Amplicon size (bp)
18s	CGCACGGCCGGTACAGTGAA	CGTCTTCTCAGCGCTCCGCC	1731
5.8s - 28s	GCGGTGGATCACTCGGCTCG	GACGAACGGCTCTCCGCACC	5768

Table 2.1 Primers used for PCR amplification of C57BL/6J rDNA templates for FISH probe synthesis

2.6.2 Metaphase chromosome spread preparation

MEFs grown in T75 flasks were cultured to ~90% confluency in complete media. Around 4 hours before harvesting, the culture media was spiked with colcemid (0.1 µg/ml) (Merck, Cat. 234109) and 2 hours before harvesting ethidium bromide (0.1 µg/ml) (ThermoFisherSci, Cat. 15585011) was also added. After incubation, media was aspirated and cells were carefully rinsed with sterile PBS. Cells were treated with 0.5 ml 1% trypsin/EDTA (ThermoFisherSci, Cat. R001100) and closely monitored under a microscope for detachment. To enrich for cells in metaphase, only the first 3rd of cells to detach were selected and were quickly isolated into a 15 ml falcon tube. Cells were spun down at 500 x g for 5 minutes and the supernatant was carefully decanted. The cell pellet was resuspended in the residual supernatant and 1 ml of pre-warmed (37°C) hypotonic solution (0.56% KCl in distilled water) was added which cells were again gently resuspended. An additional 9 ml of hypotonic solution was added and cells were incubated in a water bath at 37°C for 15 minutes. During the incubation, the cell solution was slowly inverted every 5 minutes to ensure suspension. After this point, cells were treated with extreme care to prevent them from bursting and also placed in ice when possible. Upon incubation, 1 ml of freshly prepared ice-cold fixative (methanol: glacial acetic acid, 3:1) was added and the solution was centrifuged at 300 x g for 8 minutes. The majority of the supernatant was carefully removed, leaving ~0.5 ml supernatant behind, in which the cell pellet was resuspended via gentle flicking. Ice cold fixative was added in a drop-wise manner to bring the total volume to 2.5 ml. The cell suspension was resuspended via slow pipetting and centrifuged at 300 x g for 8 minutes. The supernatant was carefully decanted and 2 ml of ice-cold fixative was added in a drop-wise manner. The cell solution was carefully resuspended via slow pipetting and incubated overnight at 4°C. The following day, cells were gently resuspended via flicking and centrifuged at 300 x g for 8 minutes. The cell pellet was carefully resuspended in 2 ml of freshly prepared ice-cold fixative and centrifuged at 300 x g for 8 minutes. The cell pellet was resuspended in 0.5 ml of ice-cold fixative and stored at 4°C until later use.

In preparation for chromosome spreading, a clean microscope slide (Merck, Cat. CLS294775X50) was briefly placed on ice and once ready for use was breathed upon to produce condensation on the surface. Quickly, 30 µl of fixed metaphase cell solution was taken up with a pipette tip and ejected onto the slide from an approximate distance of 50 cm. A single 'drop' of metaphase cells was used per slide to avoid overcrowding and allow chromosomes to adequately disperse. Sample slides were rapidly airdried and mounted in media containing DAPI (Abcam, Cat ab104139) to assess the degree and quality of chromosome spreads. Unstained metaphase chromosome spreads from successful batches were stored at -20°C for later use.

2.6.3 DNA Fluorescence *in situ* Hybridisation

For FISH of metaphase chromosomes and combed DNA fibres, pre-prepared slides were desiccated under vacuum for 2-3 days. Once adequately dry, slides were incubated for 1 hour at 37°C in a sterile glass Coplin jar (Merck, Cat. S5516) containing a 50 ml solution of RNase (20mg/ml) (ThermoFisherSci, Cat. 12091021) prepared in 2xSSC (Merck, Cat. S6639). Slides were quickly washed in 2xSSC at RT and dehydrated with a series of EtOH washes (70%, 90%, and 100 %) for 2 minutes each. Slides were dried under a vacuum for 15 minutes before being heated to 70°C in an oven for 5 minutes. Slides were quickly submerged in pre-warmed denaturing solution (70% formamide/2xSSC) (70°C) and incubated for 5 minutes at 70°C. Slides were quickly transferred to ice-cold 70% EtOH for 3 minutes and successively in 90% and 100% EtOH for 2 minutes each. Slides were dried under a vacuum for 15 minutes and then warmed to 37°C on a hot plate. Alongside slide preparation, DNA probes were prepared for hybridisation, with 50 ng of labelled probe per slide precipitated in a solution containing 1 µl salmon sperm DNA (10 mg/ml) (ThermoFisherSci, Cat. 15632011), 0.3 µl yeast tRNA (10µg/ml) (ThermofisherSci, Cat. AM7119), 1/10 (v/v) Sodium Acetate (3M, pH5.2) and 3x volumes of 100% ethanol. The solution was centrifuged at 8,000 x g for 30 minutes at 4°C. The DNA pellet was carefully washed with ice-cold 70% ethanol and centrifuged at 8,000 x g for 5 minutes at 4°C. The supernatant was carefully decanted and the pellet dried on a speedy-vac at RT for 15 minutes or until completely dry. The pellet was resuspended in 6 µl of deionised formamide (VWR, Cat. A2156.1000) per slide and incubated in a tabletop thermomixer at 42°C rotating at 14,000 RPM for 30 minutes. Probes were then denatured at 75°C for 7 minutes and placed on ice immediately for later use. Probes were mixed with 6 µl of 2x hybridisation buffer per slide (20% (v/v) 20xSSC, 20% (v/v) 50% dextran sulfate, 20% (w/v) BSA, 20% (v/v) H₂O. To each slide, 10 µl of the prepared probe was added after which a clean coverslip was positioned on top and the edges sealed with CoverGrip™ (Biotium, Cat. 23005). Slides were sealed in a moist box sealed with parafilm to prevent evaporation and incubated at 37°C overnight. The following day, coverslips were gently removed by submerging in 2xSSC and then washed, first in prewarmed 50% formamide/2xSSC for 5 minutes at 42°C and then 3 times in 2xSSC for 5 minutes at 42°C. Slides were manually and continuously agitated during wash steps. Slides were then mounted with media containing DAPI (Abcam, Cat ab104139), covered with glass coverslips, and the edges sealed with CoverGrip™ (Biotium, Cat. 23005). The slides were visualised immediately with Leica DM4000 Epi-Fluorescence Microscope using appropriate excitation parameters or stored at -20°C for later assessment.

2.7 Mouse Embryoid Body formation

2.7.1 Mouse Embryonic Stem Cell differentiation

Feeder-free MESCs maintained in 2i media were differentiated in EB media formulated with 50% neurobasal™ media (Gibco, Cat. 21103-049, and 50% DMEM/F12 GlutaMax (Gibco, Cat. 10565018) supplemented with 5 ml 50x B-27® (Gibco, Cat. 17504-044), 2.5 ml 100X N-2 (Gibco, Cat. 17502-048), 0.5 ml BSA (25mg/ml) (Sigma, A3311-10G), 2.5 ml GlutaMAX™ (35050-038), 0.25 ml Insulin 20mg/ml (Sigma, I1882-100MG), 5 ml penicillin-streptomycin, 10 % fetal bovine serum and 14 µl 1-Thioglycerol (Sigma, Cat. M6145-25ML). Cells were seeded at varying cell densities (250, 500, 1000, and 2000 cells) in a Corning® Costar® Ultra-Low Attachment Multiple Well Plate (Merck, Cat. CLS7007) and incubated at 37°C in 5% CO₂. Cells were allowed to settle at the bottom of the conical-shaped wells for 24 hours without being disturbed to promote the cellular association. After this, EB media was replenished every day for 7 days, and culture morphology was monitored under light microscopy. Once the cells had amassed into distinct spheroids, they were transferred individually into single wells of a 24-well plate using a wide-bore pipette tip to minimise mechanical disruption. Wells were either directly gelatinised with 0.1% porcine gelatine in PBS or fitted with pre-gelatinised glass coverslips. Spheroids were allowed to attach to the gelatinised surfaces and incubated undisturbed at 37°C in 5% CO₂ for 48 hours. After 48 hours of incubation, EB media was replenished every day for 14 days during the differentiation period and cells were closely monitored as they spread out of the spheroid and across the surface. After 14 days of differentiation, cells were harvested via incubation with 1% trypsin-EDTA for 5 minutes and centrifuged at 500 x g. Cell pellets were stored at -80°C until later use.

2.7.2 Embryoid Body germ layer validation with qPCR and immunofluorescence

Embryoid bodies grown directly on culture wells were harvested as described above and subjected to total RNA extraction using a total RNA extraction kit (NEB, Cat. T2010S) as per the manufacturer's instructions. RNA was used in the preparation of cDNA for qPCR gene expression analysis, in which expression of germ layer markers, Sox17 (endoderm), Brachyury (T) (mesoderm), and Otx2 (ectoderm) was compared between EBs and MESCs. Primer sequences for markers used for qPCR validation are presented in **Table 2.2**. Validation of EB germ layer development was also assessed with Immunofluorescence. For this, EBs were permitted to differentiate on gelatinised coverslips for 14 days as described above. Coverslips were carefully removed and rinsed with DPBS. Surface adhered cells were fixed in 4% PFA for 20 minutes at RT and rinsed in DPBS three times for 5 minutes each. Cells were incubated for 30 minutes at RT in a blocking buffer composed of PBS supplemented with, by volume, 0.5% goat serum, 3% 0.01x-Triton X-100, and 2% BSA (50 mg/mL). Primary Rabbit

Polyclonal antibodies against germ layer markers smooth muscle actin (Proteintech, Cat. 14395-1-AP), β -tubulin (Proteintech, Cat. 10094-1-AP), and GATA-4 (Proteintech, Cat. 19530-1-AP) were prepared individually in blocking buffer at a 1:100 dilution and incubated with cells overnight at 4°C. The next day cells were carefully rinsed with DPBS three times for 10 minutes each at RT. Cells were treated with Goat anti-Rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody-Alexa Fluor™ 488 (Invitrogen, Cat. A-11008) prepared in blocking buffer at a 1:1000 dilution. Cells were incubated for 1 hour at RT after which they were rinsed in DPBS twice for 10 minutes at RT. Cells were then mounted with media containing DAPI and visualised immediately with Leica DM4000 Epi-Fluorescence Microscope using appropriate excitation parameters.

Marker	Germ layer	Forward primer (5'-3')	Reverse primer (5'-3')
Sox17	Endoderm	ATACGCCAGTGACGACCAGAG	ACCACCTCGCCTTTCACCTTTA
Brachyury (T)	Mesoderm	TCTCTGGTCTGTGAGCAATGGT	TGCGTCAGTGGTGTGTAATGTG
Otx2	Ectoderm	GCGAAGGGAGAGGACGACTTT	CTGCTGTTGGCGGCACTTAG

Table 2.2- Embryoid Body germ layer marker qPCR validation primer sequences

2.8 Ribosomal RNA processing inhibition

Here, ribosomal RNA precursor processing was inhibited to preserve full-length rRNA precursor transcripts for nanopore direct RNA sequencing. Several chemotherapeutic drugs have been shown to exert their effects via the inhibition of ribosome biogenesis, mechanistically underpinned by the inhibition of ribosomal RNA processing. Two such compounds, Flavopiridol and 5-Fluorouracil are shown to specifically disrupt early and late rRNA processing respectively, as well as ultimately impacting cell cycle progression.

2.8.1 Drug treatment

Serial dilutions of compounds, 5-Fluorouracil (5-FU) (Merck, Cat. F6627) and Flavopiridol (BioVision, Cat. 2090-5) were made up independently in DMSO to achieve 20 mM working stock solutions. Cells were seeded at cell type appropriate densities and allowed to grow overnight under standard cell-specific conditions. Aliquots of cell-specific media were made up of a range of drug concentrations (5-Fluorouracil: 25, 50, 100, or 200 μ M or Flavopiridol: 0.25, 0.5, 1, 2 μ M) alongside vehicle controls for each cell type used. The levels of DMSO were kept constant across all conditions at 0.05% of the total culture media volume. Cells were incubated in drug spiked media for 24 hours under conditions

specific to each cell line and each condition run in biological triplicates. Cells were harvested as previously outlined during which they were partitioned into three fractions, with each 3rd used for either FACS cell cycle analysis, qPCR expression analysis, or nanopore sequencing.

2.8.2 Propidium Iodide staining and FACS cell cycle analysis

For cell cycle analysis, drug-exposed cells were harvested as previously described and cell pellets washed in sterile PBS. Cells were fixed in freshly prepared ice-cold 70% EtOH, and added drop-wise with constant vortexing to ensure effective resuspension. Cells were incubated in fixative overnight at 4°C after which they were pelleted at 500 x g for 5 minutes. Cells were washed twice in PBS and re-pelleted, after each wash the supernatant was carefully decanted to prevent loss of sample. Cells were incubated with 50µl of RNase A (20 mg/ml) (Merck, Cat. R5125) for 15 minutes at 37°C before the addition of 250 µl of staining solution composed of Propidium iodide (Pi) (50 µg/ml) (Merck, Cat. P1470) in PBS.

Cells were sorted using the ACEA Novocyt 3000 in conjunction with NovoExpress flow cytometry software (version 1.5.6) to make measurements of cellular PI-DNA. Apparatus was set to excitation with 488 nm laser and sample collection set to 30,000 individual events with a medium flow rate of 35 µl/min. Measurements were made using the default PI photodetector gain settings (520) qPCR assessment with a primary detection threshold set to 100,000. Cell cycle data were collated as an average of biological triplicates. Special attention was given to the forward scatter vs. side scatter, which determines single cells from doublets, pulse area vs. pulse width, and cell count vs. propidium iodide.

2.8.3 qPCR validation of rRNA precursor processing

Here, qPCR was used to assess the levels of intact rRNA precursor molecules, with amplification of intact ITS regions used as an indicator of this. Primers were designed to encompass known cleavage sites within the rRNA transcript, specifically located within ITS¹ and ITS². Cells were subjected to drug treatment and cell cycle analysis as outlined above and total RNA was extracted using a commercially available kit (NEB, Cat. T2010S). RNA was processed for qPCR analysis and amplification across the target sites monitored using SYBR[™] Green detection. The amplification of intact ITS cleavage sites was measured alongside total rRNA precursor expression by using levels of 5'ETS with data normalised to genes selected for non-variable expression from the RNA-seq data MAPK1 and ITGB1. All primers used in the assessment are displayed in **Table 2.3**.

Target	Forward primer (5'-3')	Reverse primer (5'-3')
ITS ¹	GCGGAAGGATCATTAACGGG	TCACCTCACTCCAGACACCT
ITS ²	GACACTTCGAACGCACTTG	TGCAGGACACATTGATCATCGA
5'ETS	GGTTGAGGGCCACCTTATTT	GGAAGAAAGACCGGGAAGAG
MAPK1	CACCAACCTCTCGTACATCG	AGGTCTGGTGCTCAAAAGGA
ITGB1	GGTTTCCTGGATTGGATTGA	ACATTCTCCGAAGATTGG

Table 2.3- qPCR primers used for rRNA precursor processing assessment

2.9 Oxford Nanopore Sequencing methods

2.9.1 Oxford Nanopore Sequencing Technology

Oxford Nanopore Sequencing Technology is a 3rd-generation sequencing technology that uses biological pores inserted into an insulated membrane, over which an electrical potential is applied. Each nucleic acid molecule is individually threaded through a pore, with each nucleotide and base modification characteristically altering the current. The change in current signal across the membrane alongside information such as pore dwell time is in turn used to infer the nucleotide sequence and any potential modifications. The recent development of Nanopore technology now permits the sequencing of native RNA without the need for prior amplification, cDNA synthesis, or fragmentation, permitting the direct identification of RNA base modifications alongside the nucleotide sequence of full-length transcripts.

2.9.2 Total and nuclear RNA extraction

Total RNA was extracted from either frozen cell pellets or freshly harvested cell culture material using a commercial total RNA extraction kit (NEB, Cat. T2010S) as per the manufacturer's instructions. For extraction of nuclear RNA, nuclei were first isolated by incubation of material in nuclei isolation medium (NIM) containing 250 mM sucrose, 25 mM KCl, 5 mM MgCl₂, 10 mM Tris-EDTA (pH 8.0), 1 μM DTT, 1x protease inhibitor (Promega, Cat. G6521), 0.4 U/μl RNaseIn (Promega, Cat. N2111), 0.2 U/μl Supersasin (Invitrogen, Cat. AM2694). Cultured cells were incubated in NIM on ice for 2 minutes and the cell membrane was disturbed via pipetting. Frozen tissue stored at -80°C was briefly thawed on wet ice and equilibrated in NIM. Tissue material was physically broken apart on wet ice using a precooled Dounce homogeniser and pestle, with 5 strokes of the loose pestle and 10-15 strokes of the tight pestle, or until the solution became completely homogenous. In all cases, the homogenate was passed through a 30 μM cell strainer to remove large debris. Isolation of nuclei was visually confirmed

under light microscopy, and a small aliquot stained with trypan blue was used to quantify nuclei number. The filtered homogenate was centrifuged at 1000 x g for 8 minutes to pellet nuclei and the supernatant was discarded. Nuclei were then processed for RNA extraction using a commercial total RNA extraction kit (NEB, Cat. T2010S) as per the manufacturer's instructions. All RNA samples were subjected to column-based DNA removal and DNase digestion during the kit procedure. The purity and concentration of extracted RNA samples was primarily assessed using the Nanodrop spectroscope and Qubit RNA assay respectively, with RNA integrity assessed using the Agilent RNA bioanalyzer.

2.9.3 Pre-processing of RNA for Nanopore library preparation

Approximately 10 µg of extracted RNA were size-selected using SPRI RNAClean XP beads (Beckman, Cat E7490S) using a 0.35:1 bead to sample ratio and eluted in a minimal volume of nuclease-free water (~20 µl). Bead size selection was assessed using gel electrophoresis by comparing smear and banding patterns of input RNA to eluted size selected fractions. Size selected RNA was *in vitro* poly(A) tailed using E. coli poly(A) Polymerase and ATP donor as instructed by a commercial kit (NEB, Cat. M0276), with 5 µg of RNA input into each 20 µl reaction. Samples were then purified using SPRI RNAClean XP beads (Beckman, Cat E7490S) and further size selected using a 0.35:1 bead to sample ratio. To ensure the removal of any non-poly(A) RNA, samples were then subjected to poly(A) selection using NEBNext® poly(A) mRNA Magnetic Isolation Module (NEB, Cat. E7490S) as per the manufacturer's instructions. The purity and concentration of extracted RNA was primarily assessed using the Nanodrop spectroscope and Qubit RNA assay respectively, with RNA integrity assessed using the Agilent RNA bioanalyzer.

2.9.4 Nanopore cDNA and Direct RNA library preparation

ONT library preparation was performed as per the manufacturer's instructions using either the direct cDNA sequencing kit (SQK-DCS109) or the direct RNA sequencing kit (SQK-RNA002). The RNA input was altered depending on the flow cell used, with libraries run on Flongle flow cells generally requiring 250 ng RNA, half the amount of RNA starting material compared to both MinION and PromethION flow cells (500 ng). Library concentration before loading was determined either with the Qubit DNA or RNA High Sensitivity assays with total yields generally between 20-30 ng for Flongle flow cell preparations and 40-60 ng for MinION and PromethION flow cells. For loading flow cells, the manufactures guidelines were followed without any notable changes.

2.9.5 Nanopore cDNA and Direct RNA Sequencing data analysis

Raw FAST5 files were base called using ONT Guppy (version 4.2.2) with read quality threshold set to QScore > Q7 and fastq reads generated. Reads with a quality score of greater than the threshold were classified as 'pass' and those that were assessed to fall below were classified as 'fail', with only passed reads being used for subsequent analysis. Passed reads were mapped to the genome using Minimap2 (Li, 2018), and aligned to reference genome assembly GRCm38 (mm10). Reads mapping to rDNA (reference BK000964.3) were extracted from those mapping to the rest of the genome for downstream analysis.

2.9.6 Ribosomal RNA modification calling with Nanocompore

For modification detection of nanopore direct RNA sequencing libraries, the community-developed software 'Nanocompore' was employed (Leger et al., 2021). Publicly available code in conjunction with experimental-specific scripts was used for analysis. Before modification calling with Nanocompore, data were pre-processed as follows. Raw fast5 files were first base called using ONT Guppy (version 4.2.2) with read quality assessed and reads classified as either 'pass' or 'fail' determined by the pre-set quality threshold, with all passed fastq files per library concatenated into a single fastq file. Base called reads were then mapped to the transcriptome using Minimap2 (version 2.16) using reference genome assembly GRCm38 (mm10). Samtools (version 1.9) was then used to filter bam files to remove any unmapped, secondary, or supplementary reads, as well as those mapping on the reverse strand. Additionally, reads with an alignment score lower than MAPQ<10 were discarded, with the remaining reads sorted and indexed. Next signal level analysis was carried out using nanopolish (version 0.10.1) to calculate an improved consensus sequence and realign the raw signal to the expected reference sequence. To this end, fast5 and fastq files were first indexed with nanopolish index, and bam files indexed using Samtools. Nanopolish eventalign was then used to re-squiggle reads, in which raw electrical signal information was re-aligned to the reference, this time considering low-level signal deviation information to detect shifts in the signal that may be attributed to potential modifications. Finally, pre-processed data was then processed using Nanocompore eventalign_collapse (version 0.6.2) which was used to collapse the data per k-mer to generate tabulated files containing realigned median intensity and dwell time values for each k-mer of each read. Data was then processed using a Nanocompore sample comparison module with coverage per position down-sampled to no more than 2500.

3 Single-Molecule Analysis of rDNA Promoter Variants

3.1 Introduction

3.1.1 Aims

The work in this chapter stems from a recent study describing the existence of two distinct genetic variants of rDNA in C57BL/6J mice that demonstrate differential environmental sensitivity resulting in lifelong epigenetic regulation and phenotypic differences (Holland *et al.*, 2016). The study by Holland *et al.* defined two rDNA variants differentiated primarily by SNPs within the promoter sequence at position -104 upstream of the transcriptional start site. Specifically, an A variant is defined by an adenine at position -104 whilst a C variant is defined by a cytosine at this position. The study explored variant-specific epigenetic dynamics in response to protein restriction during early development. Within this experimental context, hypermethylation of a CpG site at position -133 within the rDNA promoter was observed for the A variant in contrast to diminished methylation of the C variant at this site, with methylation levels negatively correlating with weaning weight. The molecular and functional significance of CpG -133 has been well characterised, with studies confirming that methylation at this site hinders the binding of the POL 1 basal transcription factor UBF, the binding of which is necessary for the expression of 45S rDNA (Santoro and Grummt, 2001; Grummt and Pikaard, 2003). Overall, this study demonstrated that rDNA genetic variation can lead to differential environmental reactivity and adaptation, establishing rDNA as a genomic target for nutritional insults. Even so, it remains unknown how rDNA genetic variants are arranged within the vast multi-chromosome landscape, and how the chromosomal positioning of variants impacts the differential environmental sensitivity observed.

To further our understanding of the nature of rDNA and its environmental dynamics it is necessary to dissect the architecture and composition of individual rDNA clusters. We can begin to do this by elucidating the specific chromosomal locations of rDNA as well as the arrangement of genetic variants within individual clusters. The work in this chapter aims to

- i. Optimise and apply molecular combing to obtain ultra-long combed DNA fibres spanning multi-Mbp in length
- ii. Explore the rDNA landscape and arrangement of rDNA genetic variants in C57BL/6J
- iii. Explore the epigenetic response of genetic variants at the single-molecule level

Largely, this chapter will focus on the work carried out to refine and apply molecular combing to

isolate Mbp length DNA molecules and methods utilised for visualising the arrangement of rDNA genetic variants on the single-molecule level. The limitations and potential pitfalls of these approaches will also be discussed.

3.1.2 Molecular Combing

Molecular combing is one such alternative, a simple yet highly effective DNA fibre stretching technique that allows for the direct visualisation of unmodified DNA at the single-molecule level (Gurevich *et al.*, 2013). Introduced in the early 1990s, molecular combing provided the possibility to study large-scale genomic events like never before. Unlike sequencing methods of the time that were limited by the artificial sheering of DNA into shorter manageable fragments measuring a few hundred bp, molecular combing allowed for the analysis of single DNA molecules over a few hundred kbp (Michalet *et al.*, 1997). Even now, the method remains unchallenged when it comes to molecule length with protocol refinements permitting the visualisation of single molecules upwards of 12 Mbp (Kaykov *et al.*, 2016). This far exceeds the size limit of any 3rd generation sequencing platform including those offered by ONT and PacBio as well as of ultralong optical mapping platforms like the BioNano Genomics Saphyr system (Walt, 2013; Amarasinghe *et al.*, 2020; Logsdon, Vollger and Eichler, 2020) (**Table 3.1**). To put things into perspective the current record of ONT sequencing, the holy grail of ultra-long sequencing, falls short at just over 2 Mbp in published data sets, when compared to the length of molecules attainable with molecular combing.

For visualisation, combed DNA is routinely pre-labelled or in situ hybridised in a technique termed Fibre-FISH, to create high-resolution physical maps spanning large genomic regions (Ersfeld, 2004). It is often employed in the study of large-scale genomic events like DNA replication and rearrangements as well as in the dissection of repetitive loci. The accurate study of these events and loci demands that individual DNA molecules span the entirety of the target loci, meaning they remain particularly difficult to investigate even with ultra-long sequencing methods. Additionally, molecular combing is far from limited to exploring 1-dimensional genomic structure. As combed DNA remains unmodified and in its native state, molecular combing also provides a platform for epigenetic analysis at the single-molecule level. This is demonstrated in a method termed Methyl-combing proposed by A. Nemeth (2014) describing a combination of dynamic molecular combing with immuno- detection of Cytosine C5 Methylation to explore the methylation profile of combed DNA fibres (Németh, 2014). Additionally, 100's of copies of a single genome can be combed onto a single surface, providing far greater potential for genome depth and coverage at a fraction of the cost when compared to ultra-long sequencing methods (Liu, Wang and Dou, 2007). Overall, molecular combing is a cost-effective, efficient, and user-

friendly alternative to ultra-long sequencing and optical mapping methods, providing particular advantages in the study of large-scale genomic events and highly repetitive loci.

Platform	Method	Approach	Average read/ molecule length	Maximum molecule length	Resolution	Molecule state
PacBio SMRT	Hifi Sequencing	Nucleotide sequencing	10- 25 kbp	30 kb	Single base	Replicated DNA
Oxford Nanopore Technology	Ultra-long DNA sequencing	Nucleotide sequencing	30 - 50 kbp	4 Mbp	Single base	Native state DNA
BioNano Genomics	Saphyr	Optical mapping	100 - 250 kbp	3 Mbp	Large scale Structural variation	Native state DNA
Molecular combing	Ultra-long DNA combing	Optical mapping	500 kbp -2 Mbp	12 Mbp	Large scale Structural variation	Native state DNA

Table 3.1 Comparison of 3rd generation sequencing and optical mapping platforms

3.1.3 Evolution of DNA combing methods

Numerous variations of the technique have been developed and applied since its introduction by Bensimon et al. in 1994. Regardless of technical variations, the combing process (described in **Figure 3.1**) can be broken down into three fundamental phases (i) Adsorption of the ends of coiled DNA molecules in solution to a substrate and (ii) stretching of the DNA by the pressures produced by a retreating meniscus, and (iii) relaxing of the deposited DNA on the substrate to its ultimate length. The process results in permanently fixed DNA fibres and has the benefit of aligning DNA fibres in parallel all across the surface, internal size standards are also not required since the stretching factor is constant ($1\ \mu\text{m} = 2\ \text{kbp}$) under standard lab conditions further simplifying large scale analysis ((Bensimon *et al.*, 1994))

Since its introduction, molecular combing has improved dramatically with refinements taking the average length of combed molecules from $\sim 100\ \text{kbp}$ to an excess of 1 Mbp. The original method described by A. Bensimon et al. involved a droplet of DNA solution ($\text{pH} = 5.5$) placed onto a silanised coverslip topped with another glass slide. The droplet was simply allowed to evaporate which caused the air/water interface to migrate, causing DNA to stretch perpendicular to the meniscus onto the silanised glass. Since then efforts to refine the method have focused on improving experimental reproducibility as well as maximising molecule length. Yokota et al. employed a variant of Bensimon's approach in which a glass slide was dragged over the deposited drop of DNA solution (Yokota *et al.*,

1997). The method relied on intentional mechanical movement of the meniscus at a constant speed using a motor-driven apparatus reducing irregularities in combing force associated with spontaneous evaporation.

Allemand et al. demonstrated how DNA combing could be achieved on a variety of non-silanised substrates including those coated with polystyrene and polymethylmethacrylate (Allemand *et al.*, 1997). Importantly, this study highlighted the pH specificity for combing on different surfaces and emphasised the narrow pH window in which DNA combing is possible, typically falling within 0.2 units. Michalet et al. announced dynamic molecular combing based on the Langmuir-Blodgett deposition approach for stretching whole genomic DNA. A coated surface was incubated for 5 minutes in a solution of yeast or human genomic DNA (pH = 5.5), after which it was removed from the solution at a constant speed. The anchored points on DNA moved upward with the coated surface as the coverslip was dragged out of the solution, while the stationary meniscus exerted a constant downward tension at the air-liquid interface. The use of high amounts (2–20 ml) of combing buffer solution in this study provided better results than in the original Bensimon method (5 μ L) with the larger volumes allowing for better control of pH.

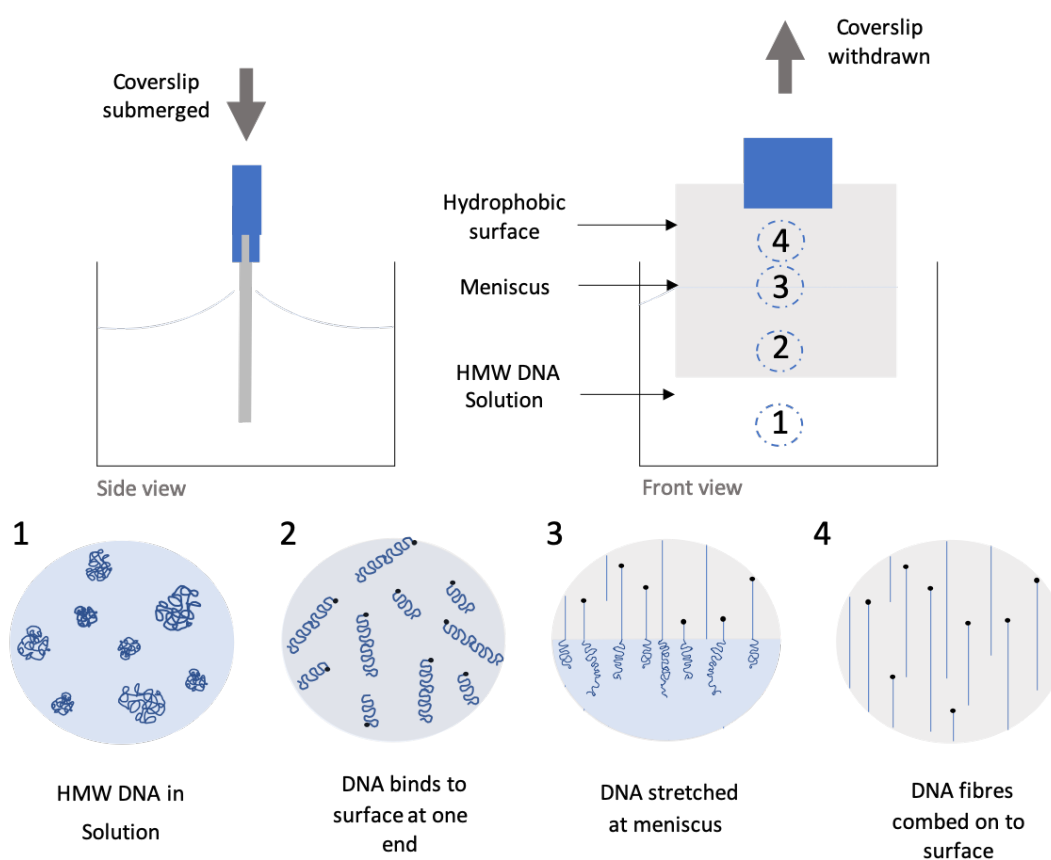


Figure 3.1 Schematic outlining the phases of dynamic DNA combing

3.1.4 Combing Ultra-long DNA molecules

Generally, DNA combing methods limit the average length of combed fibres from 200-600 kbp (Liu, Wang and Dou, 2007; Kahl *et al.*, 2020; Blin *et al.*, 2021). Optimisation of the methodology presents room for improvement and has been a focus for those wishing to permit the comprehensive mapping of larger genomic regions. Genomic vision, a company founded by A. Bensimon offers a standalone DNA extraction kit as well as complementing combing service. Genomic vision aims to streamline the combing process to reproducibly obtain ultra-long combed DNA fibres routinely exceeding a Megabases in length. Similarly, a study published by Kaykov *et al.* has aimed to maximise molecule length by considering factors impacting the physical and chemical stability of DNA, resulting in the isolation of single molecules measuring upwards of 12 Mbp in length (Kaykov *et al.*, 2016). Both methodologies demonstrate advanced methods of DNA isolation in which cells are embedded into agarose and cellular material is slowly digested and washed away over time (**Figure 3.2**). The agarose acts as a protective matrix, limiting mechanical forces that may otherwise shear DNA if it were manipulated in solution or during conventional DNA isolation methods such as phenol-chloroform extraction. The agarose matrix is then digested and the DNA liberated into a combing solution of specified ionic strength and acidity, the chemical composition of which aids in maintaining the stability of the DNA during the combing process and promotes binding to modified adherent surfaces.

The reduction of mechanical stress during the combing process is another key consideration (Kaykov *et al.*, 2016; Chanou and Hamperl, 2021). In contrast to less stable manual methods, movement of the modified substrate into and out of the combing solution is achieved with a finely tuned motorised system. This allows for precise movement in the vertical axis with minimal vibration and governs the force DNA molecules are subjected to whilst being stretched at the air-water interface. Ultimately this reduces artificial sheering of molecules and limits unwanted distortion allowing for the reproducible combing of consistently ultra- long combed fibres

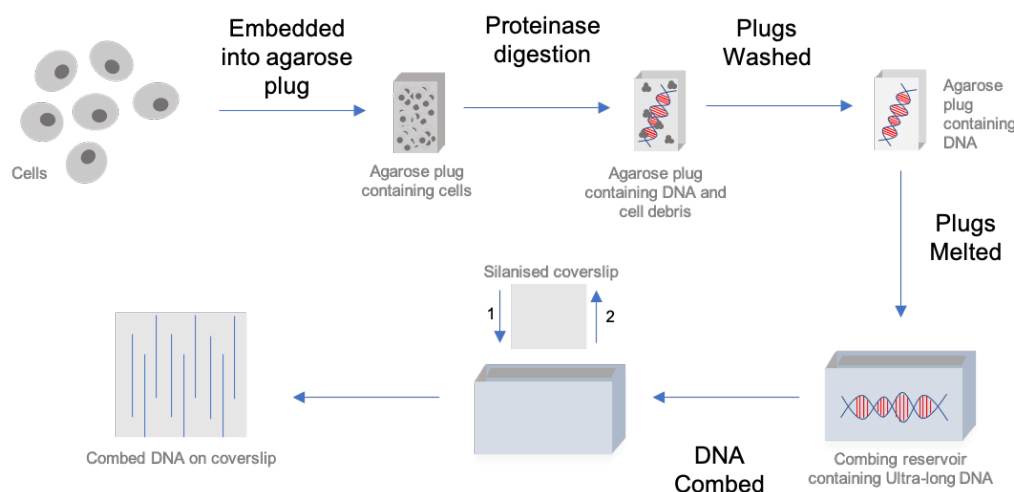


Figure 3.2 Schematic outlining the agarose plug method to obtaining ultra-long combed DNA

3.1.5 Factors influencing DNA combing

Multiple factors can impact the quality and length of combed DNA fibres including but not limited to DNA isolation methods, surface modification, ionic strength and acidity of the combing solution as well as the method, speed, and force with which fibres are combed. Changes in pH have been demonstrated to affect the density of DNA molecules adsorbed on the surface as well as the extent of stretching (Allemand *et al.*, 1997). The best pH for combing DNA on a surface is dependent on the surface composition. Most, if not all, of the data, concur that pH levels in the range of 5–6.5 are suitable for adsorption of DNA, specifically to hydrophobic surfaces (Nazari *et al.*, 2013). DNA bases are significantly protonated at very low pH values (e.g., pH = 3 and 0.1 M salt 2.2.3 equivalent to 50% protonation). This protonation lowers the melting temperature and weakens the hydrogen bonds that hold the strands together, partially exposing the hydrophobic core of the DNA helix, resulting in nonspecific DNA adsorption to the surface (Mallajosyula and Pati, 2007). As the pH is increased the melting occurs less frequently and at pH 5.5, only the extremities of DNA are sufficiently hydrophobic to attach to a hydrophobic surface (Allemand *et al.*, 1997). The ionic strength of the combing buffer is also reported to impact the efficacy of DNA combing with a 100 mM NaCl concentration resulting in better stretching, increased coverslip coverage, and reduced DNA fragmentation (Kaykov *et al.*, 2016). Generally, Na⁺ ions in the combing solution screen the negatively charged phosphate groups along the DNA backbone and act to reduce electrostatic repulsion and melting of double-stranded DNA. The

nature of substrate surface functionalization is also of importance and in general, the degree of stretching tends to be higher on hydrophobic rather than hydrophilic surfaces (Nazari and Gurevich, 2013). This is due to the strong and specific adsorption of DNA on hydrophobic surfaces, allowing for a higher meniscus force and increased combing efficacy. Silanisation is a common choice for preparing hydrophobic combing surfaces however the nature of the silane used determines the degree of substrate hydrophobicity, directly affecting the strength of DNA-substrate interactions. Other surface modifications such as those achieved through the use of several polymers containing π -conjugation units such as PVCs also result in highly-aligned DNA molecules (Labit *et al.*, 2008). As the speed at which the meniscus retracts is proportional to the force applied to DNA fibres at the water-air interface, this also dramatically impacts the quality of combing. Generally, a speed of 300–900 $\mu\text{m/s}$ is considered acceptable with speeds significantly lower or higher than this range resulting in poorly stretched fibres and an overall lower density of combed DNA (Kaykov *et al.*, 2016). Sub-optimal conditions in the parameters outlined above and during the preparation of DNA simply result in sub-optimal DNA combing. To ensure a uniform spread of linearised single DNA molecules it is important to consider all aspects of the combing protocol. Outlined in **Figure 3.3** are the pitfalls commonly observed in combing experiments and the possible causes.

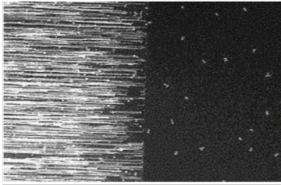

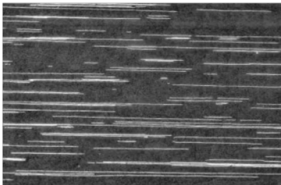
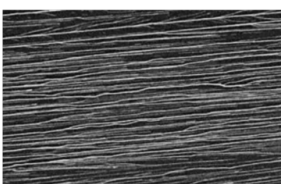
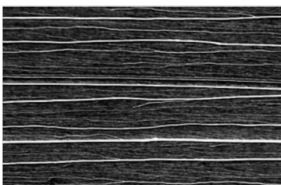
Issue	Example image	Possible causes
Non-uniform or irregular adherence to surface		<ul style="list-style-type: none"> • Defect on silanised coverslip • Substrate modification- pH incompatibility
Short fragmented DNA fibres		<ul style="list-style-type: none"> • Poor handling of DNA • Degraded DNA sample • Insufficient plug washing
Poor adherence/ low combing density		<ul style="list-style-type: none"> • Low cell count • Buffer pH degradation
High combing density, overlapping and wavy fibres		<ul style="list-style-type: none"> • High cell count • Insufficient plug washing • Mechanical stress during combing
DNA is poorly individualized. Fibres may bundle or form networks		<ul style="list-style-type: none"> • Non-efficient proteinase digestion • Poor homogenisation of cells

Figure 3.3 Common molecular combing issues and possible causes.

Adapted from Genomic visions FibrePrep® DNA extraction kit troubleshooting guide.

3.1.6 Probing genomic loci

When used in conjunction with fluorescence in situ hybridization, DNA combing allows for the direct visualisation and mapping of large-scale genomic rearrangements, DNA replication, and repetitive elements (Heiskanen *et al.*, 1995). It is particularly suited to the study of genomic structure, copy number variation, as well as the size quantification between sequence contigs in a genome assembly. An example of fibre-FISH is presented in **Figure 3.4**, where combed DNA has been probed to visualise the orientation of rDNA units within an array segment. The fluorescent probes used in such studies tend to target long stretches of DNA, typically kilobases in length. However, the binding of such large probes easily tolerates the mismatching of a few bases within the target sequences, making them

unable to distinguish between the subtle sequence variations caused by SNPs. For this reason, it is necessary to explore alternative approaches to probing combed DNA and examine the structure of rDNA arrays in a variant-specific manner.

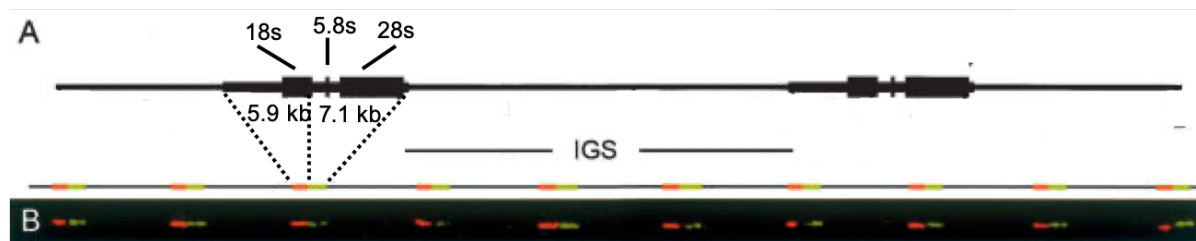


Figure 3.4 Structural analysis of human rDNA gene array with DNA combing (A) Schematic representation of two canonical rDNA units **(B)** Two colour hybridisation on combed DNA, the red 5.9 kb fragment detected with Texas Red and the green 7.1 kb fragment detected with FITC. Image displays 10 canonical rDNA units arranged in tandem, each unit composed of dual fluorescent signal and non-hybridised spacer region.

Adapted from Caburet *et al.* 2005 Figure 1.

In recent years, CRISPR-Cas9 has been leveraged as a particularly effective tool for DNA labelling. Through the use of a catalytically inactive Cas9 (dCas9) coupled with a target-specific guide RNA (gRNA), genomic loci may be labelled rather than cleaved. The direct or indirect fusion of fluorophores to dCas9 has permitted the visualisation of genomic loci dynamics in live cells (Chen *et al.*, 2013) as well as the spatial relationships of genetic elements in fixed cells (Deng *et al.*, 2015). The targeted region is usually 17-20 bp, and due to this demands high sequence specificity, though the gRNA still tolerates several base mismatches and truncations (Hiranniramol, Chen and Wang, 2020). However, the binding specificity of Cas9 is further refined by recognition of the protospacer adjacent motif (PAM) a 2-6 bp sequence directly adjacent to the target sequence which determines the binary binding of the dCas9-gRNA complex (Gleditsch *et al.*, 2019). Spatiotemporal allele organization by allele-specific CRISPR live-cell imaging (SNP-CLING) is a method described by Maass *et al.* in which fluorescently tagged dCas9 probes are employed to visualise alleles in live cells by exploiting the PAM specificity of Cas9 nucleases (Maass *et al.*, 2018). Displayed in **Figure 3.5** is an outline of the principles behind SNP-CLING and its allele labelling and resolving capabilities. The method exploits the PAM recognition specificity of dCas9 by positioning an allele-specific SNP within the PAM site which in turn promotes or hinders binding even if the adjacent targeted sequences are identical. Specifically, a SNP which creates a SpdCas9 specific PAM (3'-NRG-5') acts to promote binding of the probe whilst a SNP that creates a SpdCas9 non-specific PAM (3'-NYH-5') prevents binding. The effectiveness of this approach is demonstrated with the differential labelling of alleles for gene Ypel4 in 129S1 x CAST hybrid mice. Within a cellular environment, SNP-CLING is capable of resolving loci in close proximity,

demonstrated by the labelling of X chromosomes genes TSIX and XIST, between which there is a linear genomic distance of ~69 kb. Two distinct signals from the targeted loci are observed with a spatial displacement of ~163 nm, and with current microscopic constraints, this appears to be the resolving limit. Considering the results of this study, SNP-cling probes hold promise for discerning SNP-specific rDNA alleles. However, due to the tandem arrangement of rDNA repeats as well as the small genomic distance between each unit (~45 kb in mouse), the inherent resolution limit of SNP-CLING means rDNA arrays must be studied beyond the confines of cellular space. For this reason, this study aims to combine ultra-long DNA combing to obtain entire linearised rDNA clusters, and SNP-CLING probes to visualise rDNA SNP alleles at the single-molecule level.

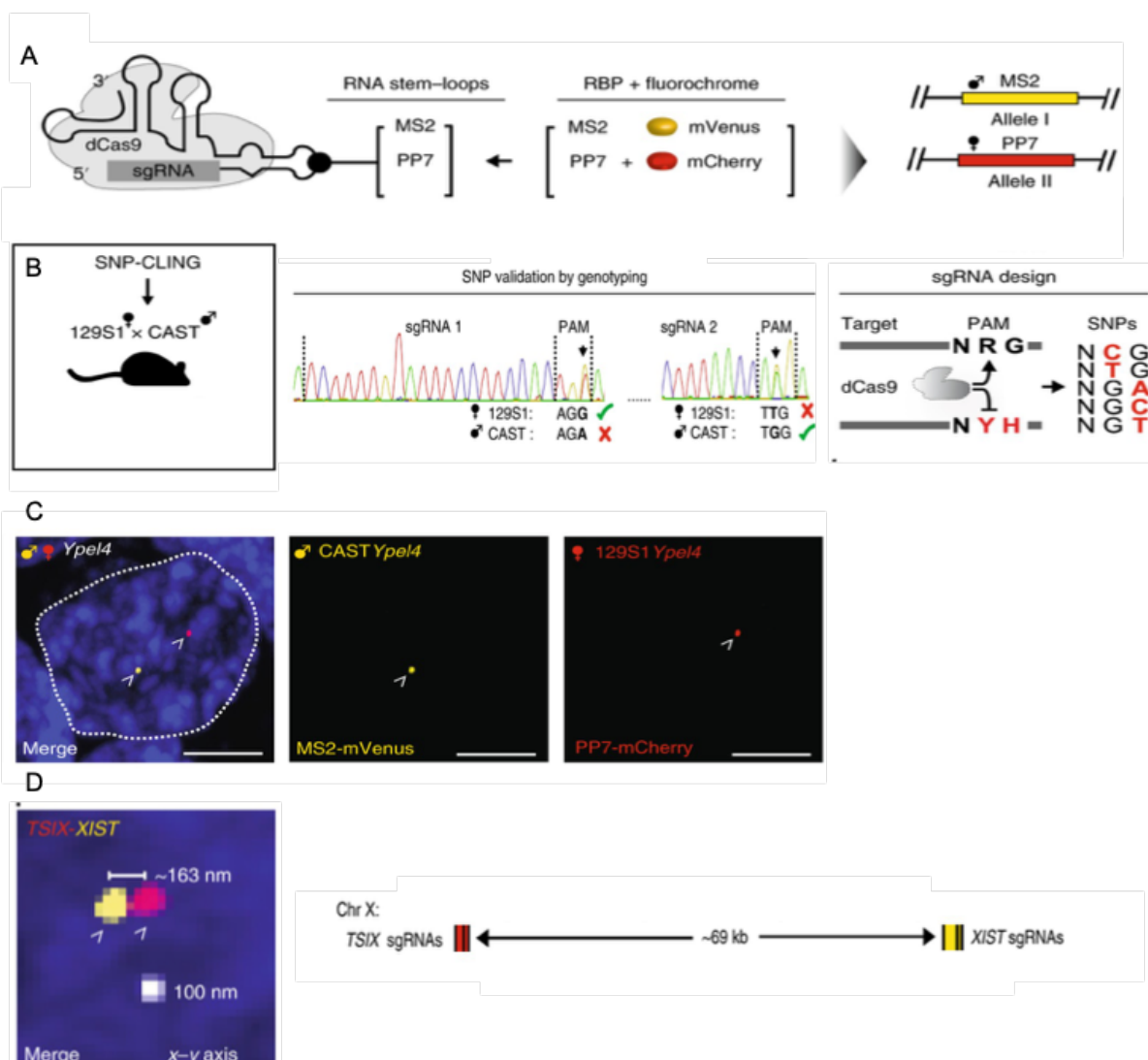


Figure 3.5 Allele specific labelling with SNP-CLING (A) dCas9 is bound to gRNA harbouring internal protein binding motifs (MS2 or PP7) that direct the complex to the target locus. Corresponding RNA binding proteins MS2 or PP7 are fused to fluorescent proteins mVenus or mCherry that differentially label 2 different alleles. (B) 129S1 x CAST hybrid mouse used for harvesting MEFs expressing 2 alleles for the Ypel4 gene (left) and sanger sequencing of selected SNP's within Ypel4 gene confirming heterozygosity and PAM presence (middle). For allele specific targeting the heterozygous SNP is positioned within the PAM. A SNP can create the SpdCas9 specific PAM 3'-NRG-5' to promote binding. Likewise, a SNP can create a SpdCas9 non-specific 3'-NYH-5' PAM if the second or third nucleotide in the spdCas9 PAM 3'-NRG-5' is replaced (right). (C) Allele specific visualisation of 129S1-Ypel4 (yellow) and CAST-Ypel4 (red) in 129S1 x CAST MEFs. (D) Resolution limits of SNP-CLING demonstrated by targeting genes TSIX (red) and XIST (yellow). XIST and TSIX are targeted by gRNA to loci 69 kb apart on chromosome X and are resolved successfully as two distinct signals with a spatial displacement of 163 nm. White spot presented as scale=100 nm.

3.2 Materials and Methods

Materials and methods section		Page number
2.1	Cell culture techniques	36
2.1.1	Mouse Embryonic fibroblasts (MEFs)	36
2.1.4	Human Embryonic Kidney Cells (HEK-293)	37
2.2	DNA and RNA techniques	37
2.2.1	Agilent bioanalyser	37
2.2.2	Gel electrophoresis	37
2.2.3	Assessing nucleic acid purity and concentration	38
2.3	Generation of control cell lines for SNP-specific probe testing	39
2.4	SNP-specific localisation of C57BL/6J rDNA promoter variants	41
2.5	Molecular combing	43
2.6	DNA labelling techniques	47

3.3 Results

The first step in achieving the aim of visualising the arrangement of rDNA variants at the single-molecule level was to establish and optimise an ultra-long DNA combing protocol. To this end, parameters impacting DNA combing including but not limited to surface functionalisation, DNA extraction methods, pH conditions, and DNA concentration were evaluated and refined and the results of this process are outlined below. For each step in the optimisation process, each unique condition was tested in triplicate, i.e. 3 separate DNA preparations were produced for each condition. To ensure findings were consistent and representative across samples from each condition, a single DNA preparation was used to generate 3 combed fibre preparations with each sample slide imaged and compared.

3.3.1 Manufactured silanised slides outperform silanised slides produced in-house

The success of any DNA combing experiment is greatly impacted by the surface on which DNA is combed. Silanisation of glass is a modification commonly employed in combing studies, however, the exact chemical composition of the silane compound used as well as its means of application can alter combing efficacy dramatically. To establish the most suitable surface modification several silane-based chemicals were tested, and various processes with which to achieve this were evaluated. According to published protocols, glass slides were treated with either Trimethoxy-octenylsilane (Labit *et al.*, 2008) or 7-Octenyltrichlorosilane (Kaykov *et al.*, 2016), using liquid or gas phase silanisation respectively. The successful modification of glass surfaces after treatment was crudely verified using the water droplet method, in which a droplet of water placed onto the treated hydrophobic surface assumes a characteristic dome-shaped configuration, in contrast to a flatter configuration on untreated glass (**Figure 3.6A**).

Upon visual confirmation of surface modification, the DNA combing suitability of in-house produced slides was assessed alongside aminoalkyl silane prep slides purchased from Sigma-Aldrich. To achieve this, high molecular weight DNA (HMW DNA) was extracted from a culture of adherent 7bl7bl/6 MEF cells and approximately 100 ng was dissolved in 2 ml of combing buffer (50 mM MES, 100 mM NaCl, pH 6). The DNA solution was combed onto each set of coverslips and factors including DNA attachment, surface coverage, and fibre linearization were evaluated.

Coverslips modified with Trimethoxy-octenylsilane via liquid phase modification resulted in poor DNA combing. Though DNA adherence was observed, fibres were not stretched and remained coiled and tightly bound upon the surface (**Figure 3.6B**). Adhered DNA was identified as spots of high

fluorescence intensity visible across the surface, the presence of these without any linearised fibres may suggest the strength of molecule adherence prevents adequate uncoiling and combing or that the combing solution of pH 6 that is incompatible with Trimethoxy-octenylsilane modification. Slides treated with 7-Octenyltrichlorosilane via gas-phase modification allowed for both adherence and of DNA **Figure 3.6C**. Molecules were linearised well and arranged in parallel across the surface however significant physical surface irregularities were also observed. Coverslips treated using this method were non-homogeneously modified with imperfections where surface silanisation had failed to be prevalent across the surface. Such surface irregularity can often negatively impact both the adherence and stretching of DNA onto the surface limiting combing capabilities and affecting downstream analyses. The use of Sigma-Aldrich aminoalkyl silane prep slides allowed for both the adherence and stretching of DNA molecules. Individual molecules are distinguishable as thin streaks of fluorescence intensity abundantly visible across the surface (**Figure 3.6D**). The surface is additionally free of imperfections such as those noted with 7-Octenyltrichlorosilane treatment. However, many DNA molecules appear to be attached to the surface at both ends resulting in the 'U' shaped fibres. DNA attachment of this nature is sub-optimal as it hinders the accurate analysis of fibre lengths and may impact molecule probing downstream. Even so, due to the acceptable extent of DNA binding and consistent surface modification, as well as 'U' shaped binding likely being due to the pH environment Sigma silane prep slides were used for all subsequent DNA combing experiments.

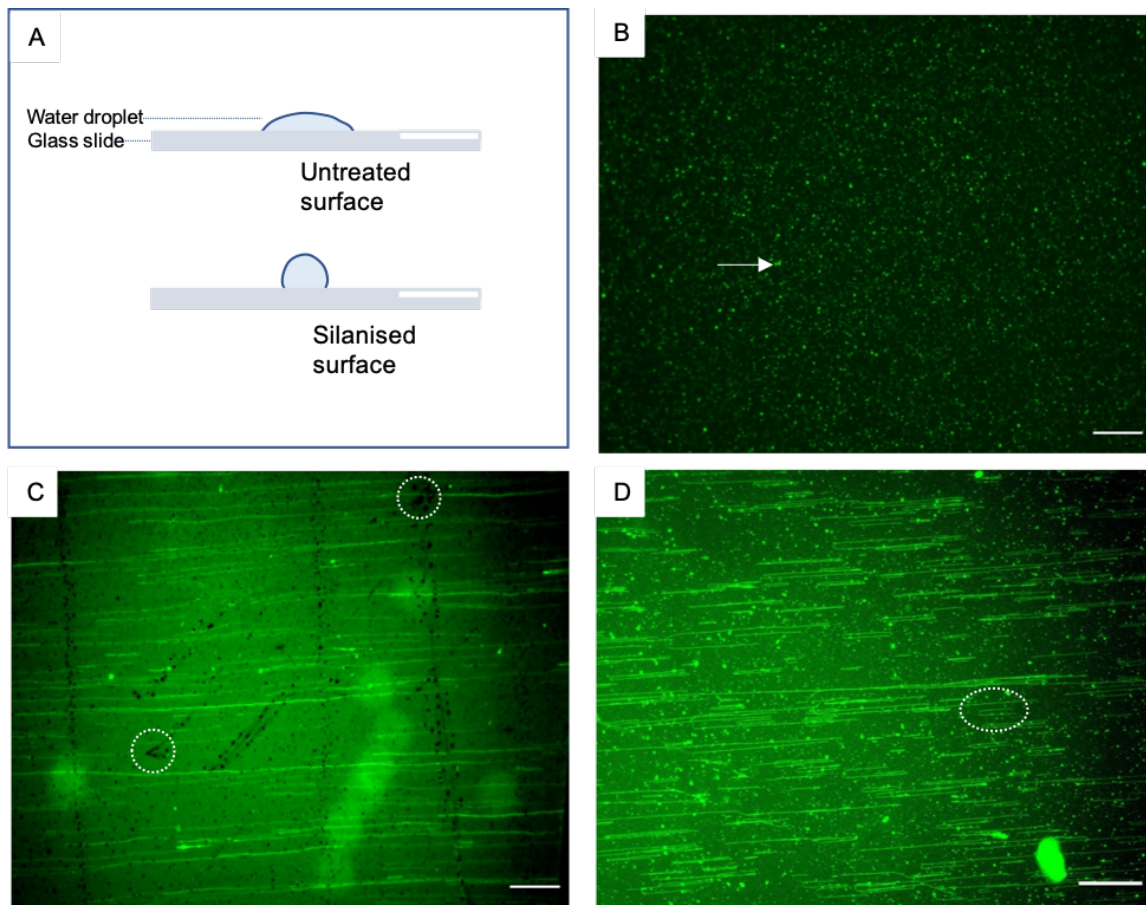


Figure 3.6 DNA combing efficiency on inhouse and manufactured silanised slides. (A) Schematic demonstrating glass surface modification confirmed via water droplet configuration. DNA combing compatibility of slides was tested with HMW MEF DNA combed on three silane treated surface (B) Liquid phase Trimethoxy-octenylsilane treatment allows DNA binding (indicated) but prevents molecule stretching (C) Gas phase 7-Octenyltrichlorosilane treated slides allow adherence and combing of linearised DNA molecules, however silanisation is non-homogenous and imperfections are prevalent across the surface (indicated) (D) Sigma-Aldrich aminoalkyl silane prep slides allow binding and stretching of DNA. Molecules are bound at both end and arranged in a 'U' shaped configuration (indicated).

B-D are representative of each condition outlined, conducted as experimental triplicates.

Per sample, 100 ng of HMW MEF DNA dissolved in 2 ml of 50 mM MES, 100 mM NaCl, pH 6. Combed DNA stained with 0.1 μ M YOYO-1 iodide, visualised by Epi-Fluorescence microscopy, scale bar= 20 μ m/40 kb, 40 X Magnification

3.3.2 Effect of pH on DNA combing

Having found a suitable surface for DNA combing it was now important to establish a compatible buffer pH. The impact of varying pH strengths on combing efficacy was investigated, evaluating factors including DNA adherence, fibre linearisation, and molecule length. To explore this 100 ng HMW MEF DNA was resuspended in combing buffer with pH values ranging from pH 5 – 6.5, increasing at 0.5-unit intervals. DNA combing was executed using Sigma aminoalkyl silane prep slides. At pH 5 DNA molecules adhered to and were stretched onto the surface (**Figure 3.7A**). However, fibres were highly fragmented and visually smaller than those combed at pH 6, with an average length between 100-200 kb. At pH 6 DNA molecules effectively adhered to and were combed across the surface, with fibres consistently linearised whilst retaining long lengths (**Figure 3.7B**). The high molecule adherence and surface coverage resulted in many fibres aligned in close proximity, preventing the accurate analysis of fibre length, however, the average fibre length was estimated to be >400 kbp with some fibres spanning the entire field of view measuring > 700 kb. As previously described, DNA fibres combed from a combing buffer (composition previously described) of pH 6 onto Sigma Aldrich aminoalkyl silane prep slides result in sub-optimal combing (**Figure 3.7C**). DNA molecules were bound at both ends preventing fibre linearisation, rather fibres adopted a 'U' shaped confirmation (**Figure 3.6D**), even so, a noticeable increase in general fibre length was observed when compared to DNA combed at pH 5, with the average length >400 kb. The testing buffer of pH 6.5 showed minimal adherence to the surface, with the few attached molecules being poorly linearised (**Figure 3.7D**). Considering these results, a combing buffer of pH 5.5 was selected as ideal for subsequent combing attempts. It is however important to note that though the combing outcomes described above are representative of the conditions outlined, DNA combing remained highly variable with combing results being inconsistent between experimental repeats even at the established optimal pH. Contrastingly, the use of MES buffer of pre-established pH 5.5 purchased from Thermofisher Scientific allowed for substantially greater reproducibility and minimised future irregularities in the combing procedure. **Figure 3.8** present a representative image of combed DNA fibres using Thermofisher scientific pH 5.5 MES buffer. Fibres are bound uniformly across the surface, are well linearised, and arranged in parallel, whilst retaining their expected high molecular length. This pre-prepared buffer was used in the preparation of the combing buffer for all subsequent experiments.

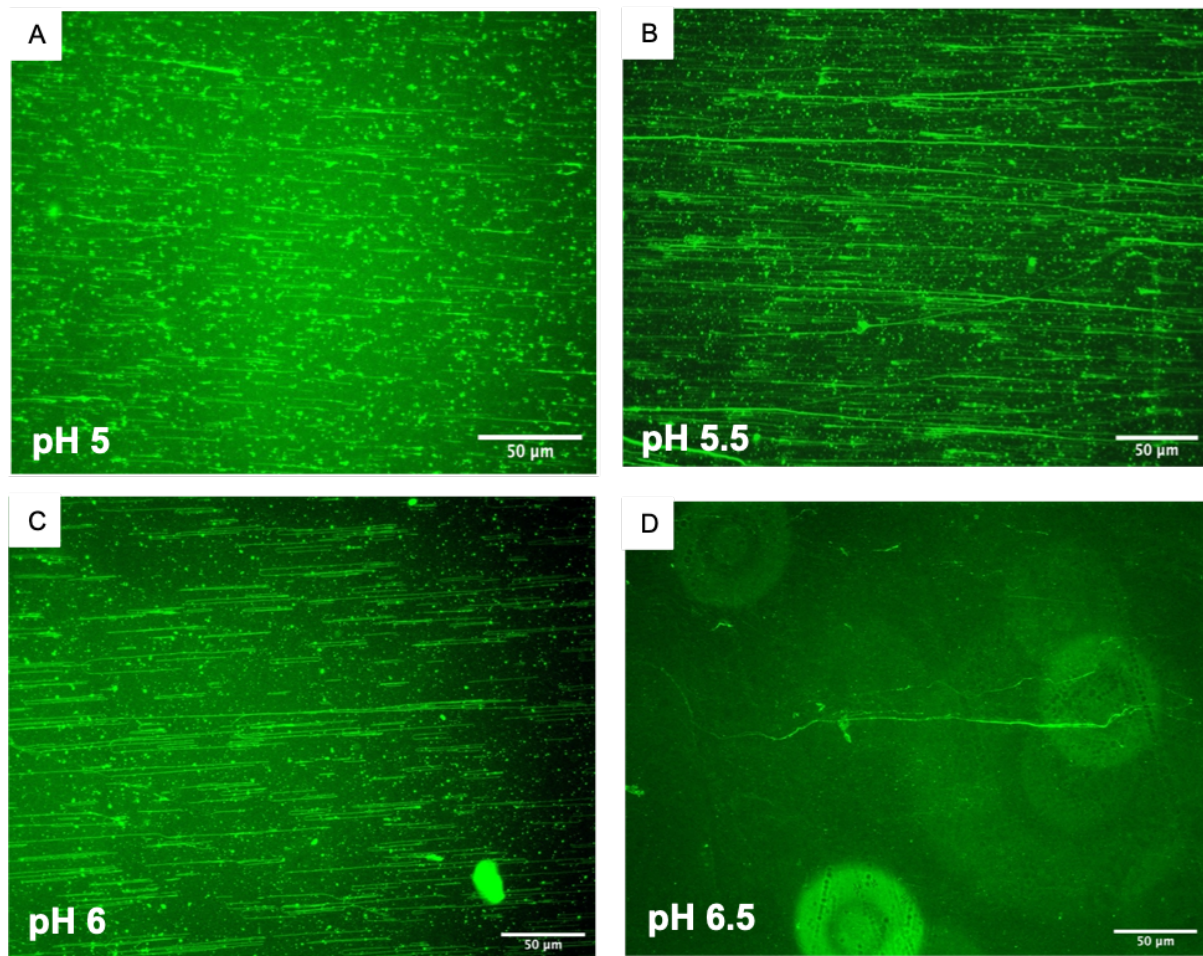


Figure 3.7 Effect of pH on DNA combing. DNA was combed from a buffer composed of (50 mM MES, 100 mM NaCl) at (A) pH 5 (B) pH 5.5 (C) pH 6 (D) pH 6.5

Per sample, 100 ng of HMW MEF DNA dissolved in 2 ml of buffer (50 mM MES, 100 mM NaCl). DNA is stained with 0.1 μM YOYO-1 iodide, visualised by Epi-Fluorescence microscopy, scale bar= 50 μm/100 kb, 40 X Magnification

A-D are representative of each condition outlined, conducted as experimental triplicates.

Figure 3.7C is the same as Figure 3.6D, presented here in the context of buffer pH optimisation.

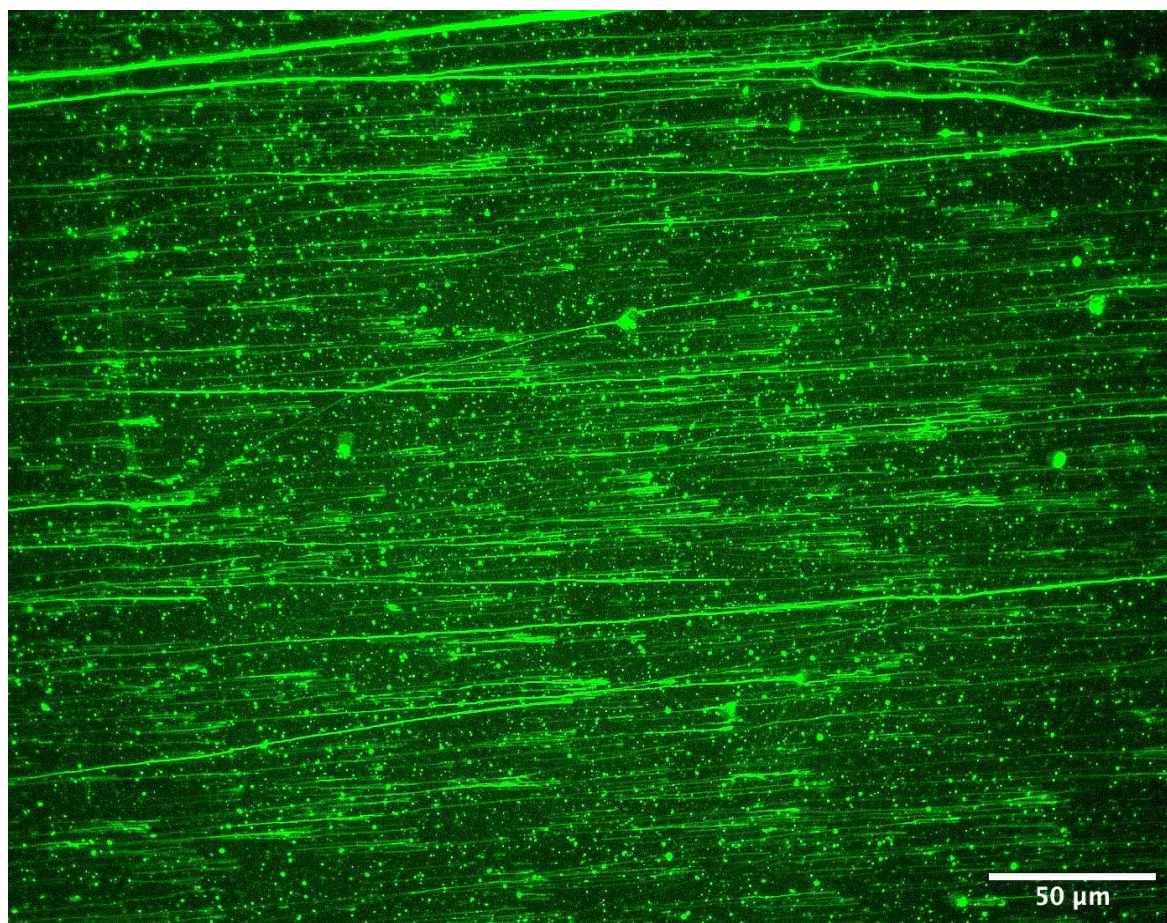


Figure 3.8 DNA combing using pre-prepared MES pH 5.5 buffer. DNA was combed from pH 5.5 buffer purchased from Thermofisher Scientific.

100 ng of HMW MEF DNA dissolved in 2 ml of 50 mM MES + 100 mM NaCl, pH 5.5. Combed DNA stained with 0.1μM YOYO-1 iodide, visualised by Epi-Fluorescence microscopy, scale bar= 50 μm/100 kb, 40 X Magnification

3.3.3 Kaykov et al. and Genomic Vision DNA extraction protocols result in the inadequate combing of MEF DNA

Having established a suitable combing surface and a compatible pH to obtain well-adhered, linearised DNA fibres it was important to next attempt the combing of ultra-long DNA molecules. The DNA combing protocol outlined by Kaykov et al. (2016), was fundamental in this study as a starting point for obtaining ultra-long combed DNA, defined here as molecules measuring in excess of a Megabase (>500 μm in length). The protocol outlines the method and reagents necessary to obtain ultra-long DNA suitable for DNA combing without the need for specialised manufactured kits. The protocol outlining DNA extraction methodology by Kaykov et al., (2016), was tested alongside Genomic Vision's 'Fibre prep DNA extraction kit' and the associated protocol for the combing of MEF DNA. Both protocols utilise LMP agarose as a matrix to stabilise cells, from which DNA is slowly extracted and refined. This approach limits the mechanical sheering of DNA and aims to maximise molecule size. Both extraction protocols were executed as directed however proved unsuccessful for the effective combing of MEF DNA with neither protocol allowing for the effective combing of individualised, and well-separated DNA. In both cases, the majority of the DNA appears aggregated and clumped, indicated by areas of intense fluorescence without the presence of individual molecules. Employing the DNA extraction protocol outlined by Kaykov et al. resulted in a mass of aggregated DNA within which a network of fibres can be seen (**Figure 3.9A**). Similarly, though to a lesser degree, the Genomic Vision DNA extraction kit and protocol results in masses of DNA streaked across the slide surface, with some fibres emerging from and within the streak (**Figure 3.9B**). These results would suggest that the DNA is poorly suspended within the combing solution and likely remains trapped within a partially digested agarose matrix. Though there was comparative unsuccess observed with both protocols, in the interest of cost it was decided to proceed with optimising the Kaykov et al. (2016) protocol which is from this point adapted for all subsequent combing experiments.

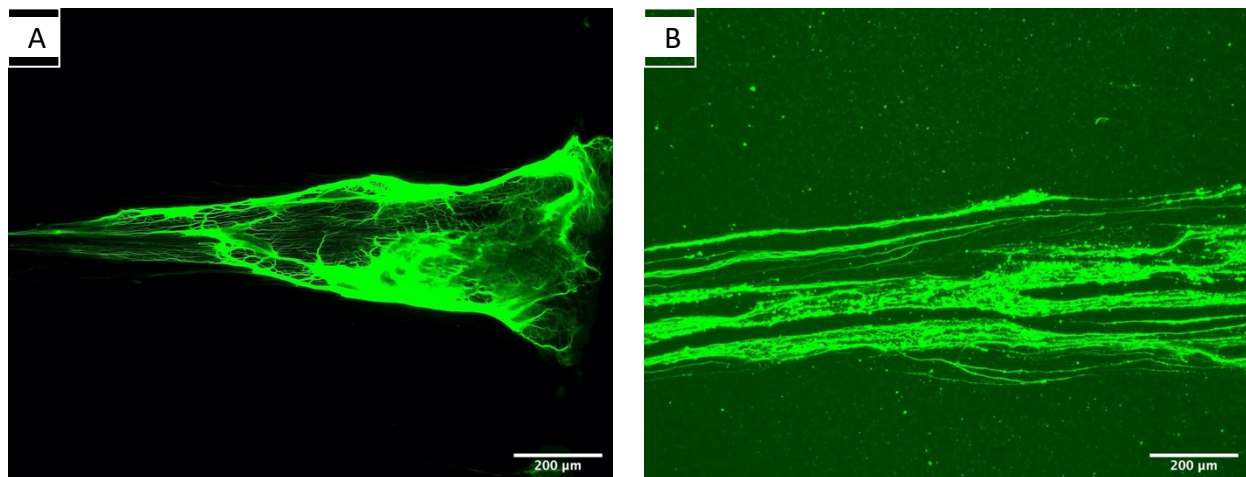


Figure 3.9 Comparing Kaykov et al. and Genomic Vision's DNA extraction protocols. (A) DNA extracted and combed according to conditions and reagents recommended in Kaykov et al. (2016). (B) DNA extracted using Genomic Vision's DNA extraction kit and associated protocol, combed according to kaykov et al. (2016).

Per sample, DNA extracted from 1×10^5 MEF cells embedded in LMP agarose plugs, dissolved in 2 ml of 50 mM MES, pH 5.5. Combed DNA stained with 0.1 μ M YOYO-1 iodide, visualised by Epi-Fluorescence microscopy, scale bar= 200 μ m, 10 X Magnification

A-B are representative of each condition outlined, conducted as experimental triplicates.

3.3.4 Adapting Kaykov et al. (2016) for combing MEF DNA

As the unaltered DNA extraction protocol outlined by Kaykov et al. (2016) led to aggregated DNA, without individualised molecules (**Figure 3.10A**), the protocol was adapted and refined to allow for effectively combing MEF DNA. In the first iteration of the Kaykov et al. DNA extraction protocol, DNA appears condensed and trapped within a partially digested agarose matrix. Strands which are perhaps individual DNA fibres can be seen within the core of the aggregate and emerging from it, though remain largely trapped within. To remedy this, two changes were made during the agarose digestion step. The first was an increase in agarose melting time from 15 to 30 minutes at 70 °C compounded with an increase in β -agarase digestion from overnight (typically 16 hours) to 24 hours. As a result, the agarose matrix appears largely digested with the DNA dispersed across the coverslip leading to an overall improvement in the spread of DNA (**Figure 3.10B**). However, DNA remained inadequately suspended in solution, with molecules adhering to the coverslip as a network of interconnected fibres rather than individualised molecules. To promote effective suspension of DNA in the combing solution and ultimately a uniform spread of individualised DNA molecules, it was paramount to ensure that cells were singularised before embedding in agarose plugs. It was speculated that any cell clumps at this step could likely lead to aggregation of DNA during extraction and negatively impact the combing efficacy. To ensure this, the cell solution was filtered through a 20 μ m cell strainer before embedding

in agarose, with the single-cell suspension confirmed via microscopy. DNA was then extracted using the aforementioned adapted protocol and combed. Ensuring a single cell suspension indeed served to improve combing outcomes, with DNA no longer aggregated but instead linearised and combed across the coverslip (**Figure 3.10C**). Though some singularised fibres can be observed, there are many areas of relatively intense fluorescence indicative of 'bundled' DNA that has not effectively separated. It was postulated that the observed bundling may be due to inadequate proteinase digestion. To improve proteinase digestion gentle but constant agitation during digestion was introduced. Agarose plugs were submerged in proteinase digestion solution and rotated at 10 RPM at 50 °C. This addition of constant rotation improved fibre resolution, with bundled DNA effectively separated into individual DNA fibres (**Figure 3.10D**).

Though the majority of DNA molecules are individualised, fibres were inadequately stretched, appearing kinked with suboptimal linearization. Such poor stretching and excessive fibre overlapping are likely to impact the downstream probing of DNA as well as any fibre length analysis. These issues were considered to be related to a high concentration of DNA in the combing solution leading to an oversaturation of the coverslip with molecules and subsequently, improper stretching. To improve the linearisation of the combed DNA and to produce a spread of uniform individualised fibres a range of DNA concentrations were tested. DNA concentration of the combing solution is dependent on the initial cell densities during agarose plug formation. The initial cell density of 1×10^5 cells per agarose plug results in poorly linearised fibres that appear with regular kinks and heavily overlap (**Figure 3.11A**). A cell density of 0.5×10^5 improves the linearization of fibres, with DNA stretched straighter, though significant overlap between fibres remains, preventing the confident identification of fibre start and end sites, which is necessary for identifying individual molecules. (**Figure 3.11B**). A cell density of 0.25×10^5 further improves fibre resolution, with noticeably less overlap between molecules observed, with fibre ends distinguishable (**Figure 3.11C**). After evaluating the results from these conditions, it was decided that a cell density of 0.25×10^5 was appropriate to achieve a uniform spread of singularised DNA fibres, and was thus used for all subsequent combing experiments.

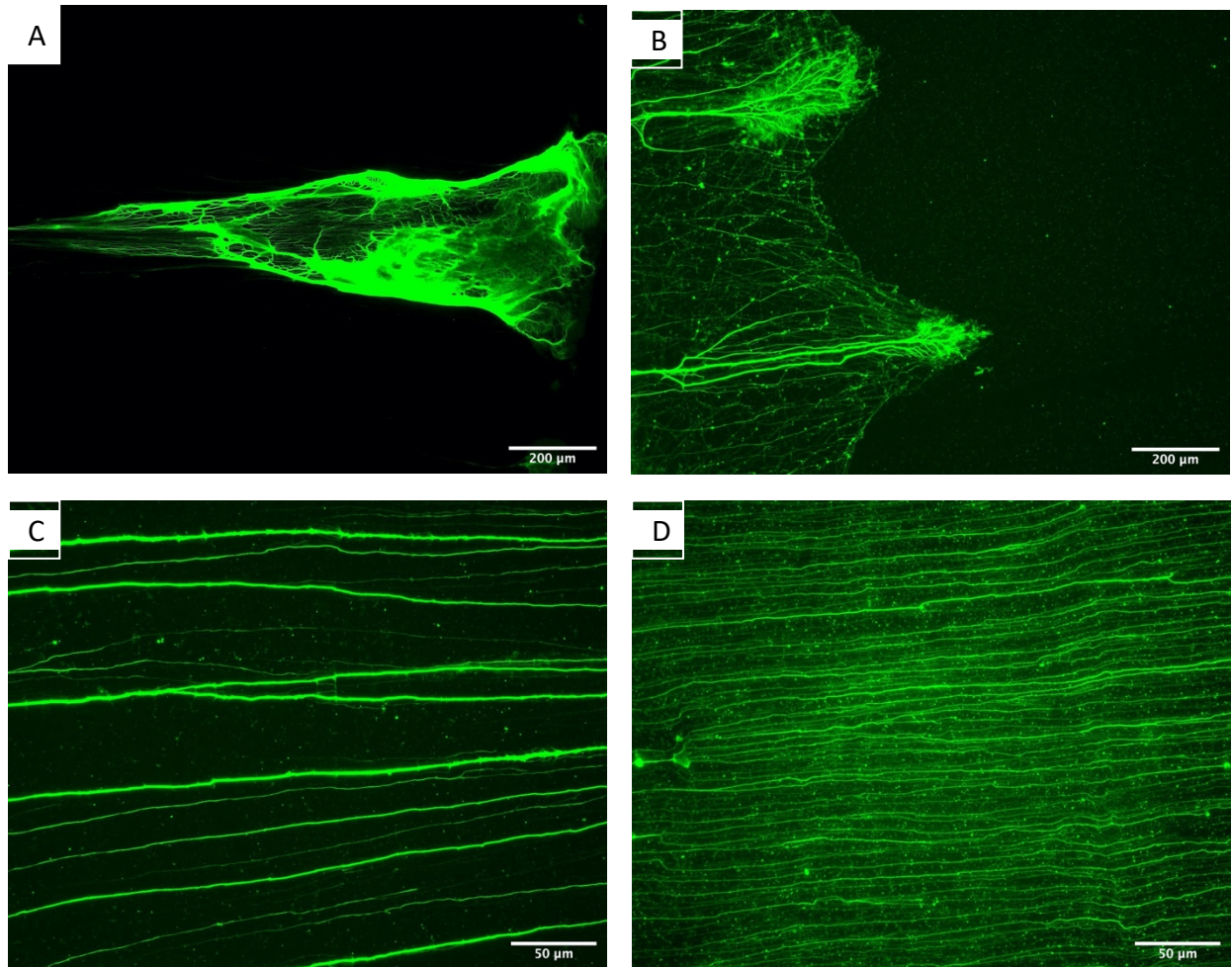


Figure 3.10 Optimising Kaykov et al. protocol parameters. Kaykov et al. protocol was altered with each iteration of the combing experiment with the aim of achieving a uniform spread of individualised combed DNA fibres. (A) Following the kaykov et al. protocol, DNA remains condensed and trapped within the incompletely digested agarose matrix (B) An increase in agarose digestion time from 16 to 24 hours results in a significant increasing in matrix digestion and DNA liberation, though DNA fibres still remain largely aggregated (C) Prior to agarose plug embedding, cell suspension is filtered through 20μm cell filter to ensure single cell suspension. DNA is now largely dispersed though a large proportion of fibres remain bundled together. Individual (i) and bundled (ii) DNA fibres are distinguished by relative fluorescence intensities and thickness. (D) Employing gentle agitation throughout proteinase K digestion further improves fibre resolution and DNA is largely individualised, though fibres are kinked, and overlap

Per sample, DNA extracted from 1×10^5 MEF cells embedded in LMP agarose plugs was dissolved in 2 ml of 50 mM MES, 100 mM NaCl, pH 5.5. Combed DNA stained with 0.1μm YOYO-1 iodide, visualised by Epi-Fluorescence microscopy, scale bar= (a, b) 200 μm/400 kb, (c, d) 50 μm/100 kb.

A-D are representative of each condition outlined, conducted as experimental triplicates.

Figure 3.10A is the same as Figure 3.9A, presented here in the context of protocol development.

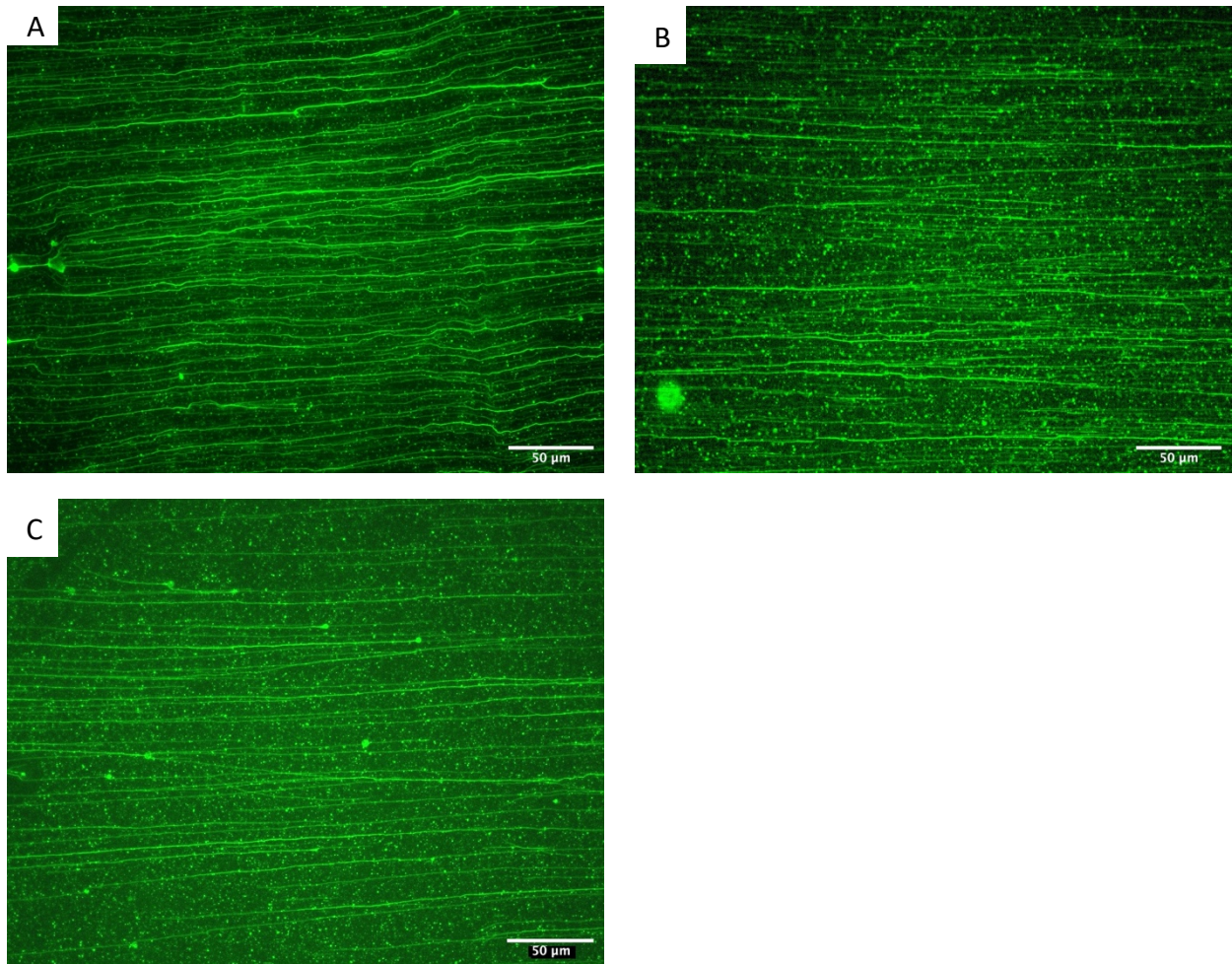


Figure 3.11 Optimising cell number and DNA density. Varying cell densities were tested to achieve optimal individualisation and uniform spread of DNA fibres. Per agarose plug (A) 1×10^5 (B) 0.5×10^5 (C) 0.25×10^5 cells, were used.

Per sample, DNA extracted from MEF cells embedded in LMP agarose plugs was dissolved in 2 ml of 50 mM MES, 100mM NaCl, pH 5.5. Combed DNA stained with 0.1µM YOYO-1 iodide, visualised by Epi-Fluorescence microscopy, scale bar= 50 µm/100 kb, 40 X Magnification.

A-C are representative of each condition outlined, conducted as experimental triplicates.

3.3.5 Adapted combing protocol allows for isolation of DNA molecules measuring >5 Megabases in length

Introducing changes outlined in (Figure 3.12A) to the original Kaykov protocol allows for the preparation of a uniform spread of ultra-long linearised DNA fibres suitable for probing and molecule length analysis. Additionally, the protocol permits the isolation of ultra-long DNA fibres regularly measuring >500 kb. A representative image of combed DNA in Figure 3.12B shows 6 DNA fibres spanning the entirety of the presented field of view and beyond. At x 40 magnification, a single field of view measures ~350 µm which equates to ~700 kbp (Figure 3.12C). Quantitative assessment of combed DNA fibre lengths that molecule length varied significantly ranging from a few 100 kbp to over 6 Mbp. To quantify the length distribution of ultra-long combed molecules after protocol

optimisation, entire sample slides were imaged in a tile-like manner using an automated scanning microscope. Tiled images were 'stitched' *in silico* to create a single continuous image of each slide. Fibres were subjected to a rudimentary random sampling process in which a total of 5 fibres were selected from each tile, 1 fibre from each of the 4 corners and 1 at the center. To prevent resampling, fibres were 'marked' and length measurements recorded similarly from adjacent tiles. Due to the unavoidable presence of small fragmented DNA commonly found in DNA combing preparations only fibres measuring >1 Mbp were included in the count. Additionally, fibres without definable start and end points were discarded from the analyses. A total of 100 fibres were selected from 3 separate slides and the lengths measured using TissueFAX measurement toolkit. The length distribution obtained from 3 separate preparations is presented in **Figure 3.13**. The average length of combed fibres after protocol optimisation was measured to be ~2.5-3 Mbp with the longest individual fibre captured measuring ~6 Mbp (**Figure 3.14**).

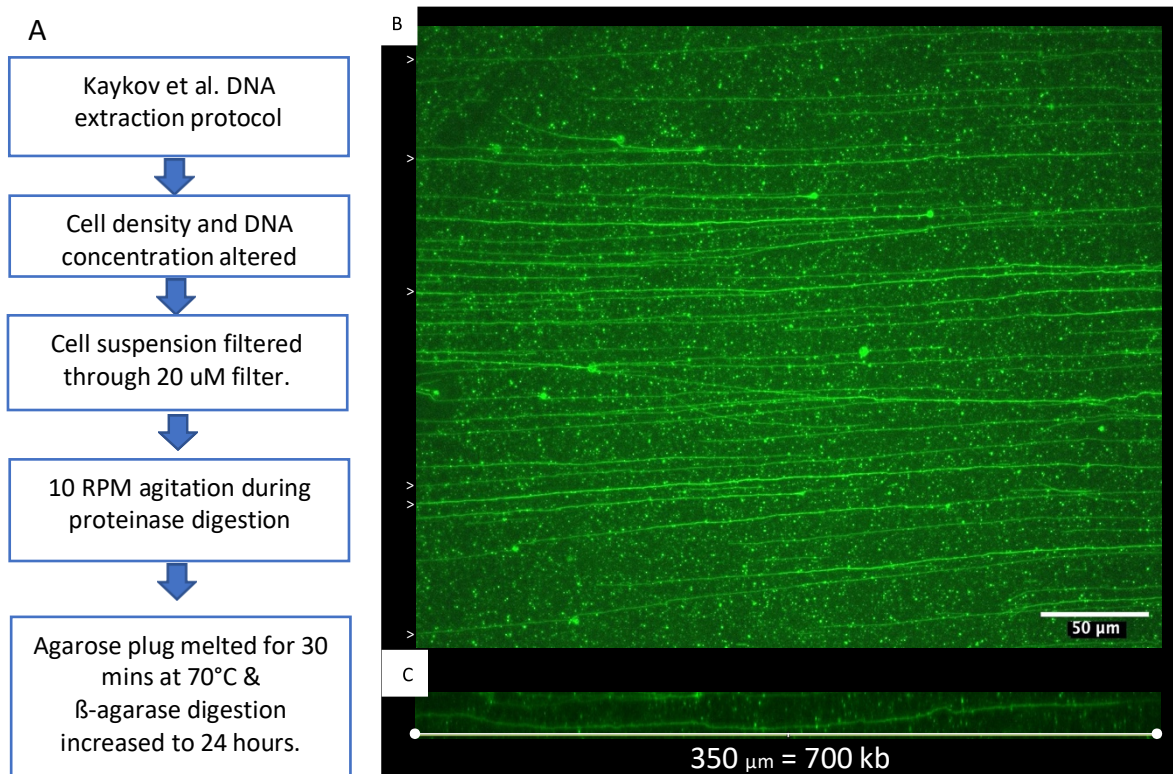


Figure 3.12 Optimised combing protocol allows for the isolation of ultra-long DNA fibres. (A) A schematic outlining the development of the DNA combing protocol (B) A representative image of combed DNA showing linear, individualised fibres. Marked are 6 fibres spanning the entire field of view (C) A single fibre spanning the entire field of view measures ~350 µm/700 kbp).

DNA extracted from 0.25×10^6 MEF cells embedded in LMP agarose plugs, dissolved in 2 ml of 50 mM MES (pH 5.5). Combed DNA stained with 0.1µm YOYO-1 iodide, visualised by Epi-Fluorescence microscopy, scale bar= 50 µm/100 kb, 40 X Magnification.

Figure 3.12B is the same as Figure 3.11C, presented here in the context of protocol development.

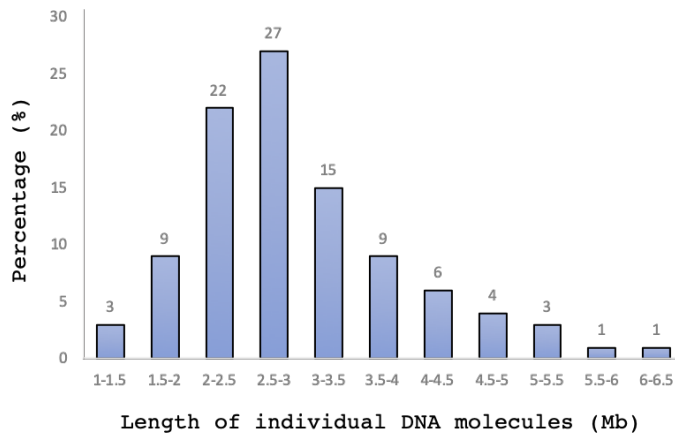


Figure 3.13 Histogram displaying size distribution of combed DNA fibres. The length of 100 randomly selected combed fibres was measured and the frequency plotted. Fibres measuring < 1 Mbp were omitted during analysis. Bin size=0.5 Mb. Length assessment was carried out using Image J measurement toolbox

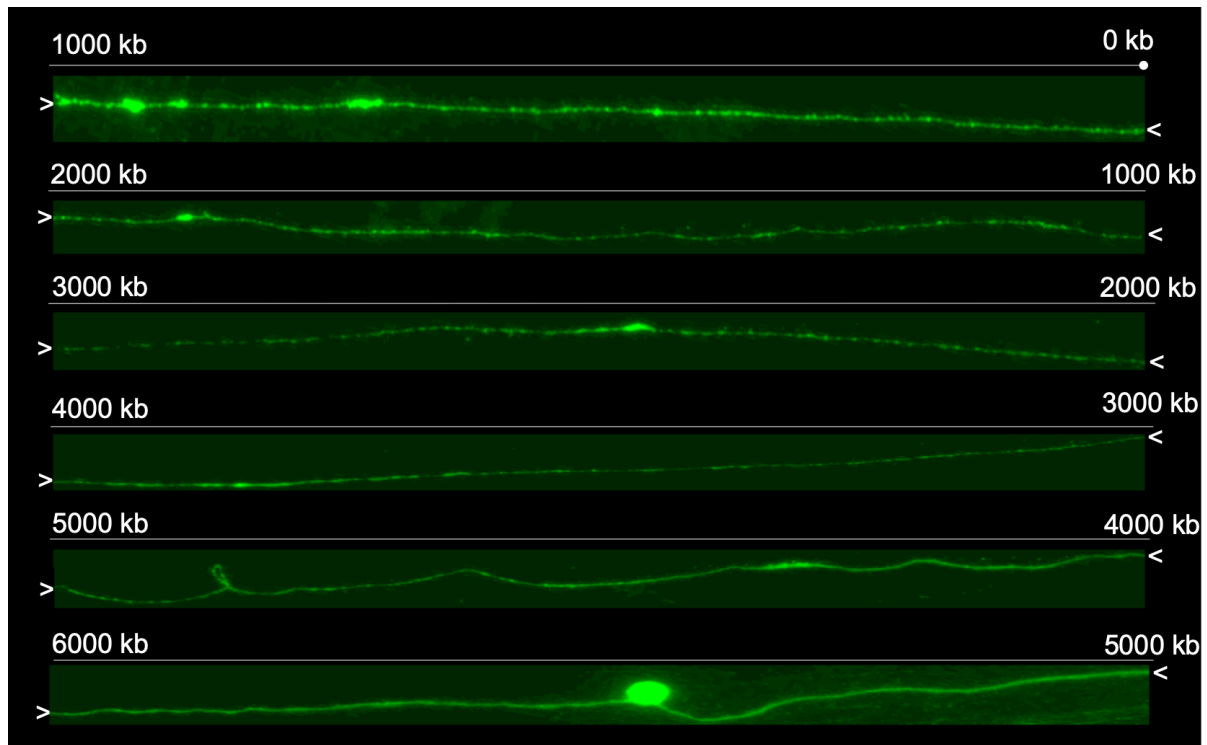


Figure 3.14 Composite image of longest combed DNA fibre. A representation of the longest continuous, non-overlapping fibre captured, measuring ~ 6000 kbp (6 Mb). To present in its entirety the molecule was "cut" in-silico into 6 fragments to construct a composite picture.

Combed DNA stained with 0.1µm YOYO-1 iodide, visualised by Fluorescence microscopy using TissueFaxs Inverted Plus slide scanner, scale bar= 50 µm/100 kb, 63 X Magnification.

Fibre length analysis assessed using image J measurement toolbox.

3.3.6 Visualising genomic loci

To visualise the capture of entire rDNA clusters and explore the orientation, cluster size and chromosome-specific arrangement of rDNA in C57BL/6J, DNA combing was combined with fluorescent in situ hybridisation (FISH). The rDNA coding unit was dissected into two portions with a fluorescently labelled probe designed against the 18S coding unit (~1.8 kbp) and another targeting the sequence encompassing 5.8S and 28S coding units as well as spanning ITS² (~5.9 kbp) (**Figure 3.15A**). Additionally, chromosome-specific probes were generated to map the chromosome-specific positioning of rDNA in C57BL/6J.

The segments of rDNA designated as 18S and 5.8s-28S were successfully PCR amplified, and the amplicon sizes were confirmed via gel electrophoresis (**Figure 3.15B**). Chromosome 12 (Chr 12), one of 6 mouse chromosomes considered to retain rDNA was selected as an initial target and a Chr 12 bacterial artificial chromosome (BAC) was used here as a template. To generate sufficient quantities of material for rDNA probe synthesis, both 18S and 5.8S-28S were individually ligated into pCR™-Blunt II-TOPO™ Vectors to yield recombinant plasmids 18S-TOPO and 5.8S-28S-TOPO respectively. Recombined vectors were transformed into chemically competent bacterial cells and ligation was confirmed via blue-white colony selection. Colonies transformed with successfully ligated plasmids were selected for expansion, and from the resulting cultures, plasmids were extracted and isolated. To confirm amplicon integration, isolated plasmids were subjected to EcoRI restriction digest. Digestion of 18S-TOPO yielded a 3.5 kbp and 1.8 kbp band corresponding to the vector backbone and 18S fragment respectively, whilst digestion of 5.8S-28S-TOPO yielded a 3.5 kbp and 5.9 kbp band corresponding to the vector backbone and 5.8S-28S fragment respectively (**Figure 3.15C**).

Probes utilised in FISH experiments were generated through nick translation of template plasmids 18S-TOPO, 5.8S-28S-TOPO and Chr 12 BAC and subsequent fluorescent dUTP incorporation. Plasmid 5.8S-28S-TOPO was labelled with SpectrumRed dUTP whilst both 18S-TOPO and Chr 12 BAC were labelled with SpectrumGreen dUTP. Labelling efficacy was determined by assessing the size distribution of labelled DNA via gel electrophoresis in addition to fluorophore incorporation via spectroscopy. Labelling via nick translation results in the partial digestion of the template DNA, and is expected to produce probes with a fragment size of 50-500 bp, confirmed via gel electrophoresis (**Figure 3.15D**). Labelled 5.8S-28S-TOPO and 18S-TOPO appeared as a smear within the expected range. Labelled Chr 12 BAC presented as 2 smears, 1 at the lower end of the expected range, and an additional band at >10 kb, perhaps suggesting sub-optimal nick translation and incomplete labelling of template DNA.

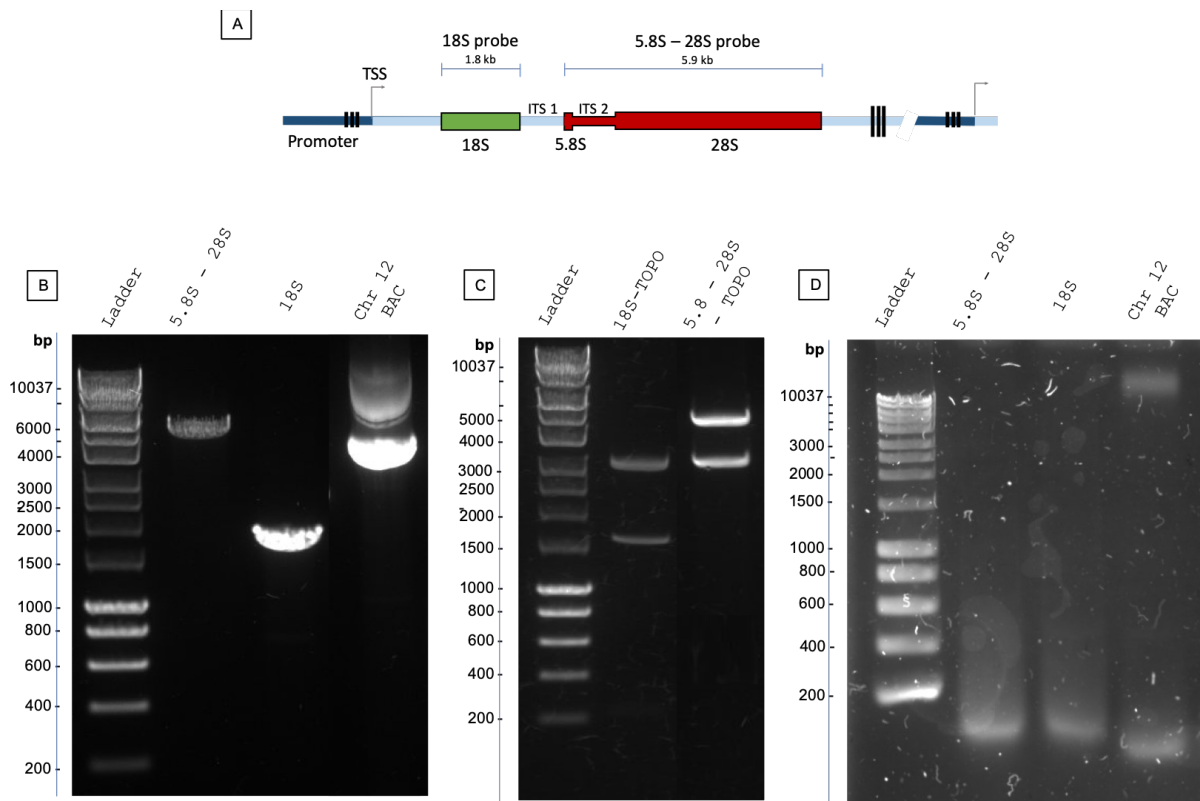


Figure 3.15 Synthesis of DNA FISH probes. (A) Schematic of a single rDNA coding unit depicting the sequence coverage of DNA FISH probes designed to target the 18S (green) and 5.8-28S (red) (B) 1 % agarose gel image showing 5.8S-28S (5.9 kbp) and 18S (1.8 kbp) PCR amplicons in addition to banding confirming the integrity of chromosome 12 BAC (C) 1 % agarose gel image showing the products of EcoRI digest of 18S- TOPO™ Vector and 5.8S-28S- TOPO™ Vector (D) 1 % agarose gel image evaluating Nick translation efficiency and the size distribution of 5.8S-28S, 18S, chromosome 12 fluorescently labelled probes.

To assess the binding capabilities of labelled DNA probes in a simpler experimental context than combed DNA, FISH was initially carried out on fixed MEF metaphase nuclei. Fixed nuclei were initially probed with 5.8S-28S (red) and Xist (green) dUTP labelled probes, the latter was used as an experimentally established positive control. As expected, Xist localisation was observed as two intense spots of pinpoint fluorescence at the nuclear periphery with the localisation of 5.8S-28S seen largely dispersed within the nuclear body (**Figure 3.16A**). Due to the high level of background fluorescence and the dispersed nature of DNA, the confident quantification of rDNA cluster number proved difficult within nuclei. However, multiple regions of varying fluorescence intensity were observed with 4-6 distinct regions assigned to 5.8S-28S rDNA localisation (**Figure 3.16B**). To further ensure the observed signal for labelled 5.8S-28S probes was not due to non-specific binding and background fluorescence, the probes were hybridised to metaphase chromosome spreads. **Figure 3.16C** presents a representative metaphase chromosome spread displaying over 40 chromosomes, likely originating from 2 separate cell nuclei. The chromosome number though not exactly quantified due to significant

overlapping can be estimated by counting the number of distinct chromosomes and also intensely fluorescent DAPI-stained centromeres which exist at a 1:1 ratio to chromosomes. Overlapped with this, is the co-localised signal detected from the hybridisation of the 5.8S-28S probes. In total, there are 16 distinct regions of fluorescence circled which indicate the 45S rDNA loci to which the 5.8S-28S probe localises, of which there are 8 for each of the 2 metaphase cells represented in the spread. To further evaluate the 45S rDNA specificity of the 5.8S-28S probe and assess the effectiveness of the generated 18S probe, nuclei were co-hybridised with both probes. **Figure 3.16Di** shows the distinct yet disperse nuclear localisation of the 5.8S-28S probe, with **Figure 3.16Dii** showing similar localisation patterns for the 18S probe. For clarity, the fluorescence channels have been separately presented for each probe, however, it is evident that regions of fluorescence intensity show a great degree of overlap between 5.8S-28S and 18S localisation, indicating that both rDNA probes likely target the same genomic regions. Confident that rDNA could be targeted with the designed set of probes, it was important to next identify the chromosomes from which the observed rDNA signals originated. A mixed preparation of metaphase nuclei and chromosomes was hybridised with probes for 5.8S-28S and Chr12 (**Figure 3.16E**). Whilst localisation of 5.8S-28S was observed both within nuclei and on individual chromosomes, no visualisation of chromosome 12 localisation was observed.

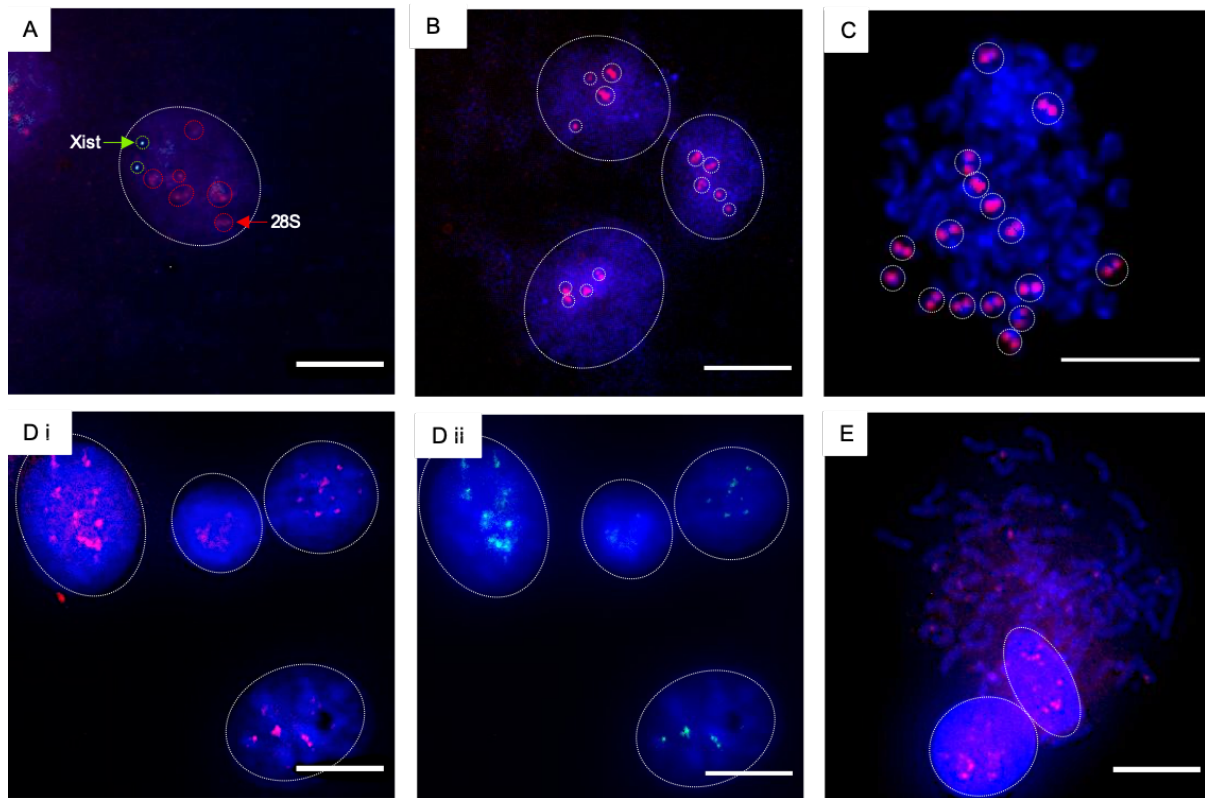


Figure 3.16 Visualisation of rDNA in fixed MEF nuclei and chromosome spreads. Fluorescent *In situ* hybridisation on fixed metaphase nuclei and chromosomes (A) Nuclei (circled) labelled with Xist (SpectrumGreen dUTP) and 5.8S-28S (SpectrumRed dUTP) probes. Representative signal for respective probes circled. (B) Varying 5.8S-28S signals clusters (C) Metaphase Chromosome spread labelled with 5.8S-28S (SpectrumRed dUTP) (Di) Nuclei (circled) labelled with 5.8S-28S (SpectrumRed dUTP) and 18S (SpectrumGreen dUTP) probes. Split channel images for each probe are presented for clarity (i) image filtered for 5.8S-28S (SpectrumRed dUTP) (ii) image filtered for 18S (SpectrumGreen dUTP) probes (E) Fixed metaphase nuclei and chromosomes labelled with chromosome 12 (SpectrumGreen dUTP) probes and 5.8S-28S (SpectrumRed dUTP) probes. Signal for chromosome 12 localisation is not observed, 5.8S-28 probe localisation observed within nuclei and on individual chromosomes. (A-E) counterstained with DAPI. Scalebar =20 μ m.

3.3.7 Generation of control cell lines for testing specificity of SNP-CLING probes

To evaluate the specificity of SNP-CLING probes for SNP recognition and rDNA variant differentiation it was important to first create a controlled environment in which probes could be tested. This would be achieved through the generation of chimeric cell lines, specifically HEK-293T cells expressing either variant of the C57BL/6J rDNA promoter. Firstly, it was necessary to isolate the individual promoter variants, to do this a 216 bp fragment of the C57BL/6J rDNA promoter spanning from position -48 to -264 upstream of the TSS was selected for recombination (**Figure 3.17A**). The promoter sequence encompasses sites of interest including SNPs at -178, -104 and CpG -133, and was PCR amplified with additional flanking restriction sites for EcoNI and MfeI and TA overhang generating a 235 bp amplicon (**Figure 3.17B**). The amplicon was then cloned using PCRTM 4-TOPOTM vector (**Figure 3.17C**) and competent bacteria were transformed. Bacterial colonies transformed with successfully recombined

plasmid were identified via blue-white selection and 12 individual colonies were selected at random for sequencing. M13 primer sites flanking the integrated sequence were utilized for Sanger sequencing, with representative sequence alignments presented in Figure 1.15D. An 80 bp sequence (-179 to -100) for isolated A and C promoter variants is aligned to reference BK000964.3 with key motifs highlighted. For the sequence presented the Isolated C variant showed 100% sequence similarity to ref BK000964.3 whilst the A variant shows both G to C and C to A SNPs at positions -178 and -104 respectively. Out of the 12 colonies sequenced, 3 were identified as C variant and 6 as A variant transformed, whilst 3 were matched unsuccessfully, with one colony being selected in the case of each variant for plasmid cloning and promoter fragment isolation.

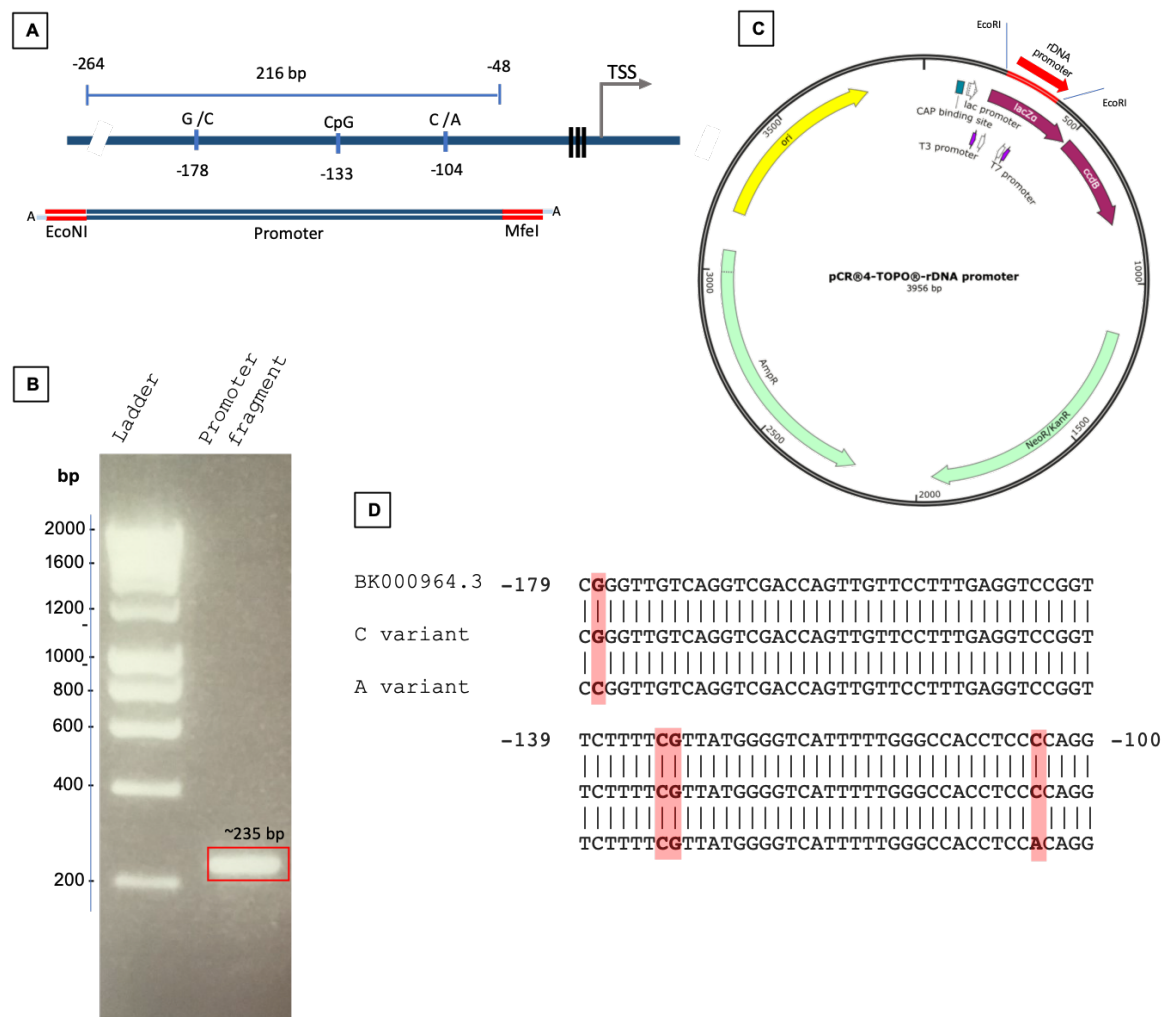


Figure 3.17 Isolation of rDNA genetic variants. (A) Schematic of rDNA promoter with annotated 216 bp section encompassing -178, -104 SNPs and -133 CpG site, alongside a schematic of the rDNA promoter fragment flanked with artificially introduced restriction sites for EcoNI and MfeI with additional 3' A overhangs (B) Annotated vector map for recombinant PCR™ 4-TOPO™ with cloned rDNA promoter fragment (highlighted in red) (C) Image of 1% Agarose gel showing the PCR amplification of rDNA promoter fragment + flanking restriction sites for EcoNI and MfeI (235 bp) (D) Sequence alignment showing 80 bp sequence for isolated A and C promoter variants against reference BK000954.3. SNPs at -178, -104 and -133 CpG highlighted. A and C rDNA promoter variants sequenced using Sanger sequencing.

To integrate the isolated C57BL/6J rDNA promoter variants into HEK-293T cells and generate stable control cell lines, lentiviral transduction was chosen as the most suitable approach. For each variant, lentiviral transfer plasmids containing either an antibiotic selection marker or fluorescent report were selected, with A and C variant fragments assembled into vectors pLenti-puro and pLJM1-EGFP respectively (**Figure 3.18A**). To confirm sequence integration, native and recombined vectors were subjected to a double restriction digest with either MfeI and NdeI or EcoNI and MfeI respectively, with digest products run on a 1% agarose gel for confirmation (**Figure 3.18B**). Digestion of native pLJM1-EGFP and pLenti-puro with MfeI and NdeI yielded fragments measuring ~7031/1052bp and ~6093/1027 respectively as expected, whilst digestion of pLJM1-EGFP-C Variant and pLenti-puro-A Variant with EcoNI and MfeI yielded fragments measuring ~8064/221 bp and ~7047/221 bp respectively. In the case of recombined transfer vectors, the liberation of fragments measuring ~221 bp indicated successful rDNA promoter fragment integration.

For the generation of transduction efficient lentiviral particles carrying either A or C rDNA promoter variants, HEK-293T packaging cells were transfected with recombined transfer vectors pLJM1-EGFP-C-Var and pLenti-puro-A-Var, alongside packaging and envelope plasmids psPAX2, pMD2.G yielding particles termed Lenti-C and Lenti-A respectively. Packaging cell supernatant was collected at 24, 48 and 72 hours post-transfection and the presence of lentivirus was qualitatively assessed using Lenti-X™ GoStix™. Lentiviral particles were confirmed in all 3 collections with the supernatant collected at 48 hours used for transduction. HEK-293T cells transduced with Lenti-A, termed HEK-293T-A were subjected to puromycin selection 24 hours post-transduction with growth monitored for an additional 3 days. Cell survival was crudely assessed using trypan blue exclusion, on days 1,2, 3 and 4 post puromycin selection initiation and compared to both untreated WT HEK-293T and treated WT HEK-293T cells (**Figure 3.18Ci**). As expected normal cell growth is observed in all 3 conditions from day 0-1 with continuous steady growth observed for untreated WT HEK-293T from day 1-4. Treatment of WT HEK-293T cell with 2.5 µg/ml puromycin results in a noticeable reduction in cell viability on day 2 with complete death observed by day 4. Treatment of HEK-293T-A with 2.5 µg/ml puromycin results in a slight reduction in cell viability between days 1-2 accounting for the death of non-transduced cells within the population, however, cell growth picks up from days 2-4 as those cells successfully transduced continue proliferating. Non-transduced cells were further eliminated by the continuation of puromycin selection for an additional 3 days. HEK-293T cells transduced with Lenti-C termed HEK-293T-C were assessed for GFP expression as an indicator of successful transduction (**Figure 3.18C ii**). Additionally, 24 hours after transduction cells were subjected to puromycin selection for 3 days to ensure the elimination of any non-transduced cells within the population.

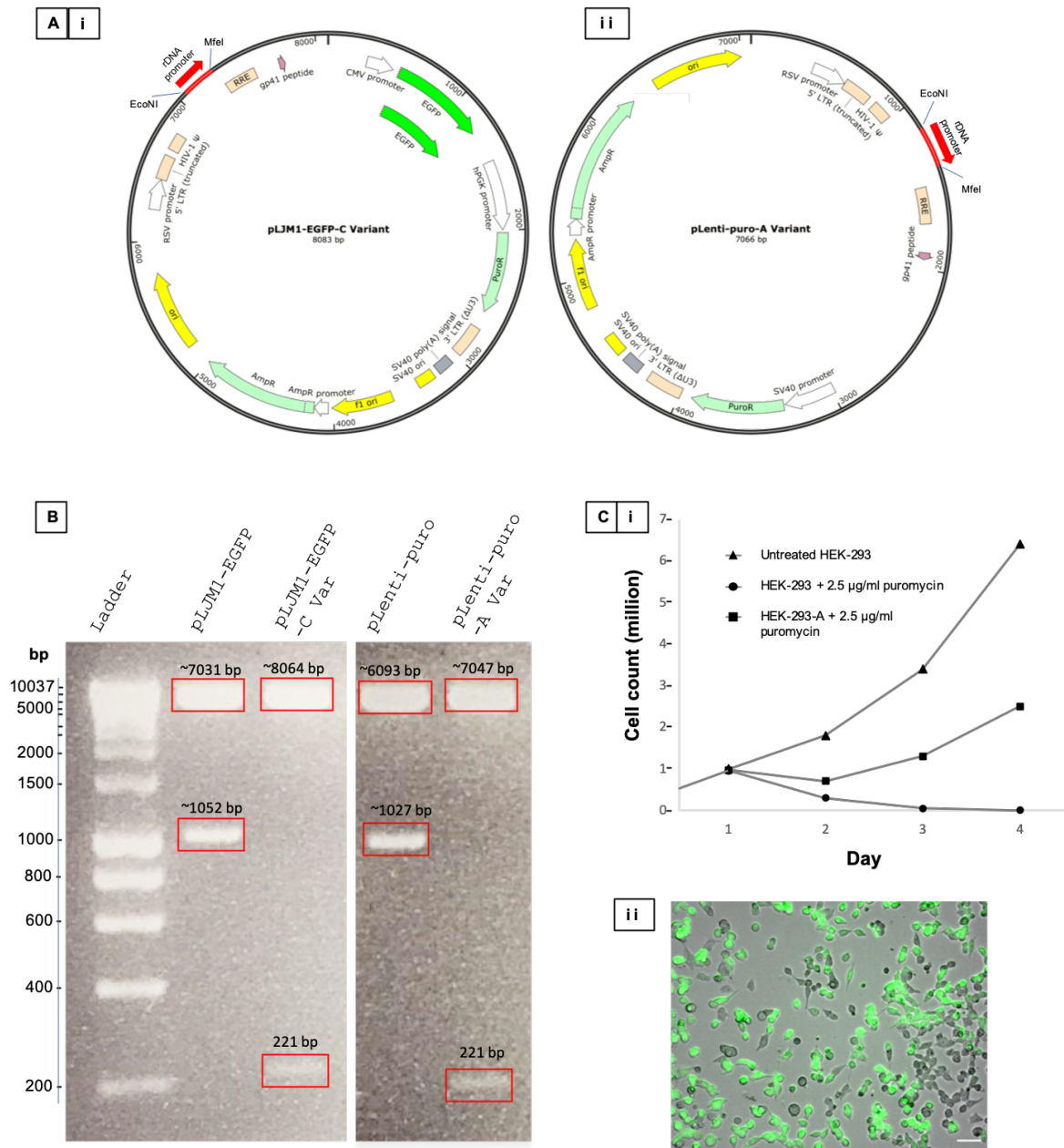


Figure 3.18 Generation of stable HEK-293T cell lines with integrated rDNA promoter sequences. (A) Schematic of 3rd Gen Lentiviral transfer constructs (i) Simplified schematic of pLJM1-EGFP transfer vector with rDNA promoter C variant cloned in (ii) Simplified schematic of pLenti-puro transfer vector with rDNA promoter A variant cloned in (B) 1% agarose gel image showing double digest of lentiviral transfer vector constructs. pLJM1-EGFP digested with MfeI + NdeI, pLJM1-EGFP-C Variant digested with EcoNI + MfeI, pLenti-puro digested with MfeI + NdeI, pLenti-puro-A Variant digested with EcoNI + MfeI. (C) Successful transduction of HEK293T cells with lentiviral particles confirmed for (i) Lenti-A Var via puromycin selection (data presented as an average of 2 biological replicates) (ii) Lenti-C Var via GFP expression visualised by Epi-Fluorescence microscopy, scale bar= 50 µm, 10 X Magnification.

3.3.8 Generation of SNP-CLING probes for distinguishing rDNA promoter variants

To utilise SNP-CLING probes on combed DNA fibres, it was necessary to first express and purify the fluorescently tagged dCas9. Two lentiviral expression vectors were selected, pHAGE-TO-dCas9-3xGFP encoding dCas9 fused to 3X GFP (**Figure 1.19Ai**) and pHAGE-TO-dCas9-3xmCherry encoding dCas9 fused to 3X mCherry (**Figure 1.19ii**). Initially, it was hypothesised that the quickest and simplest method of obtaining large amounts of purified fluorophore fused dCas9 would be via bacterial expression followed by affinity purification. To achieve this pHAGE-TO-dCas9-3xGFP vector was double digested with EcoRI and XbaI, liberating the ~6.3 kbp insert encoding dCas9-3xGFP. This coding sequence was then inserted into pSF-OXB20-NH2-6His-TEV, a non-inducible, high-expression bacterial vector with a cleavable N-terminal Hexa-Histidine tag. Once vector recombination was confirmed via gel electrophoresis, competent *E. coli* were transformed and cultured in the presence of kanamycin for selection. This, however, yielded no noticeable expression of dCas9-3xGFP, the absence of which was indicated via the lack of any observable GFP fluorescence as well as a lack of GFP detection via western blot analysis. This approach was quickly abandoned and replaced with mammalian system expression through lentiviral transduction.

For expression of dCas9-3xGFP, HEK-293T packaging cells were first transfected with pHAGE-TO-dCas9-3xGFP, psPAX2, pMD2.G and supernatant collected at 24, 48, and 72 hours. Lentiviral presence was qualitatively confirmed in all three collections using Lenti-X™ GoStix™, with particles collected at 48 hours used for transducing HEK-293T and the expression of dCas9-3xGFP. Transduced cells were cultured for 3 days before the expression was confirmed with Epi-fluorescence microscopy (**Figure 3.19B**) and dCas9-3xGFP was purified using ChromoTek GFP-Trap™ magnetic beads.

The effective targeting of probes to SNP alleles and distinguishing between A and C variants is dependent on the PAM recognition specificity of Cas9. SpCas9 binding is highly dependent on its recognition of a specific protospacer adjacent motif (PAM: 5'-NGR-3') where 'N' is any nucleotide (**Figure 1.19Cii**). By positioning SNPs within the PAM site, binding and subsequent labelling of the respective alleles can either be promoted or hindered. This necessitated designing gRNAs which targeted SNP-CLING probes to sites directly adjacent to SNPs at positions -178 and -104 (**Figure 3.19D**). When considering the SNP at -178 and an additional base on either side of this position, the A variant sequence reads as 5'-CGG-3' on the antisense strand, maintaining SpdCas9 PAM recognition. However, considering this locus in the context of the C variant, the sequence reads 5'-CCG-3' as a result disrupting the SpdCas9 PAM recognition at position -178. Hence SNP -178 may be utilised to target the A variant but not the C variant. Contrastingly, when considering the SNP at -104 and an

additional base on either side of this position, the C variant sequence reads as 5'-GGG-3' on the antisense strand maintaining SpdCas9 PAM recognition. However, considering this locus in the context of the A variant, the sequence reads 5'-GTG-3', disrupting the PAM recognition at position -104. Hence SNP -104 may be utilised to target the C variant but not the A variant.

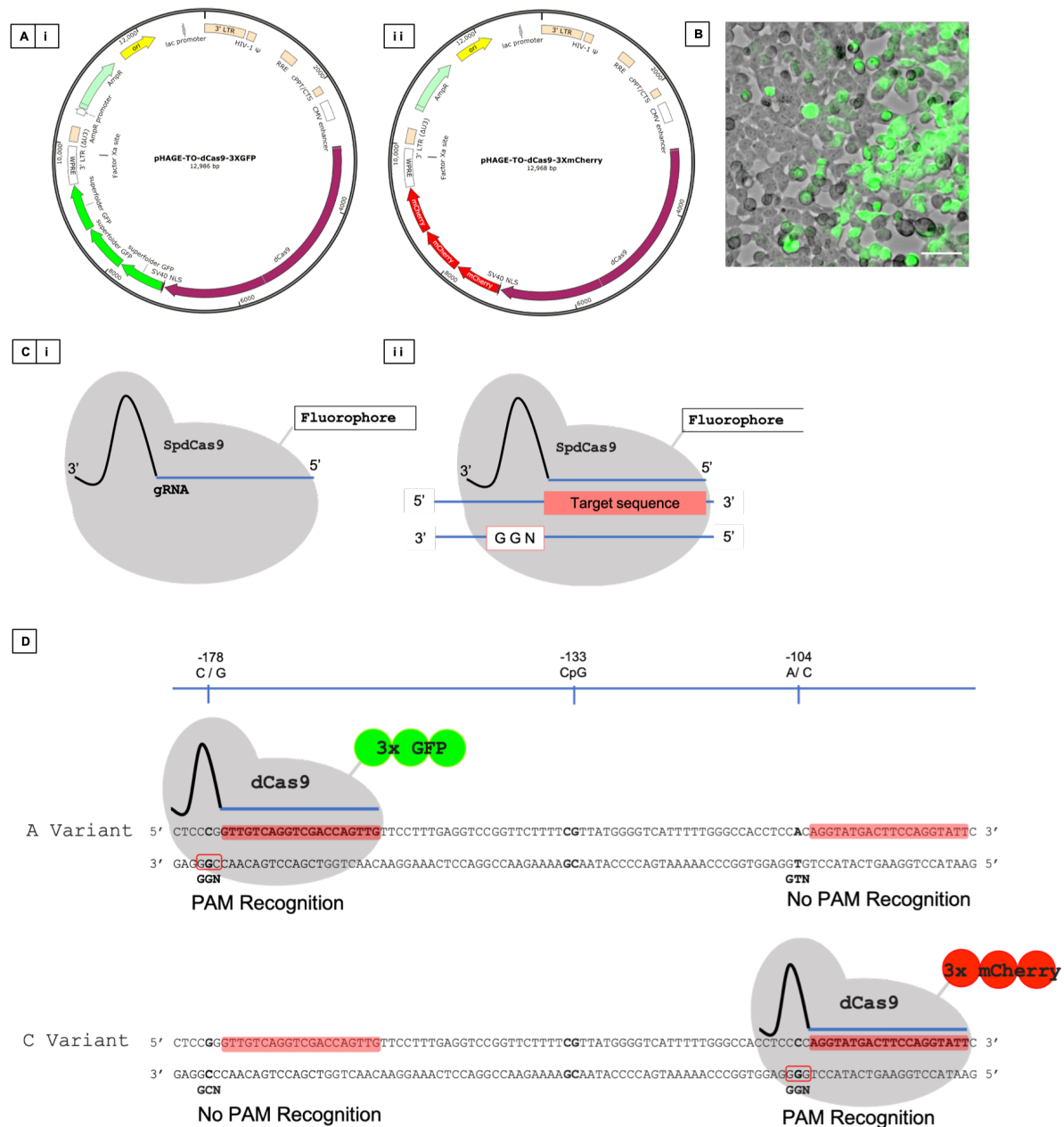


Figure 3.19 SNP-CLING probes generation. (A) Lentiviral vectors for mammalian expression of (i) SpdCas9 tagged with 3X GFP fusion protein (ii) SpdCas9 tagged with 3X mCherry fusion protein. (B) HEK-293T cells transduced with lentiviral particles generated from pHAGE-TO-dCas9-3XGFP, psPAX2, pMD2.G. Expression visualized using Epi-fluorescence microscopy, GFP signal overlapped on brightfield image. (C) Schematic of (i) SpdCas9-gRNA complex, black segment= tracrRNA/ blue segment = crRNA and (ii) SpdCas9-gRNA mechanism of PAM (3'-GGN-5') dependent binding. (D) Schematic of SpdCas9 SNP specific binding to rDNA A- and C- promoter variants. A 101 bp sequence for both promoter variants is shown, with nucleotides of interest (-178, -133, -104) annotated. A and C variants may be visualised by exploiting the SNPs at positions -178 and -104 respectively.

3.4 Discussion

The work in this chapter has fulfilled the first aim of optimising and applying the method of DNA combing to isolate ultralong multi MB length combed DNA molecules. Furthermore, the progress made toward dissecting the allele-specific landscape of ribosomal DNA clusters has also been outlined. Here the challenges faced in achieving these aims will be examined, as well as any potential experimental improvements that may have been implemented. Also, there will be an exploration of the next steps that may be taken to achieve the outlined project aims and a discussion of any potential pitfalls.

3.4.1 DNA combing optimisation

DNA combing can serve as an invaluable tool to visualise and study genomic loci and is particularly useful in the exploration of long repetitive stretches of DNA, regions which remain elusive to modern sequencing technologies and in theory complete chromosomes may be captured in their entirety. This potentially powerful method though theoretically simple can be highly variable, and success is largely dependent on a variety of specific chemical and physical conditions. This chapter outlines the optimisation of the DNAcombing protocol for the isolation of MEF genomic DNA molecules routinely measuring more than a Mbp. Using a protocol published by Kaykov et al., (2016) as a framework for further optimisation, single molecules routinely measuring > 2.5 Mbp and occasionally exceeding 5-6 Mbp in length were obtained. These results have highlighted the importance of surface functionalisation and buffer pH compatibility, as well as the need for effective proteinase digestion and DNA resuspension before combing, to obtain a highly homogenous, linearised and individualised fibre array.

The effectiveness of contemporary molecular combing applications primarily relies on the density and alignment of the combed DNA. Both are reliant on the quality of the coverslip surface modification, which must be uniform to support DNA binding and stretching throughout the entire surface. Different surface modifications have been investigated, however, the coating of the surface with an octenyl carbon chain is thought to provide the best potential stretching of the DNA fibre (Bensimon *et al.*, 1994; Allemand *et al.*, 1997). This prompted the evaluation of Trimethoxy-octenylsilane and 7-Octenyltrichlorosilane modified surfaces for their compatibility with DNA combing, with both compounds being used with success in previous studies. Liquid phase Trimethoxy-octenylsilane treatment was demonstrated by Labit et al., (2008), to be a quick and simple method of producing

DNA combing compatible surfaces permitting both the adherence and stretching of DNA molecules. However, when tested, this method led to non-specific DNA binding and a complete lack of molecule linearisation. This could either be due to poor surface modification resulting from a heterogeneously modified surface, or the use of an incompatible pH.

Next, an attempt using gas-phase 7-Octenyltrichlorosilane modified surfaces was made, previously utilised by Kaykov et al., (2016). Though this method produced well-adhered and linearised fibres at the tested pH, the surface was non-homogeneously modified, and difficult to reproduce accurately. Gas-phase silanisation requires the use of specialised incubators with regulated anhydrous conditions, which are difficult to come by in standard biology labs, with sub-optimal equipment likely leading to the observed outcome. For this reason, professionally manufacture slides, specifically sigma Aldrich aminoalkyl silane prep slides were tested and were shown to permit binding and fibre combing to some degree. The use of a buffer with pH 6 proved compatible with 7-Octenyltrichlorosilane modified surfaces however resulted in the described 'U' shaped attachment of DNA fibres when combined with aminoalkyl silane-modified surfaces. Testing combing buffers ranging from pH 5- 6.5 demonstrated that pH 5.5 was compatible with aminoalkyl silane-modified surfaces. This aligns with previous observations showing that combing efficiency is greatly dependent on a combination of specific buffer pH and surface functionalisation (Allemand et al., 1997) DNA combing is extremely sensitive to pH fluctuations, and a uniform spread of linearised, parallel DNA molecules can only be obtained within a small pH window, the optimal pH further varies depending on the surface chemical modification used.

Secondly, DNA homogenisation is of paramount importance in obtaining a spread of well-separated fibres. This study demonstrates the importance of factors including singularisation of cells in suspension, effective proteinase digestion as well as the complete release of DNA from the agarose matrix. Failure at any of these steps may lead to sub-optimal DNA dispersal. These are a collection of technical particularities that need to be optimised within each laboratory setting and may be highly variable depending on the practitioner as well as the tools used. For instance, a single cell suspension may be achieved simply by carefully resuspending a cell pellet in a volume of solution by pipetting, however, the result can vary depending on the pipetting speed and time. To overcome potential pipetting irregularities, resuspension via size dependant filtration was shown to be highly effective, consistently preventing DNA clumping. Regarding proteinase digestion, the duration of this step varies between protocols, with Kaykov et al. (2016) recommending 48 hours of digestion at 50°. Most protocols make no mention of physical aid to assist digestion. This study has observed that the addition of constant gentle agitation throughout the proteinase digestion step with the use of a

rotator at 10 RPM aided in noticeably improving fibre resolution and minimising fibre bundling. It is important to note however that increased agitation, for instance, 20 RPM led to the destruction of the agarose plug and subsequently negatively impacted the combing process. Finally, the release of DNA from the agarose matrix relies on the complete melting and digestion of the agarose matrix, steps which are both time and temperature-sensitive. The Genomic vision protocol recommends an initial melting step at 68°C for 20 minutes whilst Kaykov et al. (2016) suggests 70°C for 15 minutes, both followed by overnight β -agarase digestion at 42°C. These however lead to poor melting and digestion when tested here. Instead, increasing the melting time to 30 minutes and β -agarase digestion to 24 hours led to a better release of DNA from the agarose matrix. Overall, these observations align with previous studies showing that DNA at high concentrations behaves like a polymer, and unbound molecules tend to entangle into larger polymer meshes which deposit as poorly combed patches onto the surface (Michalet *et al.*, 1997). Considering this it is reasonable to imagine that factors influencing DNA concentration, directly or indirectly could lead to poor combing outcomes.

The amendments to the combing protocol noted in this chapter include an increase in mechanical and thermal forces with increased incubation times. Though these amendments may have contributed to the reduction in fibre length when compared to those reported by Kaykov et al., (2016), this study has achieved the capture of DNA molecules often measuring >2 Mbp and up to ~6 Mbp. Considering the varying size of rDNA clusters, it is difficult to say with certainty if molecules of this length will be sufficient for capturing entire rDNA clusters. A fairly sizeable rDNA cluster containing 100 repeats (~45 kbp per unit) can be expected to span ~4.5 Mbp. It is therefore likely that the fibres captured here would allow for the study of the large-scale arrangements of rDNA cluster. In future iterations of the combing experiment, the length of DNA molecules captured may be increased by controlling the mechanical and chemical forces applied to DNA throughout the extraction and combing process. Additionally, combed DNA may be enriched for rDNA via the restriction digest of DNA in agarose plugs with specific nucleases that cleave non-rDNA sequences. For instance, the restriction enzyme, PmeI, recognises 5'-GTTT^AAAC-3' sites which are not found within the rDNA gene consensus sequence however this sequence is found in multiple places in the mouse genome. Therefore, digesting DNA with PmeI would theoretically cut most of the DNA into small fragments but leave each rDNA cluster intact. This may enrich combed DNA for rDNA clusters and facilitate probing and analysis, however, could come at the cost of potentially digesting chromosomes specific sequences which could hinder the identification of the chromosomes from which the clusters originate.

3.4.2 Probing genomic loci

Ribosomal DNA comprises a highly variable region of a genome and is organised as clusters of variable size, location and arrangement. The literature remains in disagreement regarding these aspects of rDNA distribution, with the number of clusters, the loci and the copy number known to differ among species, population, and even individuals (Jhanwar, Prensky and Chaganti, 1981; Kopp, Mayr and Schleger, 1988). Considering the high variability and lack of consensus between studies it was important to establish these features within the cell line used in this study and gain a more thorough understanding of the rDNA landscape in C57BL/6J.

Specifically, we sought to identify the rDNA carrying chromosomes with chromosome-FISH as well as probe the size of each array and the arrangement of individual units within entire clusters captured with molecular combing. Here, 45S rDNA was successfully localised in fixed nuclei as well as on metaphase MEF chromosomes. As expected, localisation was observed to be paracentral on multiple chromosomes. A simplified approach to identifying the rDNA-containing chromosomes was made by labelling certain chromosomes of interest individually and sequentially alongside 45S rDNA to deduce the specific chromosome localisation. However, attempts at identifying the specific chromosomes to which the 45S rDNA probes hybridised proved unsuccessful. Additionally, initial attempts made at probing rDNA arrays on combed DNA were also unsuccessful. These unsatisfactory outcomes are likely due to the ineffectiveness of the probes used as well as a lack of experience and skill with the methodologies, the development of which could likely lead to increased success at visualising rDNA both on metaphase chromosome spread and at the single-molecule level on combed DNA.

Regarding the identification of rDNA loci on metaphase chromosomes the approach used here may be considered crude and on a large scale is both time and resource-intensive. As alternatives to the approach used in this study, methods such as multicolour fluorescence in situ hybridization (M-FISH) (Speicher, Ballard and Ward, 1996) or Multi-colour banding (M-Band) (Chudoba *et al.*, 2004) could also be employed to identify the rDNA carrying chromosomes in C57BL/6J. To achieve this 45S rDNA labelling may be coupled with the 'painting' of chromosomes with a spread of spectrally distinct probes creating a unique chromosome-specific chromo signature. These signatures can in turn be used to identify specific chromosomes with the additional 45S rDNA probes localising rDNA carrying chromosomes. Alternatively, concurrent FISH labelling of 45S rDNA with traditional karyotyping methods may also be used to identify 45S rDNA chromosomal localisation. Karyotyping relies on staining chromosomes with dyes which bind in a chromosome-specific manner to create banding patterns unique to each chromosome. Specifically, the use of fluorochrome dyes has permitted

chromosomes to be simultaneously banded and hybridized in situ with probes to gain complete insights (Christian *et al.*, 1998). DAPI is one such tool in the fluorescent karyotyping toolbox, at low DAPI: DNA concentration DAPI binds preferentially to the minor groove AT-rich sequences of DNA, with this sequence selectivity creating unique chromosome banding patterns used for identification (Heng and Tsui, 1993). This approach though seemingly straightforward requires a near-perfect spread of metaphase chromosomes, as well as staining methods which minimise background coupled with advanced microscopy analysis, and an experienced eye to discern the subtle banding differences between chromosomes. To bypass these issues, flow cytometry could be employed initially to isolate specific chromosomes of interest according to optical parameters. This is an invaluable method for chromosome enrichment and is widely used to facilitate the study of genome sequencing, target development of DNA markers and gene cloning (Kovářová *et al.*, 2007; Kuderna *et al.*, 2020). Flow cytometry-based enrichment of chromosomes 1 has been used in combination with long-read nanopore sequencing to facilitate and simplify genome assembly (Kuderna *et al.*, 2020). Theoretically, chromosomes sorted in this way could be used as the starting material for both metaphase spreading and DNA combing, circumventing the need for hybridisation-dependent chromosome identification. Though many methods are available for discerning the 45S rDNA chromosomes, ultimately the approach adopted is dependent on the resources and expertise available.

With regards to visualising rDNA genetic variants, control HEK-293T cell lines with stably integrated C57BL/6J promoter variants A and C were established. These cell lines would serve as a controlled environment in which to test the binding capacity and SNP specificity of dCas9 probes, however, this remains to be tested. Furthermore, fluorescent molecule fused dCas9 complexes were synthesised and purified to be used in an extracellular environment, and gRNAs exploiting SNPs at -104 and -178 were designed. These probes were intended for use alongside dCas9 probes designed for chromosomes specific targeting, however, this remains to be tested.

Though the specific experimental pitfalls remain to be observed, potential challenges are outlined below. One limitation of allele-specific visualisation using this method is the limited targets which can be exploited. In the context of this experimental set-up, the “A” and “C” promoter variants may only be visualised through the exploitation of SNP -178 and -104 respectively limiting the pool of potential gRNAs, with any hindrances to gRNA binding challenging this approach. Fortunately, since the report of distinct rDNA promoter variants A and C, distinct genetic haplotypes have been categorised in C57BL/6J mice (Rodriguez-Algarra *et al.*, 2022), with further characterisation of these haplotypes

strongly linking them to the promoter variants. This approach provides a greater degree of freedom for dCas9 targeting and may be utilised secondary to directly probing promoter SNPs.

Additionally, though dCas9 nuclease labelling has proven useful in visualising genomic loci in fixed cells (Deng *et al.* etc) and chromatin dynamics in live cell (Maass *et al.*, 2018) its use on combed DNA molecules is yet to be explored. Regardless, the employment of Cas9 nuclease for *in vitro* labelling and DNA editing (Liu *et al.*, 2015) has been reported with great success and may be used for the development of a method compatible with combed DNA fibres. A study by Mikheikin *et al.*, (2016) describes a labelling technique (CRISPR-Cas9 nanoparticles) for high-speed AFM-based physical mapping of DNA in which dCas9 is preassembled with gRNA and the complex is then crosslinked to target DNA and imaged by AFM (Mikheikin *et al.*, 2017). Alternatively, another study reports CRISPR/dCas9-mediated labelling of genomic DNA for optical mapping in conjunction with BioNano Genomic technology (Zhang *et al.*, 2018). It is unclear how these approaches will fare when used in conjunction with combed DNA, although they remain promising alternatives to explore.

3.4.3 Conclusions

This chapter aimed to characterise mouse rDNA clusters with respect to the arrangement of environmentally sensitive genetic variants. By using a combination of molecular combing and SNP-specific CRISPR/dCas9 probes I sought to gain a deeper understanding of the structure of rDNA clusters and the arrangements of rDNA variants at the single-molecule level. Progress was made in establishing a combing protocol allowing for the isolation of DNA molecules measuring 5-6 Mbp and SNP-specific CRISPR/dCas9 probes were generated for labelling variants. However, due to time constraints and hindrances resulting from the global pandemic, the project was halted prematurely and these methods were not fully optimised or applied.

Establishing this methodology could allow for bypassing the current limitations imposed by long read sequencing methods, providing a cost-effective alternative to studying both the structure and dynamics of rDNA arrays. If given the opportunity, further optimization of the combing process could possibly lead to the capture of entire rDNA clusters which could be visually probed at the single molecule level. By probing isolated rDNA arrays with SNP specific rDNA allele probes in conjunction with chromosome specific probes, I would hope to explore the arrangement of environmentally sensitive rDNA alleles within intact arrays and their chromosomal positioning in our C57BL/6J strain. Employing this methodology would hopefully gain us deeper insights into the arrangement of and interplay between distinct rDNA variants. Coupling single molecule analyses with FISH may also allow the quantification of rDNA array copy number in a chromosome specific manner, elucidate

Long-Read Sequencing Analysis of Ribosomal RNA Modifications array morphology and provide a means to study chromosome specific copy number dynamics. This pioneering work could serve to not only establish, once and for all, the large scale structural features of rDNA arrays in complex organisms, but also form the basis for future studies exploring the spatial, structural and epigenetic state of rDNA at the cluster level.

4 Long-Read Sequencing Analysis of Ribosomal RNA Modifications

4.1 Introduction

4.1.1 Aims

The work in this chapter aims to build on the finding of a recent study conducted by our lab, demonstrating the existence of distinct ribosomal DNA haplotypes, linked to differential environmental sensitivity in a tested experimental context. Sequencing efforts have characterised these haplotypes, distinguished by a handful of distinct nucleotide variations. Haplotype-specific variations occur across the rDNA locus, with the most notable of these found to occur within both the transcribed spacer regions and coding sequencing of the mature rRNAs. The recent development of direct RNA sequencing by Oxford Nanopore Technologies (ONT), now presents the opportunity to directly sequence RNA molecules and explore the epitranscriptomic profiles of rRNA variants, and ascertain any differential RNA modifications which may gain them functional differences. This chapter intends to explore the differences in rRNA haplotype-specific modifications, as well as rRNA-related epitranscriptome changes in a developmental and cell type-specific context.

Specifically, the work in this chapter aims to:

1. Establish a working protocol for the sequencing of ribosomal RNA with ONT sequencing platforms
2. Develop a protocol for the capture of full-length rRNA primary transcript, or short-lived rRNA precursor processing intermediates
3. Explore the modification profiles of specific rRNA haplotypes, as well as differences in rRNA modification in a cell-specific and developmental context.

4.1.2 Detecting and mapping RNA modifications

Mapping RNA modifications transcriptome-wide is integral to unravelling their role in the dynamic cellular response. However, the historical inaccessibility of RNA modifications to molecular study

Long-Read Sequencing Analysis of Ribosomal RNA Modifications coupled with technical limitations has hindered these efforts. For many years, RNA modifications were largely detected using chromatographic methods and mass spectrometry by exploiting a modifications' distinct physicochemical property (Kellner *et al.*, 2014). These techniques proved well enough in detecting abundant modifications on abundantly present RNA species, however,

underperformed in detecting rare modifications on lowly expressed RNAs. Additionally, these traditional methods only allow for the examination of ensembles of molecules, from which the fraction of modified sites are estimated. For instance, the commonly utilized method of liquid chromatography-tandem mass spectrometry (LC-MS/MS) has the power to identify all types of nucleotide modifications, but only as a fraction of entire transcript populations (Taoka et al. 2018), all whilst providing minimal sequence context.

The technological advancements in next-generation sequencing (NGS) have served to expand the scope of RNA modification research, suddenly allowing for detection of rarely occurring modifications all whilst preserving RNA sequence context. Assays such as m₆A-Seq, based on the coupling of NGS and immunoprecipitation, allowed for the first time, transcriptome-wide analysis of m₆A sites, with pioneering work revealing that RNA modifications are much more widespread than previously thought (Dominissini *et al.*, 2012). In recent years, NGS-based methods have become the gold standard for RNA modification mapping with common approaches including methods relying on chemo-selective alterations (Chem-Seq) (Ramaswami *et al.*, 2013), RNA immunoprecipitation (RIP-Seq) (Dominissini *et al.*, 2012), or the detection of specific mismatch signatures (Mismatch-Seq) (**Figure 4.1**). Specifically, RIP-Seq involves the antibody selection and enrichment of specific RNA modifications. In Chem-Seq, RNA samples are pretreated with chemical reagents, which inhibit the reverse transcription reaction beyond the chemically modified position. Mismatch-Seq is based on the increased mismatch rates that occur upon reverse transcription at certain RNA-modified positions. In each case, these methods enrich for RNA harboring specific modifications, after which NGS methods can then be used to determine their sequence of origin. Such approaches have expanded our understanding of the role of RNA modifications in RNA stability (Boo and Kim, 2020), processing, localisation (Madugalle *et al.*, 2020), and translation (Mao *et al.*, 2019), as well as revealing them to be key regulators of a variety of biological process (Geula *et al.*, 2015; Jin *et al.*, 2019).

Even so, a major hindrance to rapid progress in NGS-based RNA modification research is the general lack of detection methods. Whilst over 150 naturally occurring RNA modifications have been described (Boccaletto *et al.*, 2018), only a handful can currently be detected, identified and quantified (Linder and Jaffrey, 2019; Anreiter *et al.*, 2021). NGS-methods rely on the conversion of RNA to complementary DNA (cDNA) which essentially strips the RNA of its modifications, necessitating the indirect detection of modified bases via antibody or chemical-based methods. Due to the limited repertoire of commercially available antibodies (Novoa, Mason and Mattick, 2017) and a lack of chemical compounds selective for specific modified ribonucleotides (Helm, Lyko and Motorin, 2019), ~90 % of known RNA modifications remain unmappable with NGS-based detection

Long-Read Sequencing Analysis of Ribosomal RNA Modifications methods. Additionally, the majority of current NGS based methods remain highly specific for a single type of modified nucleotide, for instance RiboMeth-seq (Birkedal *et al.*, 2015)) which can only detect ribose methylations, and Ψ -seq ('psi-seq') which can only detect sites of pseudouridylation (Schwartz *et al.*, 2014). Additionally, owing to NGS read size limitations, such methods cannot inform on the associations between modification at distant sites in large RNA molecules and do not allow for the capture of modification profiles across all positions of a full length transcript. Such whole molecule information is necessary to assess the interplay between distally located sites. Additionally, whilst providing sequence-specific information, these methods neglect RNA isoform-specific modification, are often not quantitative and do not provide accurate nucleotide level resolution (Meyer *et al.*, 2012). Moreover, NGS based methods often require complex multi-step protocols which can themselves introduce biases and artefacts in the data (Linder *et al.*, 2015; Linder and Jaffrey, 2019).

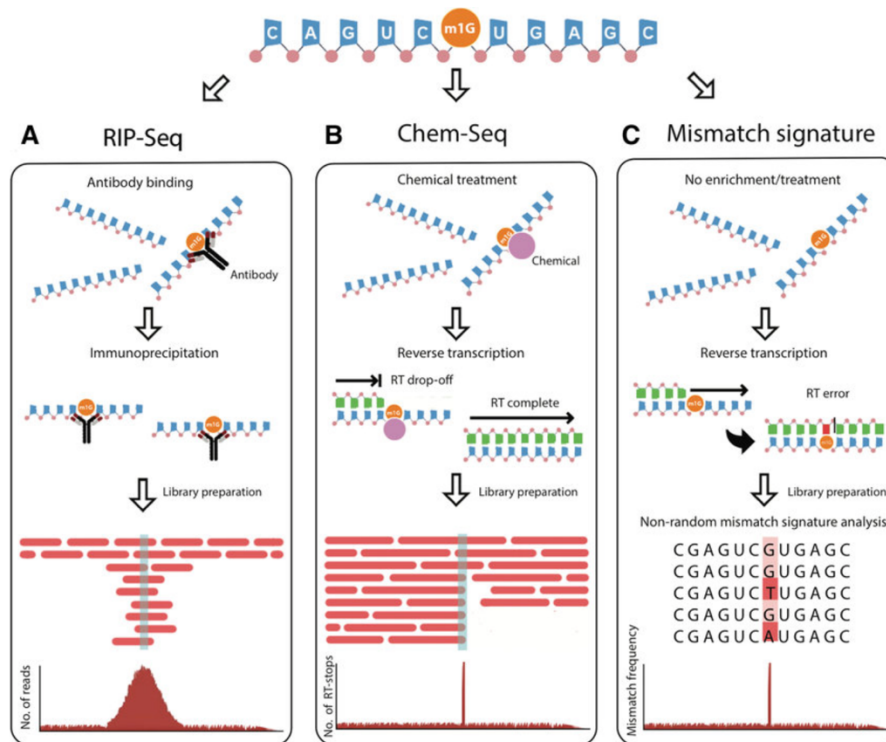


Figure 4.1 Current genome-wide detection methods used to identify RNA modifications. (A) Antibody-based method (RIP-seq) shows how RNA-modification enriched fragments are selected using pool-down, and compared to a total fragmented sample (input), which is used for normalization, obtaining genome-wide maps with peak resolution. (B) In Chem-Seq, RNA samples are pretreated with chemical reagents, which inhibit the reverse transcription reaction beyond the chemically modified position. (C) Mismatch signature-based methods are based on the increased mismatch rates that occur upon reverse transcription at certain RNA-modified positions.

(Jonkhout *et al.*, 2017)

4.1.3 Nanopore RNA sequencing

A promising alternative to both traditional and NGS-based methods is the direct sequencing of native RNA molecules to detect and quantify RNA modifications in a sequence specific manner. The development of Oxford Nanopore Technologies' direct RNA sequencing (DRS) does just this, allowing

Long-Read Sequencing Analysis of Ribosomal RNA Modifications for the direct sequencing of full-length native RNA without PCR amplification or cDNA conversion (Gralde *et al.*, 2018). As a result, Nanopore DRS preserves RNA modifications across the length of the transcript, permitting the direct and simultaneous detection of RNA modifications alongside nucleotide sequence. Additionally, unlike NGS-based methods, there is no size limitation imposed during library preparations permitting the study of the epitranscriptome across full-length transcripts and the exploration of distally located modifications at the single- molecule level.

Simply, Nanopore sequencing detects modified nucleotides according to differences in the current signals between modified and unmodified bases. During nanopore sequencing, at any one point in time, ~5 nucleotides (k-mer) occupy the sequencing pore, resulting in a distinct, k-mer-specific change in the current signal being detected. Modifications fundamentally change the physio-chemical properties of a base and alter how it interreacts with the sequencing nanopores. Crucially, they can cause discernible shifts in current intensity, as well as impact the time a nucleic acid sequence occupies a pore (dwell time). Current signals that deviate from the norm, or from what is expected from canonical bases can be used in conjunction with changes in dwell time to infer potential modification position (Anreiter *et al.*, 2021). Furthermore, as different modifications alter the current profile uniquely, distinct current signatures can be used to determine the exact identity of a base modification (Gralde *et al.*, 2018). Nonetheless, RNA modification detection from DRS signals is far from simple

and presents various challenges. The differences between current signals of individual modified and unmodified bases are often subtle and highly sequence-specific. Additionally, the variable translocation rate of nucleotide molecules through the sequencing pores, as well as the possible pore-to-pore variability, different copies of an identical molecule display considerable signal variations (Rang, Kloosterman and de Ridder, 2018). These challenges make the application of sophisticated computational models necessary to interpret the signals and modification status.

4.1.4 Mapping RNA modifications using Nanopore data

Nanopore sequencing has been successfully applied in the study of several commonly occurring RNA modifications including 2'-O-methylations, pseudouridine, N6-methyladenosine (m⁶A), 5-methylcytosine (m⁵C), 5-hydroxymethylcytosine (hm5C) and N7-methylguanosine (m⁷G) (Jonkhout *et al.*, 2017; Garalde *et al.*, 2018; Parker *et al.*, 2020; Jenjaroenpun *et al.*, 2021; Leger *et al.*, 2021). A substantial effort by the scientific community has also led to the development of several analytical tools to identify and map specific modifications (Furlan *et al.*, 2021). Modification detection tools for nanopore data can be primarily grouped into two types: those that detect modifications based on aberrations to the raw electrical signal, or those that rely on modification-induced base-calling errors. The first type includes tools like Eligos (Jenjaroenpun *et al.*, 2021), EpiNano (Liu, Begik and Novoa, 2021), and DiffErr (Parker *et al.*, 2020) which rely on base-calling errors introduced by RNA modifications. Considering the current improvements in nanopore base-calling algorithms, as they become increasingly less sensitive to common modifications, this increases the risk of false negatives and underrepresentation of modification sites. The second type includes tools such as Tombo (Stoiber *et al.*, 2016), and xPore (Pratanwanich *et al.*, 2021), which detect changes in the differential ionic current intensity between modified and unmodified positions. Although current fluctuation for some modifications can be too small to detect, the coupling of current fluctuations with other features such as k-mer 'dwell time' can allow for a richer comparative analysis. Modification tools can be further categorised as either comparative methods that infer modifications from differences between two samples, or *de novo* detection methods that utilise models trained for detecting specific RNA modifications. Currently, *de novo* methods remain encumbered by the difficulty to produce the extensive training sets containing all possible k-mer combinations with and without modifications. As a result, the majority of these tools are limited in their scope of detection and are primarily adept at detecting a limited set of specific RNA modifications. For instance, MINES (Lorenz *et al.*, 2020), Nanom6A (Gao *et al.*, 2021), and m6Anet (Hendra *et al.*, 2021) all predict m6A. Additionally, the vast majority of these tools do not provide information at single-molecule resolution, preventing the profiling of RNA modifications within specific transcript variants. To thoroughly examine the complete

modification profile of rRNA, as well as to elucidate any allele-specific differences we turn our attention to Nanocompore (Leger *et al.*, 2021).

4.1.5 Nanocompore

Nanocompore is an example of a model-free, comparative modification assessment tool which predicts modification sites based on differences in current signals. It is based on a 2 components Gaussian mixture model, in which an experimental or ‘test’ sample is compared directly against a control sample, in which there are substantially fewer or no modifications. Ideally, the control sample is RNA isolated from a cell line in which the expression of a specific RNA modification enzyme has been reduced or completely inhibited. Alternatively, for the study of specific transcripts, *in vitro* transcribed RNA may be used. Nanocompore is not restricted by the requirement of modification-specific models and so has the potential to detect any given modification across the length of a transcript, provided an appropriate control is used and the modification results in a significant alteration in the current signal. The tool has shown effectiveness in the detection of several different RNA modifications including m⁶A, Inosine, pseudouridine, m⁶₂A, m⁵C, m⁶G and 2’OmeA, and has been applied to inferring RNA modification at single-molecule resolution (Leger *et al.*, 2021). Due to its potential for detecting a wide range of modifications and allowing for single-molecule comparisons, Nanocompore may serve as a valuable tool in discerning differential RNA modification patterns of ribosomal RNAs in an allele-specific manner.

4.1.6 Predicting ribosomal RNA modification on full-length transcripts

The majority of eukaryotic rRNA modifications are sites of 2^O-O-methylation and pseudouridine, both of which are installed, largely by small nucleolar (sno)RNPs (Watkins and Bohnsack, 2012). Eukaryotic snoRNPs, are ribonucleoprotein complexes containing a snoRNA, which base-pairs with the rRNA and directs the catalytic protein component of the snoRNP to modify a target residue. Base modifications largely occur before rRNA maturation, during transcription in a large precursor particle composed of over 200 biogenesis factors, called the preribosome (Grandi *et al.*, 2002; Tschochner and Hurt, 2003). Processing and base modification in the preribosome, are both dependent on the spatial coordination of the 2D conformation adopted by the pre-rRNA molecule, which results from extensive sequence-specific interaction between transcribed spacer elements and coding subunit sequences (Dutca *et al.*, 2011; Zang *et al.*, 2016). Considering the dependence of this process on sequence-specific interactions, coordinated by the precise 2D confirmation of rRNA, it is of interest to explore the impact

of sequence variation on rRNA modification profiles and elucidate the specific modification profiles of unique rRNA alleles.

Several studies have exemplified the use of ONT DRS in the specific study of ribosomal RNA and demonstrated its capabilities in detecting, identifying and mapping a range of rRNA modifications. Using a comparative method based on shifts in the electrical signal, Stephenson *et al.* (2002) demonstrated the detection of 2'-O-methyl and pseudouridine sites in rRNAs from both yeast and bacteria (Stephenson *et al.*, 2022). Whilst a study by Smith *et al.*, (2019), demonstrated the detection of conserved 16S rRNA 7-methylguanosine and pseudouridine modifications in *E. coli*, from DRS datasets (Smith *et al.*, 2019). Along with this, the identification of a 7-methylguanosine modification conferring aminoglycoside resistance in certain pathological *E. coli* strains was obtained for full-length 16S rRNA (Smith *et al.*, 2019). Till now, ONT DRS-based studies have focused exclusively on elucidating modification profiles across mature transcripts (Smith *et al.*, 2019; Jain *et al.*, 2021; Stephenson *et al.*, 2022), without consideration of transcribed spacer elements. Mature transcripts which are in relative abundance to immature pre-rRNA, are highly conserved sequences which show little inter- and intra-individual genetic variation, therefore, hold little value for the study of rRNA allele dynamics. In C57BL/6J inbred mice, a large proportion of allele-defining SNVs are concentrated in transcribed spacer elements (Rodriguez-Algarra *et al.*, 2022). Considering the vast sequence variation in these regions, and potential contribution to rRNA heterogeneity, approaches which are limited to the assessment of mature transcripts greatly underestimate this inherent complexity. Therefore, the analysis of reads spanning across these regions and beyond into coding subunits is a prerequisite for the study of allele-specific RNA modifications.

The literature presented here provides support for the potential of nanopore DRS and complementary analytical tools, in elucidating rRNA allele-specific modifications at the single-molecule level. Here, the outcomes from sequencing rRNA with nanopore DRS will be discussed, along with the results of modification calling across cell-specific DRS data sets using Nanocompore.

4.2 Methods and Materials

Materials and methods section		Page number
2.1	Cell culture techniques	36
2.1.1	Mouse Embryonic fibroblasts (MEFs)	36
2.1.2	Mouse Embryonic Stem Cells (MESC)s	37
2.1.3	Human Lymphoblastoid Cell line (LCLs)	37
2.2	DNA and RNA techniques	37
2.2.1	Agilent bioanalyser	37
2.2.3	Assessing nucleic acid purity and concentration	38
2.2.4	Quantitative polymerase chain reaction	38
2.7	Mouse Embryoid Body formation	50
2.8	Ribosomal RNA processing inhibition	52
2.9	Oxford Nanopore Sequencing methods	54

4.3 Results

4.3.1 Nanopore cDNA sequencing of *in vitro* poly-adenylated ribosomal RNA

A key prerequisite for ONT RNA sequencing is the presence of a poly(A) tail (minimum length of 8 nucleotides) at the 3' end of the RNA molecule. This is needed so that a poly(T) adapter can be bound to the molecule, and the sequencing motor protein can be subsequently fixed to the molecule. Only then is the RNA molecule threaded through the nanopore allowing for sequencing information to be gathered. Unfortunately, the majority of rRNA molecules are not endogenously synthesised with a 3'-poly(A) tail, necessitating its artificial addition. The *in vitro* poly(A) tailing of *in vitro* transcribed RNA is routine and methodically simple. However, considering the sizable costs of nanopore sequencing experiments, it was important to first confirm and assess poly(A) tail addition to rRNA molecules before proceeding with sequencing.

Total RNA extracted from MEFs was subjected to *in vitro* poly(A) tailing using an NEB poly(A) tailing kit. According to the associated protocol, 1 µg of RNA input yields a poly(A) tail measuring on average ~150 bp per molecule. To confirm the addition and quantify the size of the poly(A) tail added to rRNA molecules, two nucleotide size quantification methods were employed: RNA gel electrophoresis and Agilent RNA bioanalyser. The rationale was that upon successful poly(A)-tailing, a noticeable size shift in 18S and 28S bands would be observed between test and control samples. Unfortunately, due to the low sensitivity of both methods to detect the relatively small shift in size expected after 150 bp poly(A) tail addition (1869 to 2019 bp for 18S, and 4729 to 4979 for 28S), no noticeable difference in size was detected.

Instead, poly(A) treated RNA was subjected to oligo dT-bead purification to assess poly(A)-tailed RNA enrichment. Here, a magnetic bead linked to a poly(T) oligo complementarily binds to the poly(A) tail of an RNA molecule, extracting it from a heterogeneous sample. Considering that rRNA constitutes 80-95% of total RNA and is not endogenously poly-adenylated, it was reasoned that the same amount of RNA input would result in significantly different yields when comparing total RNA (control) to poly(A) treated total RNA (test). According to the product protocol, 1 mg of oligo dT beads can isolate up to 1 µg of poly(A) mRNA from 5 µg of total RNA input. Here, a 5-fold increase in poly(A) enrichment was observed between control and test samples, with 1 mg of oligo d(T)beads yielding ~100 ng of poly(A) RNA from 5 µg of total RNA, and ~500 ng of poly(A) RNA from 5 µg of *in vitro* poly adenylated total RNA. Though the size and extent of rRNA poly(A) tailing remain unquantified, the considerable difference in poly(A) RNA enrichment indicated some level of *in vitro* poly(A) tailing success.

Long-Read Sequencing Analysis of Ribosomal RNA Modifications

Next, to evaluate the result of *in vitro* poly(A) tailing on Nanopore sequencing of rRNA, a series of pilot sequencing experiments were conducted. For the pilot study, sequencing was conducted on the Flongle sequencing device, owing to the considerably lower cost of sequencing flow cells compared to all other ONT platforms. However, due to the incompatibility of Flongle flow cells with direct RNA sequencing, cDNA sequencing was selected as an appropriate alternative for protocol optimisation. **Figure 4.2** illustrates the strategy used to prepare rRNA for Nanopore sequencing. Three separate runs were conducted (noted pilot 1, 2, and 3 respectively), using *in vitro* poly-adenylated C57BL/6J muscle total RNA. Common to all runs, 1 µg of DNase I treated total RNA was *in vitro* poly(A) tailed as described previously. For pilot 1, 100 ng of poly(A) RNA was used as input for cDNA library preparation (as recommended for ONT cDNA sequencing kit SQKDCS109), yielding 20 ng of cDNA library. This falls below the ONT recommended cDNA input (60 ng). For pilot 2, poly(A) RNA input was doubled to 200 ng, yielding ~40 ng of cDNA library. For pilot 3, RNA input was further increased to 400 ng, with poly(A) RNA being subjected to an additional oligo (dT) bead isolation, yielding ~100 ng of enriched poly(A) RNA (poly(A)+), which subsequently yielded ~60 ng of cDNA library.

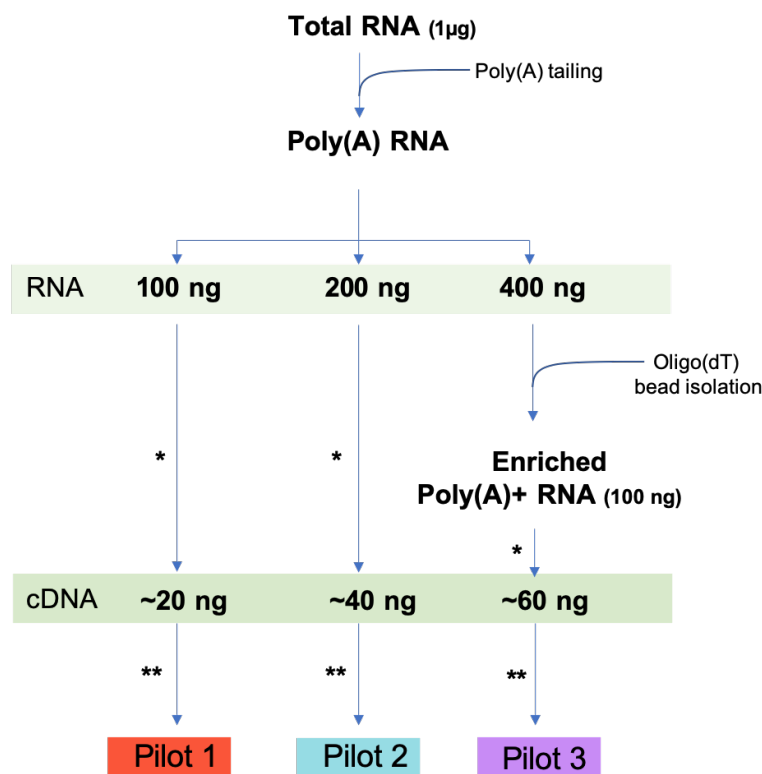


Figure 4.2 ONT cDNA Sequencing pilot study sample preparation. 3 ONT cDNA sequencing pilot runs were conducted, with variations in sample preparation. In each instance, 1 µg of DNase I treated total RNA, extracted from C57BL/6J muscle tissue, was subjected to poly(A) tail addition. For 'Pilot 1', 100 ng of poly(A) RNA (ONT recommended RNA input) was taken forward for SQKDCS109 library preparation, yielding ~20 ng of cDNA library. For 'Pilot 2', 200 ng of poly(A) RNA was taken forward for SQKDCS109 library preparation, yielding ~40 ng of cDNA library. For 'Pilot 3', 400 ng of poly(A) RNA was subjected to oligo (dT) bead isolation, yielding ~100 ng of enriched poly(A)+ RNA, which was taken forward for SQKDCS109 library preparation, yielding ~60 ng of cDNA library.

Each cDNA library was run on a separate ONT Flongle flow cell as per ONT instructions.

*Library preparation using ONT cDNA-Seq kit SQKDCS109

**Taken forward for sequencing using ONT Flongle flow cell.

Each library was run independently on a single, unused Flongle flow cell, quality control checked to have a minimum of 60 pores available for sequencing. Reads were aligned to the mouse whole-genome (WG) assembly GRCm38 (mm10)+ rDNA reference (Accession No. BK000964.3). **Figure 4.3A** presents sequencing summary statistics for pilots 1, 2, and 3. An increase in total read output and WG mapping was observed with each sequential run, positively correlating with increasing cDNA library input. Alignment to rDNA revealed an overall alignment rate of > 73% (1), 72% (2) and 65% (3). For rDNA-aligned reads, the mean read length was 973 (1), 999 (2), and 832 bp (3) with the longest read captured measuring 4601 (1), 5649 (2) and 6941 bp (3) respectively. Coverage depth across the rDNA coding unit (1-13,403 bp) is presented in **Figure 4.3B**, with coverage presented specifically for both transcribed spacer elements and rRNA coding subunits, 18S, 5.8S, and 28S. Coverage depth profiles were comparable between the three pilot runs, with coverage of 18s, 5.8S and 28S coding subunits observed for all three. A distinct 3' coverage bias is observed across the 18S and 28S subunits, for all three sequencing runs. Coverage depth appeared to gradually increase (5' to 3') In the case of the 18S, whilst a sudden increase in 3' coverage was seen at around 12,000 bp for the 28S. The coverage of the 5.8S subunit was comparatively reduced to both the 18S and 28S, likely due to the smaller size of the 5.8S mature transcript. Also, coverage of coding subunits, in particular, that of the 18S and 28S, was comparably higher than that of transcribed spacer elements, likely owing to the capture of mature rRNA transcripts.

A

		Reads Aligning To:		rDNA Alignment Rate (%)	rDNA Aligned Reads		
					Mean Read Length (bp)	Longest Read (bp)	
Pilot		cDNA Input (ng)	Whole Genome	rDNA			
	1	~20	20,329	14,965	73.6	973	4601
	2	~40	55,634	40,507	72.8	999	5649
	3	~60	109,391	71,850	65.7	832	6941

B

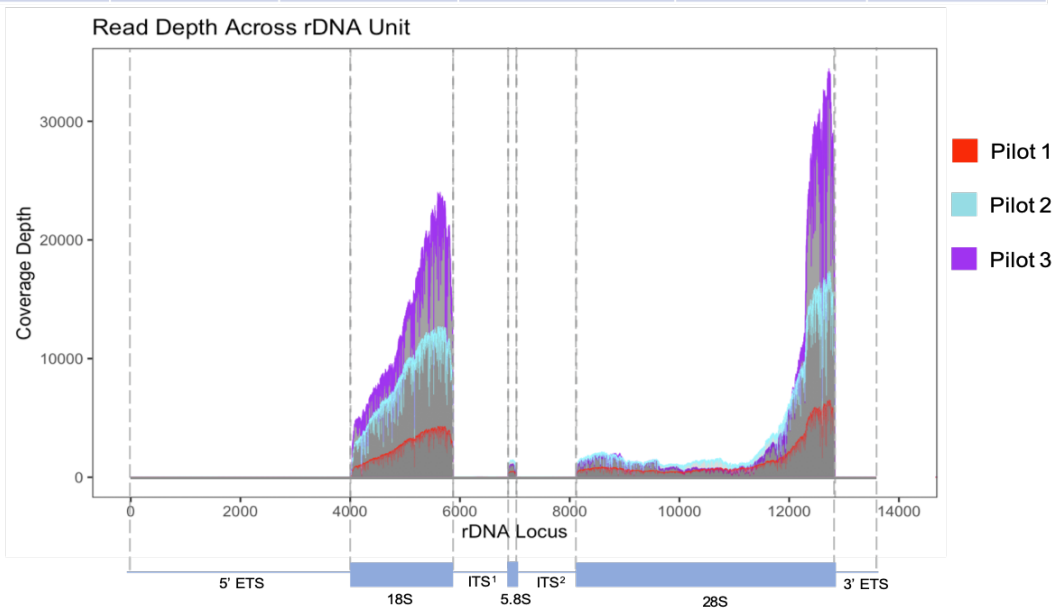


Figure 4.3 Comparison of ONT cDNA sequencing pilot runs. (A) Read coverage depth across rDNA unit presented for pilot 1 (red), 2 (blue) and 3 (purple). Coverage depth plot is aligned to a schematic of rDNA coding unit (1-13403 bp), Grey dotted lines indicate boundaries of rDNA coding unit elements, (left to right- 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS). **(B)** Summary statistics of pilot runs 1,2 and 3. Each cDNA library was run on a separate ONT Flongle flow cell as per ONT instructions. All rRNA reads are mapped to the published consensus sequence (Accession No. BK000964.3).

To assess rRNA haplotype-specific modification profiles, it was necessary to capture reads spanning across haplotype-specific SNP's (largely occurring in transcribed spacer elements) and into the coding subunits. Reads were specifically assessed for the presence of pre-rRNA (precursor rRNA transcript processing intermediates spanning across both TS and coding sub-unit elements). To this end, reads from all three pilot runs were pooled, providing 162,206 rDNA mapping reads for assessment. First, all individual reads were aligned to the rDNA coding unit (**Figure 4.4A**). From this, it emerged that there were a proportion of reads aligning to the transcribed spacer (TS) elements. To ensure these reads were not simply the by-product of rRNA precursor processing, all reads only mapping to the TS elements were disregarded. Additionally, only pre-rRNA reads, defined as those aligned to both the rRNA coding subunits and TS elements (>5 bp into each) were selected (332 reads) (**Figure 4.4B**). Next,

to select reads which would potentially span across multiple coding subunits, reads were selected for the presence of intact ITS¹ and ITS² cleavage sites. To account for the approximate positions of cleavage sites at positions ~5932, ~6712, (ITS¹) and ~7841 (ITS²), only reads spanning ± 5 bp of the indicated positions were chosen (110 reads) (**Figure 4.4C**). However, due to the multiple rRNA precursor processing pathways documented, reads were finally selected for complete coverage of ITS¹ or ITS² (+1 bp on both 5' and 3' ends), yielding 6 reads from a total of 162,206 (**Figure 4.4D**). Of these, no reads completely spanned ITS¹, the element in which the majority of identified haplotype-specific SNPs occur.

Overall, based on the outcomes of these pilot sequencing runs, *in vitro* poly(A) tailing of total RNA allowed for sequencing of rRNA with Nanopore technology, with an rDNA-alignment rate consistently >65%. Additionally, sample enrichment with oligo (dT) bead selection of poly(A)+ RNA (pilot 3) resulted in increased cDNA library yield and a subsequent increase in sequencing output. Therefore, oligo (dT) bead-based enrichment was introduced as a fundamental step in all subsequent sample preparations. Even so, the assessment of pre-rRNA capture revealed that input of total cellular RNA (extracted from preserved C57BL/6J muscle tissue), proved ineffective for capturing sufficient amounts of pre-rRNA from which haplotype-specific RNA modifications profiles could be determined. Achieving this may likely require optimising sample pre-processing for the enrichment of pre-rRNA as well as ensuring the preservation of RNA integrity during extraction and processing.

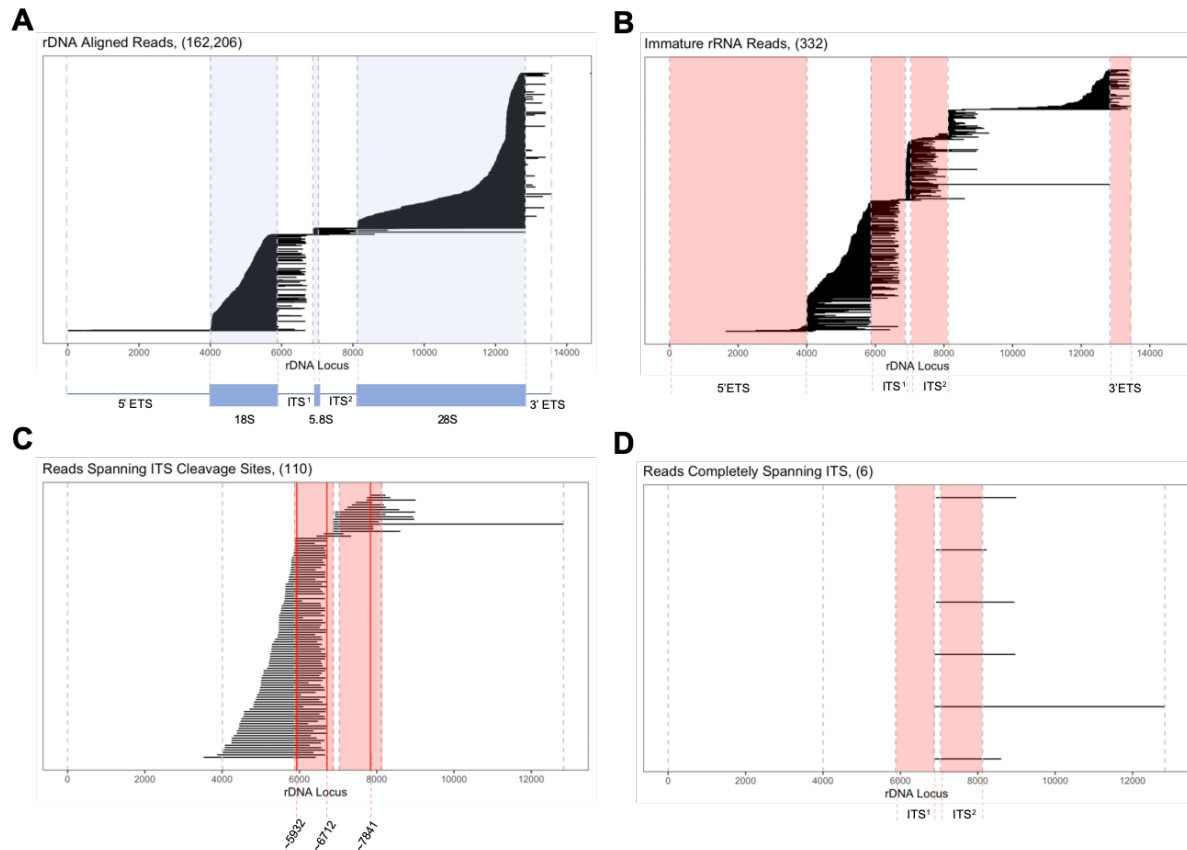


Figure 4.4 Nanopore cDNA sequencing of C57BL/6J muscle rRNA (A) A stacked plot of all individual reads mapping to rDNA (162,206 reads). Vertical dotted lines indicate boundaries of rDNA coding unit elements (left to right- 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS). (B) Stacked plot of all immature rRNA reads (332 reads). Reads are designated as 'Immature' if they map to rRNA genes (18S, 5.8S, 28S) and > 5 bp into transcribed spacer elements (red). (C) Stacked plot of all reads spanning cleavage sites in ITS¹ (~5932 bp, ~6712 bp), and ITS² (~7841bp), (110 reads). Reads are classified as such if they spanning ITS elements are highlighted in red (left to right- ITS¹- ITS²) with cleavage sites indicated by vertical red lines, labelled with approximate positions. (D) Stacked plot of all reads completely spanning ITS regions (6 reads). Reads are classified as such if they map across the entirety of ITS¹ or ITS², +1 bp beyond, on both 5' and 3' ends.

This figure presents sequencing data combined from sequencing runs, pilot 1, 2 and 3.

All rRNA reads are mapped to the published consensus sequence (Accession No. BK000964.3).

4.3.2 Nuclear RNA extracts permit increased capture of rRNA processing intermediates compared to cellular extracts

To increase the capture of pre-rRNA, sample pre-processing was refined to enrich for pre-rRNA processing intermediates. To begin, the focus was shifted from the finite and precious stock of preserved tissue to abundantly available cultured cells. C57BL/6J MEFs were selected as an appropriate cell line, owing to the ease with which they could be cultured and manipulated, as well as the extensive prior genetic characterisation. To begin protocol refinement, a baseline coverage of pre-rRNA was determined by ONT cDNA sequencing of MEF total RNA, with reads assessed for the presence of pre-rRNA. Specifically, ITS¹ and ITS² spanning reads were used as indicators of this. MinION cDNA sequencing of MEF total RNA yielded 79,949 rDNA aligned reads with a mean length of 841 bp. The coverage depth across the rDNA unit is plotted in **Figure 4.5A**, with all individual reads aligning to the rDNA coding unit presented in **Figure 4.5B**. Assessment of pre-rRNA reads yielded 2 reads completely spanning ITS elements (**Figure 4.5C**). Of these, 1 read completely spanned ITS¹ (0.00125 % of total rRNA reads), whilst 2 reads completely spanned ITS² (0.0025 % of total rRNA reads). Sequencing summary statistics are provided in **Figure 4.4D**.

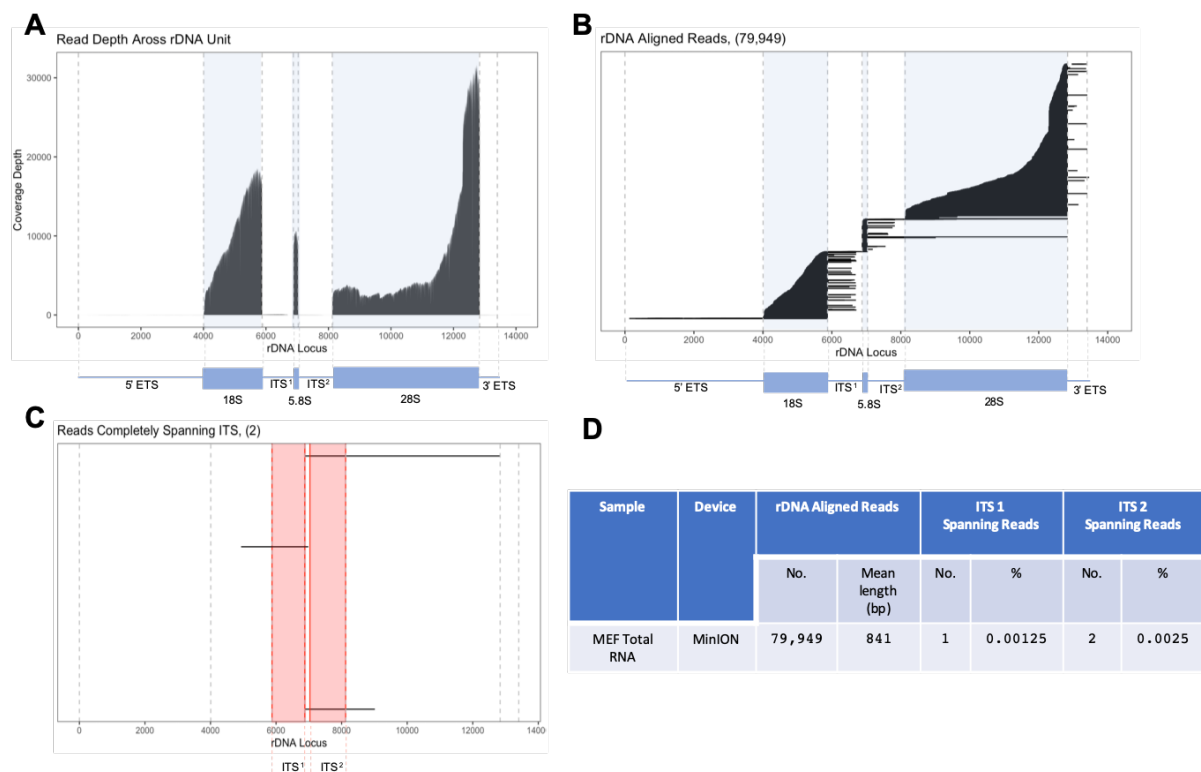


Figure 4.5 Nanopore cDNA sequencing of MEF total RNA, read analysis. (A) Read coverage depth across rDNA unit aligned to a schematic of rDNA coding unit (1-13403 bp), Grey dotted lines indicate boundaries of rDNA coding unit elements, (left to right- 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS) (B) A stacked plot of all individual reads mapping to rDNA (79,949 reads). Plot is aligned to a schematic of rDNA coding unit (1-13403 bp) Vertical dotted lines indicate boundaries of rDNA coding unit elements, (left to right: 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS). (C) Stacked plot of all reads completely spanning ITS regions (2 reads). Reads are classified as such if they map across the entirety of ITS¹ or ITS², +1 bp beyond, on both 5' and 3' ends. (D) cDNA sequencing summary statistics

All rRNA reads are mapped to the published consensus sequence (Accession No. BK000964.3).

Considering this outcome, it was hypothesised that a sub-cellular fraction of nuclear RNA may lead to increased capture of rRNA processing intermediates when compared to total cellular RNA. Though the majority of rRNA precursor processing occurs within the nucleolus, mature transcripts are readily exported out, into the cytoplasm. By removing this sizeable pool of mature cytoplasmic rRNA and only selecting for nuclear RNAs, enrichment of immature pre-rRNA transcripts could be achieved. To this end, RNA was extracted from isolated MEF nuclei and prepared for ONT cDNA-Seq. MinION cDNA sequencing of MEF nuclear RNA yielded a total of 249,086 rDNA aligned reads with a mean length of 874 bp. The coverage depth across the rDNA unit is plotted in **Figure 4.6A**, with all individual reads aligning to the rDNA coding unit presented in **Figure 4.6B**. Assessment of pre-rRNA reads yielded 18 reads completely spanning ITS elements (**Figure 4.6C**). Of these, 6 reads completely spanned ITS¹ (0.0024 % of total rRNA reads), and 14 completely spanned ITS² (0.0056 % of total rRNA reads). Additionally, 2 reads across both ITS¹ and ITS² (**Figure 4.6D**). Sequencing summary statistics are provided in **Figure 4.6E**. Compared to total cellular RNA, ONT cDNA sequencing of nuclear RNA resulted in a greater degree of complete transcribed spacer element capture with a near ~2-fold increase in reads mapping completely across ITS¹, and a similar increase in complete coverage of ITS² was observed. Though a slight increase in read length was also observed, a large proportion of the read output for both sequencing runs was made up of short reads measuring below 500 bp. This falls below the read size required to span across the entirety of ITS¹ (1,001 bp) or ITS² (1,089 bp), indicating further refinement of the sample pre-processing protocol was needed to increase read length.

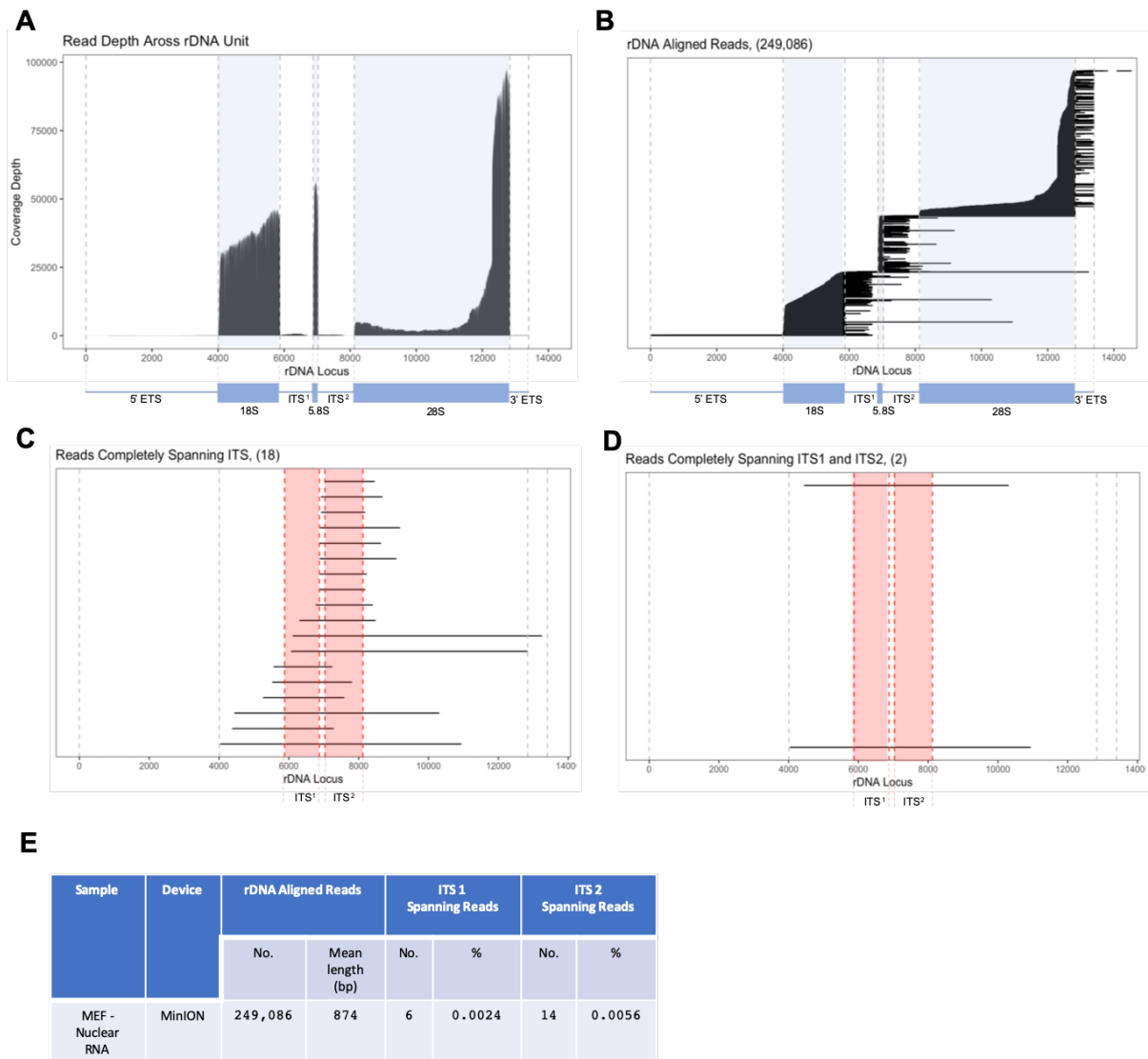


Figure 4.6 Nanopore cDNA sequencing of MEF nuclear RNA, read analysis. (A) Read coverage depth across rDNA unit aligned to a schematic of rDNA coding unit (1-13403 bp). Grey dotted lines indicate boundaries of rDNA coding unit elements, (left to right: 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS). (B) A stacked plot of all individual reads mapping to rDNA (249,086 reads). Plot is aligned to a schematic of rDNA coding unit (1-13403 bp). Vertical dotted lines indicate boundaries of rDNA coding unit elements, (left to right: 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS). (C) Stacked plot of all reads completely spanning ITS regions (18 reads). Reads are classified as such if they map across the entirety of ITS¹ or ITS², +1 bp beyond, on both 5' and 3' ends. (D) Stacked plot of all reads completely spanning ITS¹ and ITS² (2 reads). Reads are classified as such if they map across the entirety of ITS¹ and ITS², +1 bp beyond, on both 5' and 3' ends. (E) cDNA sequencing summary statistics. All rRNA reads are mapped to the published consensus sequence (Accession No. BK000964.3).

4.3.3 Size selection-based enrichment of rRNA processing intermediates

To increase the size of captured reads, a size selection step was introduced to the sample pre-processing protocol. It was hypothesised that the removal of smaller RNA fragments would enrich larger molecules, increasing the potential for capturing pre-rRNA processing intermediates. Size selection was achieved with the use of Solid Phase Reversible Immobilisation (SPRI) beads, commonly used to purify nucleic acids as well as for the size selection of both RNA and DNA short-read

Long-Read Sequencing Analysis of Ribosomal RNA Modifications

sequencing libraries. SPRI beads are paramagnetic beads coated with carboxyl molecules, resuspended in a solution of crowding reagent polyethylene glycol (PEG) and salt. The PEG causes the negatively charged nucleic acid to bind the carboxyl group on the bead surface, with the concentration of PEG determining the degree of nucleic acid immobilisation as well as the size of DNA/RNA molecules that bind. This, in turn, makes the bead-to-sample-volume ratio used in the separation critical for size selection. Generally, lowering the PEG concentration leads to the binding and capture of larger nucleic acids. MEF nuclear RNA was size selected with SPRI beads, with a bead-to-sample-volume ratio of 0.35:1. Input nuclear RNA was size selected to obtain two distinct fractions, a supernatant fraction (unbound RNA) and an eluted fraction (RNA bound to beads). The input nuclear RNA and two size-selected fractions were used to generate cDNA, which was then used to assess the degree of size selection and quantify molecule size. To do this, cDNA prepared from each fraction was run on an Agilent TapeStation 2200 genomic tape (sizing capacity of 0.2-60 kb). Results are displayed in **Figure 4.7**. Size assessment of input nuclear cDNA revealed two major peaks, at ~1933 bp and ~5625 bp. Assessment of the supernatant fraction revealed a single major peak at ~1484 bp whilst the elute fraction was shown to be composed of larger molecules with two major peaks at ~2026 bp and ~5570 bp.

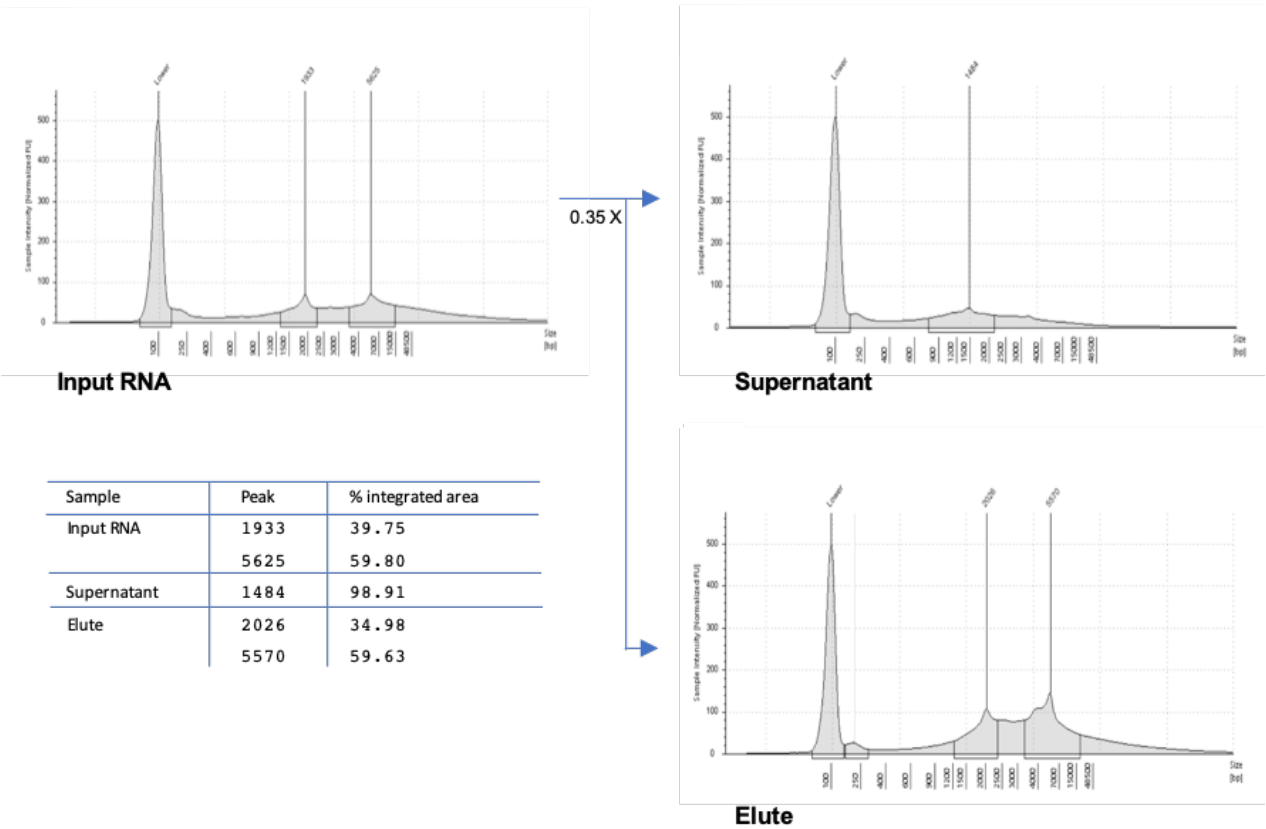


Figure 4.7 Size analysis of size-selected MEF nuclear cDNA using TapeStation 2200. cDNA generated from nuclear RNA (input) was size selected with SPRIselect beads using a 0.35:1 bead to sample volume ratio to yield output fractions, supernatant and elute. Electropherogram traces presented for input sample and output fractions, with X-axis and Y-axis representing the fragment size (bp) and sample intensity (normalised FU). Two major peaks are observed for input sample measuring at ~1,933 and ~5,625 bp. Supernatant size analysis yields a single major peak at ~1,484 bp whilst size selected elute yields two major peaks at ~2,026 and ~5,570 bp. 20 ng of each sample was analysed on Genomic DNA screen tape (0.2 - 60 kbp).

Long-Read Sequencing Analysis of Ribosomal RNA Modifications

Considering these results, MEF nuclear RNA was size-selected using a bead-to-sample volume ratio of 0.35:1. The size-selected elute was then processed and used for ONT cDNA sequencing library preparation. MinION cDNA sequencing of size-selected MEF nuclear RNA yielded a total of 23,895 rDNA aligned reads with a mean length of 962 bp. The coverage depth across the rDNA unit is plotted in **Figure 4.8A**, with all individual reads aligning to the rDNA coding unit presented in **Figure 4.8B**. Assessment of pre-rRNA reads yielded 18 reads completely spanning ITS elements (**Figure 4.8C**). Of these, 3 reads completely spanned ITS¹ (0.0126 % of total rRNA reads), and 15 completely spanned ITS² (0.0628 % of total rRNA reads). Sequencing summary statistics are presented in **Figure 4.8D**.

Compared with the MEF nuclear RNA sequencing run, only a modest 88 bp increase in mean read length was observed. However, a comparison of ITS coverage revealed a > 5-fold increase in reads completely spanning ITS¹ and a > 11-fold increase in reads completely spanning ITS². This was taken as an indicator of improved pre-rRNA enrichment. All subsequent libraries were thus prepared with an additional size selection pre-processing step as outlined above.

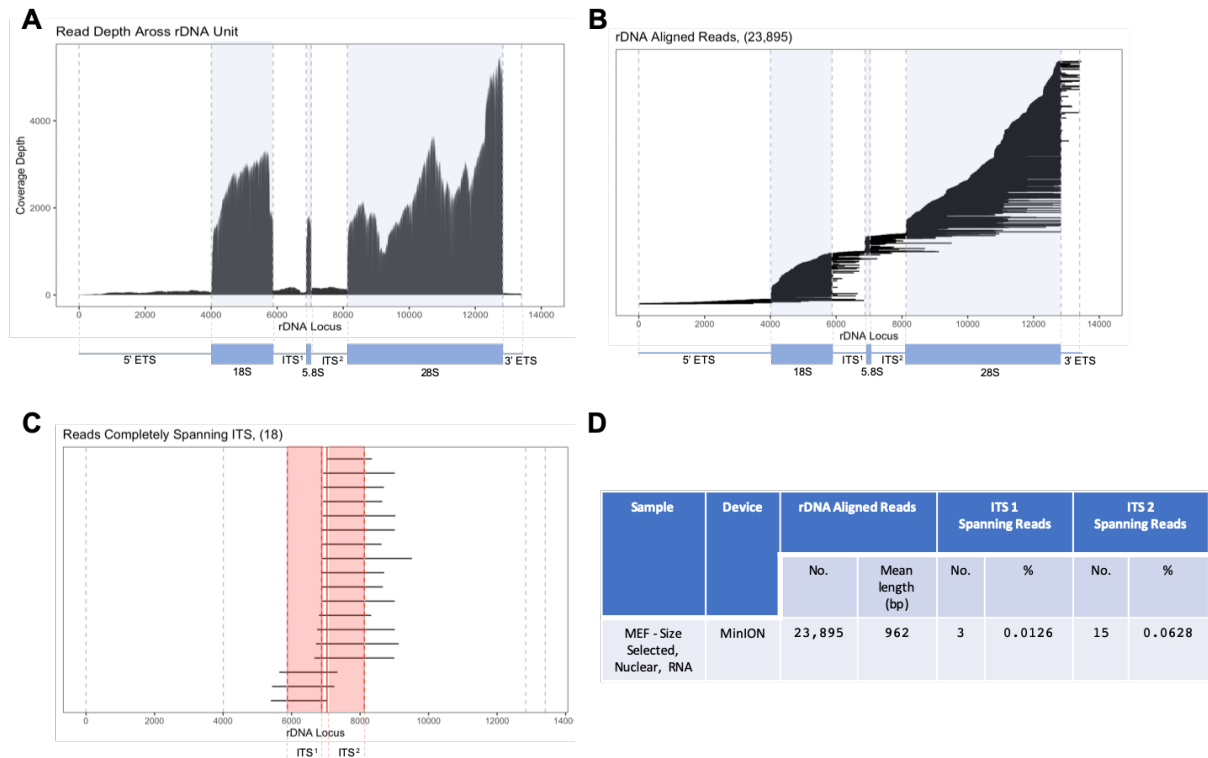


Figure 4.8 Nanopore cDNA sequencing of MEF, size selected, nuclear RNA. (A) Read coverage depth across rDNA unit is aligned to a schematic of rDNA coding unit (1-13403 bp). Grey dotted lines indicate boundaries of rDNA coding unit elements, (left to right: 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS). (B) A stacked plot of all individual reads mapping to rDNA (23,895 reads). Plot is aligned to a schematic of rDNA coding unit (1-13403 bp). Vertical dotted lines indicate boundaries of rDNA coding unit elements (left to right- 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS). (C) Stacked plot of all reads completely spanning ITS regions (18 reads). Reads are classified as such if they map across the entirety of ITS¹ or ITS², +1 bp beyond, on both 5' and 3' ends. (D) cDNA sequencing summary statistics
All rRNA reads are mapped to the published consensus sequence (Accession No. BK000964.3).

4.3.4 5-Fluorouracil exposure of MEF cells hinders rRNA processing

To further maximise the capture of pre-rRNA transcript processing intermediates, the cellular processes involved in rRNA processing were targeted. Several widely researched chemotherapeutic drugs are thought to exert their therapeutic effects through perturbing ribosome biogenesis and ultimately hindering cell cycle progression. Some of these are shown to directly impact pre-rRNA processing, preventing the maturation of rRNAs and their subsequent incorporation into ribosomes. Two such compounds are flavopiridol, a kinase inhibitor, and 5-fluorouracil (5-FU), an antimetabolic nucleotide analogue. Studies in human cell lines have demonstrated the disruptive impact of flavopiridol (0.049-0.781 μM) on the early processing of pre-rRNA, inhibiting the generation of the 32S pre-rRNA processing intermediate, without significantly inhibiting the generation of the 47S rRNA primary transcript (Burger et al., 2010). Similarly, 5-FU (6.25-100 μM) is shown to disrupt the late processing of pre-rRNA, inhibiting the generation of mature 18S and 28S rRNA without significantly inhibiting the generation of the 47S rRNA primary transcript (Burger et al., 2010).

To assess the impact of these drugs on rRNA processing in MEFs, cells were independently exposed to increasing concentrations of each drug and cell cycle progression was assessed as a general indicator of drug effect. Cell cycle progression was assessed for untreated and treated cells, via Fluorescence Activated Cell Sorting (FACS) of Propidium Iodide (PI)-stained fixed cells. Propidium iodide binds stoichiometrically to DNA, allowing for cells in different cell cycle stages to be distinguished based on the differing PI-DNA content. **Figure 4.9A** presents an example population histogram depicting the fraction of cells in each cell cycle phase (Sub-G1, G0/G1, S, G2/M), deduced by measuring the PI-DNA content of each cell within a population. **Figure 4.9B** presents the cell cycle progression analyses of MEF cells exposed to increasing concentrations of 5-FU. The percentage of cells in each cell cycle stage for each condition was quantified from 3 biological replicates and is presented in **Figure 4.9C**.

These cell cycle analyses of 5-FU treated MEF cells revealed perturbations to cell cycle progression with increasing 5-FU exposure, from 25-100 μM . Considering cells exposed to 25-50 μM 5-FU, a distinct build-up of cells in the G0-G1 and S phases was observed, alongside the clear loss of cells in the G2/M phase. Quantifying this, a ~5% and ~6% average increase in G0-G1 and S phase cells, alongside a ~11 % average decrease in G2/M cells was recorded for 50 μM 5-FU treated cells compared to control (0 μM). The greatest change in cell cycle progression was observed for 100 μM 5-FU exposed cells with a dramatic change to the population profile, characterised by a ~20 % increase in S phase cells and loss of a distinct G2/M peak. Interestingly, increasing 5-FU exposure to 200 μM appeared to exert a reduced effect on cell cycle progression compared to lower concentrations, with a reduction

of cells in the G0/G1 and S phases and a resurgence of cells in the G2/M phase. This was observed alongside a build-up of cells in Sub-G1, a potential indicator of cell death and drug toxicity. It is important to note that PI-DNA content-based analysis of cell cycle progression is limited, and can only provide approximates of cell population statistics. Furthermore, dramatic changes to a cell population profile, as observed with 100 μ M 5-FU exposure, can make the accurate differentiation of cells in different cell cycle phases difficult.

The cell cycle progression of MEF cells exposed to increasing concentrations of Flavopiridol (0.025, 0.05, 0.1, 0.2 μ M) was also assessed. However, no notable change was observed. Additionally, MESC and human LCLs were also similarly assessed for their response to 5-FU and Flavopiridol exposure. In all cases, however, no noticeable change in cell cycle progression was observed for the drug concentrations tested.

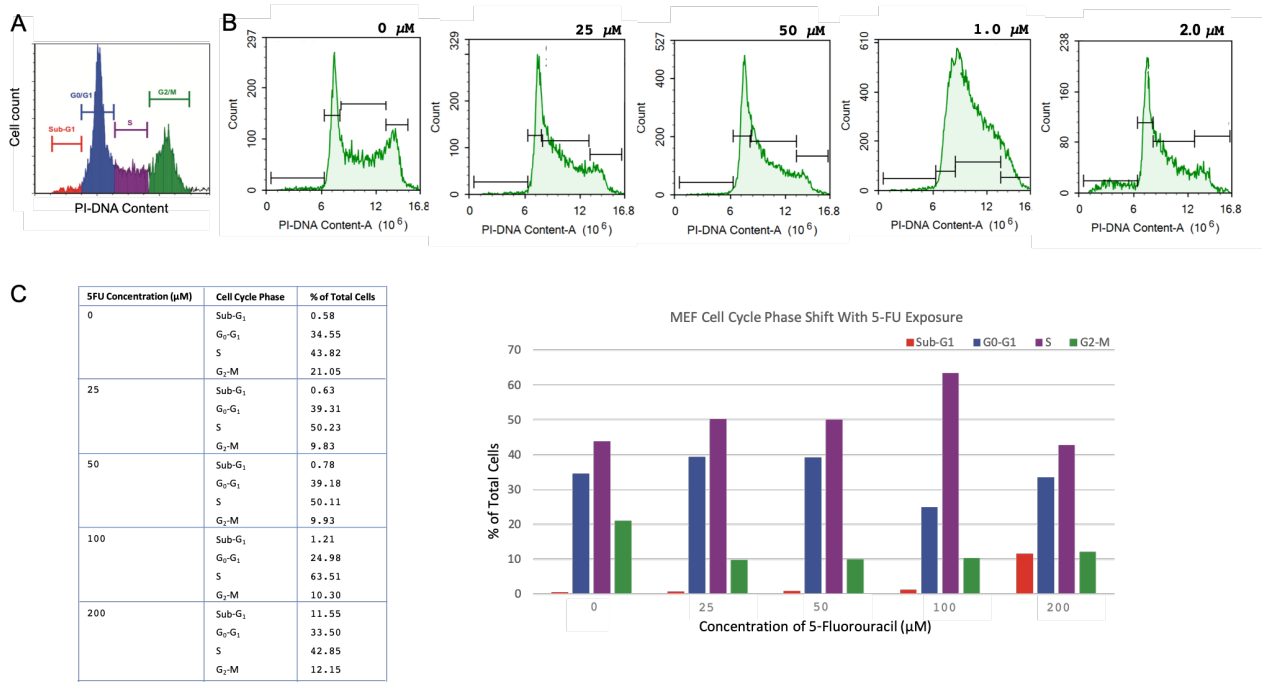


Figure 4.9 Cell cycle analysis of 5-FU treated MEF cells. (A) An example population histogram of a cell population in various stages of the cell cycle, Sub-G1 (red), G0/G1 (blue), S (purple), G2/M (green) with X-axis and Y-axis representing PI-DNA content and cell count. **(B)** MEF cell cycle progression inhibition with 5-FU treatment. Representative population histograms presented for 5 experimental conditions (5-FU concentration: 0, 0.25, 0.5, 1.0, 2.0 μ M), black bars indicate the fraction of cell population in various cell cycle phases (left to right: Sub-G1, G0/G1, S, G2/M). **(C)** Cell population cell cycle phase quantification. Percent of total cells in each cell cycle phase quantified from an average of 3 biological replicates and plotted for increasing 5-FU concentration.

Whilst the analyses from 5-FU exposure on cell cycle arrest indicated the impact of drug treatment on MEF cell cycle progression, qPCR was employed to assess and quantify 5-FU-induced inhibition of pre-rRNA processing. It was hypothesised that 5-FU induced inhibition of pre-rRNA processing would prevent the natural cleavage events involved in generating certain pre-rRNA processing intermediates. This in turn would lead to the build-up of intact ITS cleavage sites, with sequence levels quantifiably different between test and control groups. Three regions within the 47S primary transcript were targeted: 1) a site within the first 650 bp of the 5'ETS, used an indicator of total 47S levels, 2) a 110 bp sequence spanning positions 5893-6003, encompassing an ITS¹ cleavage site at position ~5932, and 3) a 118 bp sequence spanning positions 7760-7878, encompassing an ITS² cleavage site at position ~7841 (**Figure 4.10A**). To this end, MEF cells were exposed to increasing concentrations of 5-FU (25, 50, 100 μ M), total RNA was extracted, depleted of genomic DNA and processed to generate cDNA used for qPCR analyses. **Figure 4.10B** presents the detected levels of targeted sites 1,2 and 3, in cells exposed to increasing 5-FU concentrations, relative to control. Data are presented as the average fold change in detection of 2 biological replicates, each conducted as technical triplicates. A significant increase in the detection of sequences spanning intact ITS¹ and ITS² cleavage sites was observed across all concentrations of 5-FU tested when compared to levels in untreated control cells ($p < 0.01$). The greatest increase was observed for 25 μ M 5-FU exposure, resulting in a ~4.1- and ~5.8-fold increase in ITS¹ and ITS² target sequences compared with untreated control. In comparison, exposure to 50 μ M 5-FU resulted in a ~3.4- and ~4.2-fold increase in ITS¹ and ITS² target sequences, whilst 100 μ M 5-FU exposure resulted in a ~4.1- and ~3.5-fold increase in ITS¹ and ITS² target sequences. The detected levels of the 5'ETS sequence decreased with increasing 5-FU exposure. Compared with control, an initial increase was observed in cells exposed to 25 μ M 5-FU, with comparable levels for 50 μ M 5-FU exposed cells, whilst a decrease was observed in cells exposed to 100 μ M 5-FU.

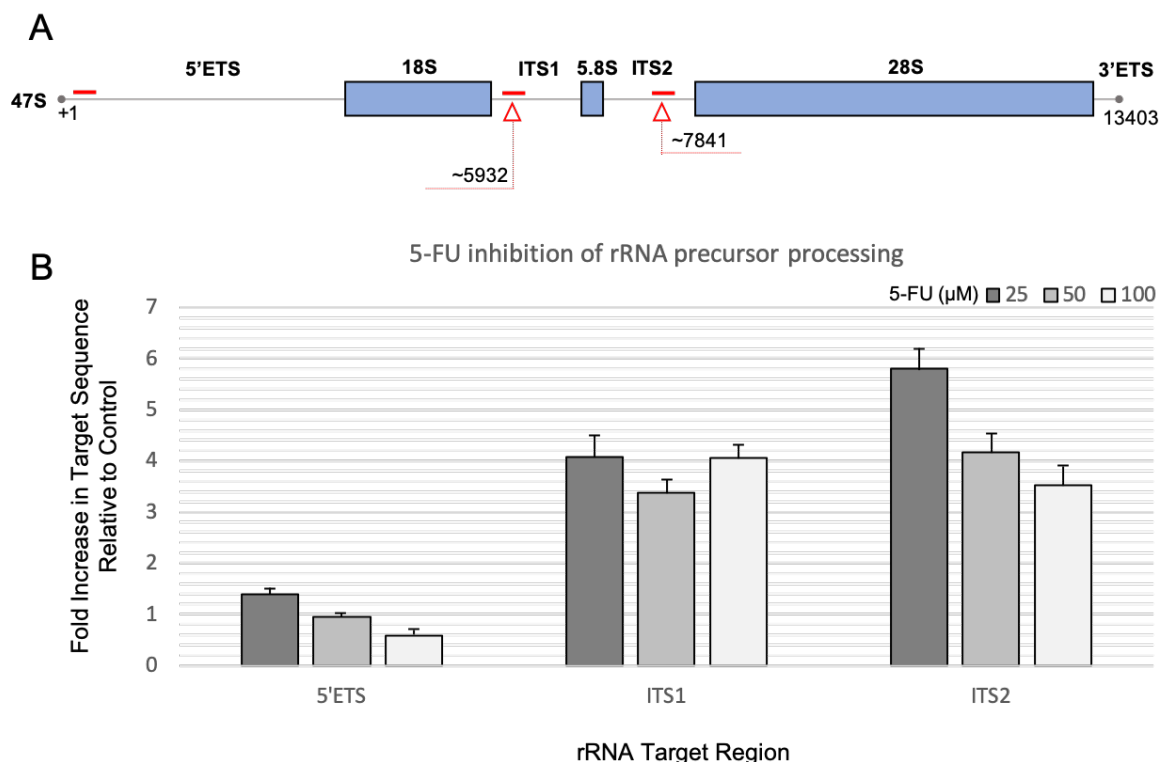


Figure 4.10 Validation of 5-FU inhibition of rRNA processing in MEF cells with qPCR. (A) Schematic of 47S rRNA precursor molecule with key elements indicated. Red bars indicate target sites in 5'ETS, ITS¹ and ITS² for qPCR amplification. Target sites in ITS¹ and ITS² span known cleavage sites at ~5932 bp and ~7841 bp respectively. (B) qPCR detection of target sites in 5-FU treated cells (25, 50, 100 μM 5-FU) relative to control. Fold increase in target sequences was determined using qRT-PCR with intercalation of SYBR Green using the $2^{-\Delta\Delta C_t}$ formula. Data are presented as an average of biological duplicates each run as technical triplicates, normalised to expression of control genes MAPK1 and ITGB1. Technical triplicates were considered reliable if $\Delta C_t < 0.5$. Error bars indicate \pm SD of biological replicates.

Compared to control, the increased detection of 5'ETS, ITS¹ and ITS² targeted sites in 25 μM 5-FU exposed MEFs would suggest effective disruption of pre-rRNA maturation at various processing stages. The greatest degree of disruption was observed for late-stage rRNA processing, indicated by the near 6-fold increase in detection of intact ITS² cleavage site compared to control. Based on the effect on pre-rRNA processing and the observed effect on cell cycle arrest, 25 μM 5-FU exposure was implemented as an additional step in the capture of rRNA processing intermediates. MEFs were treated with 25 μM 5-FU for 24 hours and harvested for their nuclei. The nuclear-extracted RNA was subjected to bead size selection before *in vitro* poly(A) tail addition. Oligo (dT) bead enriched Poly(A)+ RNA was then used for cDNA library preparation.

PromethION cDNA sequencing of size-selected, nuclear RNA from 5-FU treated MEFs, yielded 138,830 rRNA reads, with a mean read length of 974 bp. The coverage depth across the rDNA unit is plotted in **Figure 4.11A**, with all individual reads aligning to the rDNA coding unit presented in **Figure 4.11B**. Assessment of pre-rRNA reads yielded 99 reads completely spanning ITS elements (**Figure 4.11C**).

Long-Read Sequencing Analysis of Ribosomal RNA Modifications

From these, 44 reads completely spanned ITS¹ (0.0317 % of total rRNA reads), and 71 completely spanned ITS² (0.0511 % of total rRNA reads). Additionally, 16 reads were found to completely span both ITS elements (**Figure 4.11D**). Sequencing summary statistics are presented in **Figure 4.11E**. In comparison to the size selected, nuclear RNA, ONT cDNA sequencing of size-selected, nuclear RNA from 5-FU treated MEFs, resulted in a >2.5-fold increase in reads mapping completely across ITS¹. Correlating with the increase in intact ITS¹ sequence quantified via qPCR. However, a slight decrease (0.0628 % to 0.0511 %) was observed in reads mapping completely across ITS².

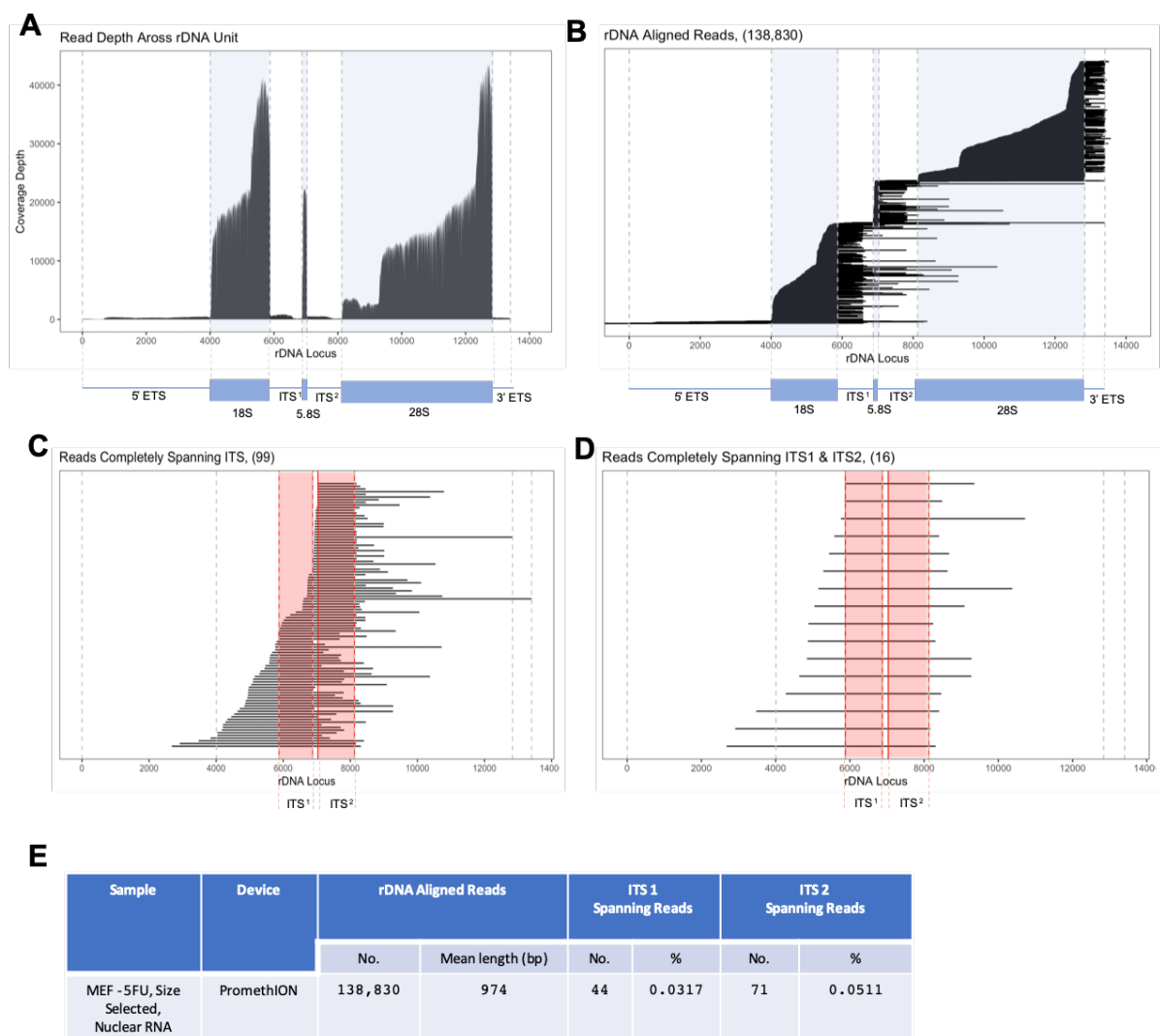


Figure 4.11 Nanopore cDNA sequencing of 5FU treated, MEF, size selected, nuclear RNA. (A) Read coverage depth across rDNA unit is aligned to a schematic of rDNA coding unit (1-13403 bp). Grey dotted lines indicate boundaries of rDNA coding unit elements (left to right: 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS). (B) A stacked plot of all individual reads mapping to rDNA (138,830 reads). Plot is aligned to a schematic of rDNA coding unit (1-13403 bp). Vertical dotted lines indicate boundaries of rDNA coding unit elements (left to right: 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS). (C) Stacked plot of all reads completely spanning ITS elements (99 reads). Reads are defined as such if they map across the entirety of ITS¹ or ITS², +1 bp beyond, on both 5' and 3' ends. (D) Stacked plot of all reads completely spanning ITS¹ and ITS² (16 reads). Reads are defined as such if they map across the entirety of ITS¹ and ITS², +1 bp beyond, on both 5' and 3' ends. (E) cDNA sequencing summary statistics. All rRNA reads are mapped to the published consensus sequence (Accession No. BK000964.3). MEF nuclear RNA was size selected using a bead-to-sample volume ratio of 0.35:1.

Considering the results from this sequencing run, it would appear that 5-FU treatment of MEFs results in the increased capture of reads spanning ITS¹, with a slight decrease in complete coverage of ITS². The reason for the reduction in ITS² mapping is contrary to the expected outcomes, and the reasons remain unclear. The observed reduction could be a true biological variation or due to any number of inconsistencies during the library preparation. As the observation is based on the comparison of only 1 sequencing run per condition, the assessment remains inconclusive. Additionally, it remains unclear how 5-FU treatment impacts downstream Nanopore sequencing analysis. Owing to the structural and chemical similarities of 5-FU and the RNA nucleotide uracil, 5-FU is actively incorporated into actively transcribed in the place of uracil. This is considered a primary mechanism by which it exerts its cellular effect, via hindering the enzymatic processing of rRNA. However, due to the sensitivity of Nanopore sequence detection methods, slight changes to the chemical composition of a nucleotide can translate to markedly different current signal readings. Generally, algorithms employed for base-calling Nanopore sequencing data are trained to identify the current signatures of the 5 basic nucleotides, A, T, U, C, and G. Base-calling algorithms can be trained to identify alternative RNA base modifications. Currently, however, there are no widely-accessible tools for accurately identifying and distinguishing 5-FU. As a result, 5-FU incorporation into RNA could likely lead to errors in base calling and present an issue with any downstream analyses. Moreover, due to the substantial increase in the capture of pre-rRNA processing intermediates achieved with size selection of nuclear RNA compared to total RNA, it was decided that 5-FU treatment would be excluded from future sequencing preparations.

Overall, the protocol refinement outlined here, including sub-cellular fractioning, size selection, and 5-FU induced rRNA processing inhibition, appears to allow for increased capture of rRNA precursor processing intermediates. **Figure 4.12A** presents a comparison of sequencing runs for each step in the sample pre-processing protocol refinement of Nanopore cDNA sequencing. Fold differences in ITS coverage between sequencing runs are presented in **Figure 4.12B**. To summarise, compared with total MEF RNA, a nuclear extract results in a ~2-fold increase in the capture of complete ITS¹ and ITS² sequences. Implementing an additional size selection step was shown to further increase the capture of these regions with a ~10- and ~24-fold increase seen for ITS¹ and ITS², respectively compared with total RNA. The exposure of cells to 25 μ M 5-FU before nuclear RNA extraction and size selection further increased the capture of ITS¹ and ITS² spanning reads ~20- and ~24-fold respectively when compared with total RNA.

A

Sample (MEF RNA)	Device	rDNA Aligned Reads		ITS 1 Spanning Reads		ITS 2 Spanning Reads	
		No.	Mean length (bp)	No.	%	No.	%
Total	MinION	79,949	841	0	0.0000	2	0.0025
Nuclear	MinION	249,086	874	6	0.0024	14	0.0056
Size selected, nuclear	MinION	23,895	962	3	0.0126	15	0.0628
5-FU treated, size selected, nuclear	PromethION	138,830	974	44	0.0317	71	0.0511

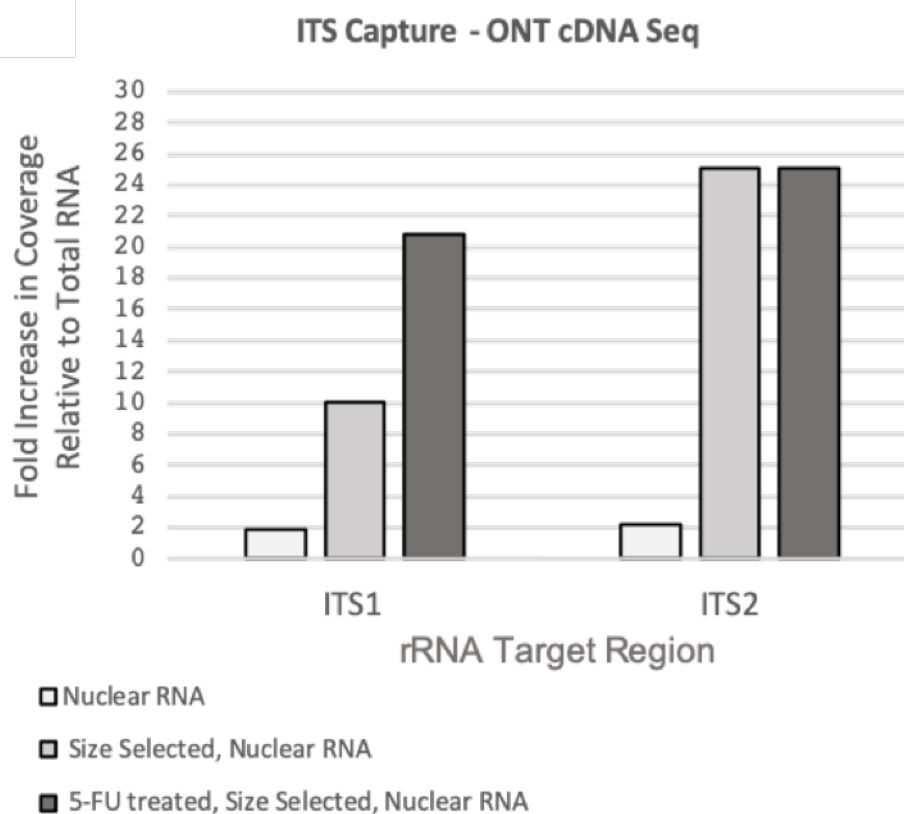
B


Figure 4.12 Comparison of ONT cDNA sequencing runs. (A) Summary statistic of ONT cDNA sequencing runs for various MEF RNA preparation tested. (B) Fold increase in ITS spanning reads across all samples sequenced, presented relative to MEF Total RNA sequencing.

4.3.5 ONT Direct RNA sequencing of ribosomal RNA

Having refined the sample pre-processing protocol for increasing the capture of pre-rRNA processing intermediates with ONT cDNA sequencing, the focus was shifted to ONT direct RNA sequencing (DRS). In contrast to ONT cDNA sequencing, which relies on reverse transcription and subsequent removal of the RNA template, ONT DRS directly sequences full-length transcripts, without prior amplification. This preserves the RNA's native state, enabling the detection of RNA modifications. To assess if the ONT cDNA protocol development yielded similar improvements in rRNA precursor coverage, comparable MEF RNA preparations were sequenced with ONT DRS: total, nuclear and size-selected nuclear RNA. Each RNA preparation was pre-processed as outlined above for its comparable cDNA counterpart.

PromethION DRS of MEF total RNA, yielded 182,658 rRNA reads, with a mean read length of 1308 bp. The coverage depth across the rDNA unit is plotted in **Figure 4.13A**, with all individual reads aligning to the rDNA coding unit presented in **Figure 4.13B**. Assessment of pre-rRNA reads yielded 22 reads completely spanning ITS elements (**Figure 4.13C**). From these, 12 reads completely spanned ITS¹ (0.0066 % of total rRNA reads), and 14 completely spanned ITS² (0.0077 % of total rRNA reads). Additionally, 4 reads were found to completely span both ITS elements (**Figure 4.13D**). Sequencing summary statistics are presented in **Figure 4.13E**.

PromethION DRS of MEF nuclear RNA, yielded 73,649 rRNA reads, with a mean read length of 1,588 bp. The coverage depth across the rDNA unit is plotted in **Figure 4.14A**, with all individual reads aligning to the rDNA coding unit presented in **Figure 4.14B**. Assessment of pre-rRNA reads yielded 16 reads completely spanning ITS elements (**Figure 4.14C**). From these, 10 reads completely spanned ITS¹ (0.0136 % of total rRNA reads), and 8 completely spanned ITS² (0.0109 % of total rRNA reads). Additionally, 4 reads were found to completely span both ITS elements (**Figure 4.14D**). Sequencing summary statistics are presented in **Figure 4.14E**. Compared to total RNA, mean read length increased by 280 bp, and a >2-fold increase in ITS¹ spanning reads was observed, alongside a ~1.4- fold increase in ITS² spanning reads.

PromethION DRS of size-selected MEF nuclear RNA yielded 237,995 rRNA reads, with a mean read length of 1,780 bp. The coverage depth across the rDNA unit is plotted in **Figure 4.14A**, with all individual reads aligning to the rDNA coding unit presented in **Figure 4.14B**. Assessment of pre-rRNA reads yielded 218 reads completely spanning ITS elements (**Figure 4.14C**). From these, 105 reads completely spanned ITS¹ (0.0441 % of total rRNA reads), and 140 completely spanned ITS² (0.0588 % of total rRNA reads). Additionally, 27 reads were found to completely span both ITS elements (**Figure**

4.15D). Sequencing summary statistics are presented in **Figure 4.15E**. Compared to nuclear RNA, mean read length increased by 192 bp, and a >3.2-fold increase in ITS¹ spanning reads was observed, alongside a >5.3- fold increase in ITS² spanning reads.

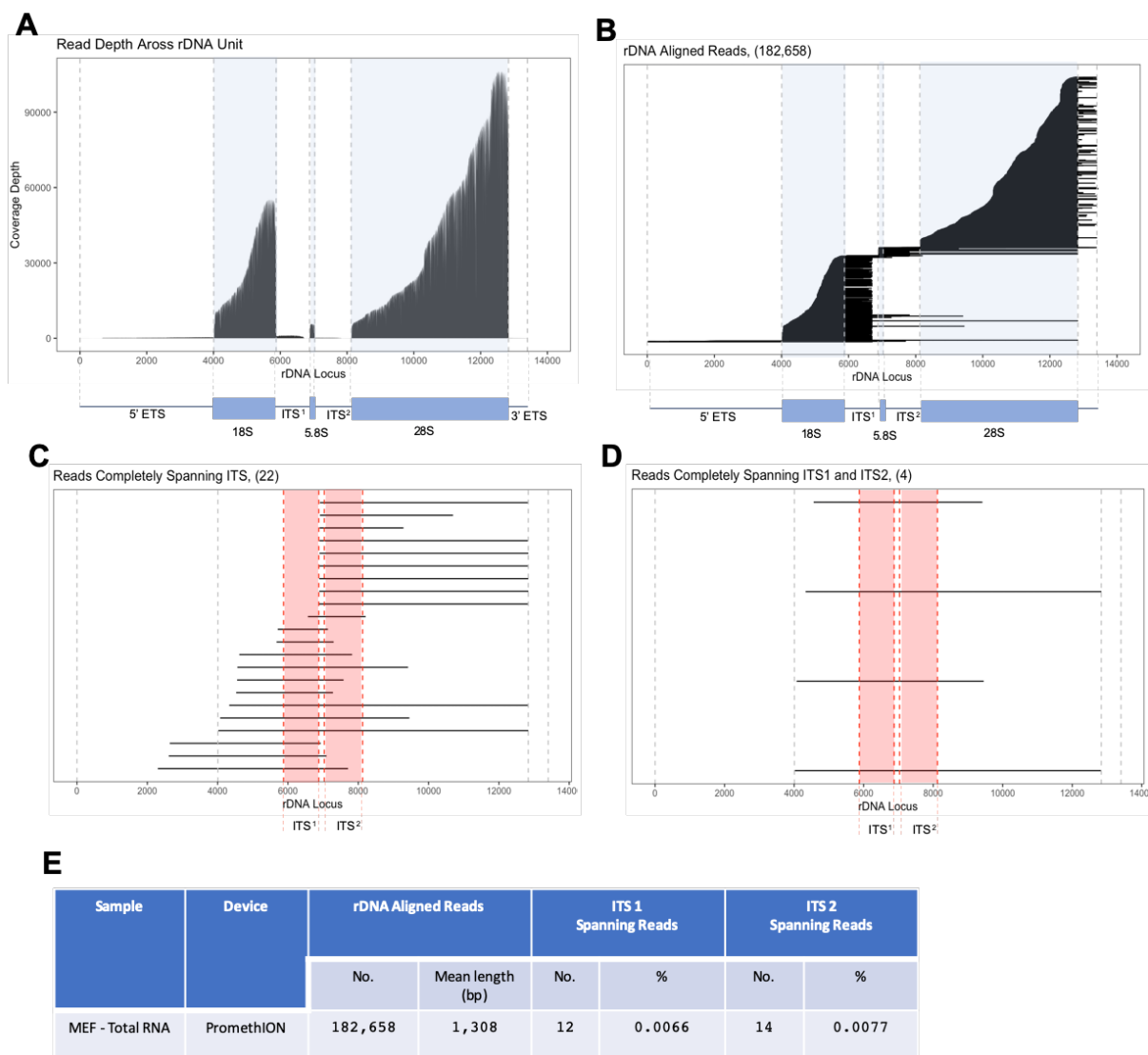


Figure 4.13 Nanopore DRS of MEF total RNA. **(A)** Read coverage depth across rDNA unit presented for pilot 1 (red), 2 (blue) and 3 (purple). Coverage depth plot is aligned to a schematic of rDNA coding unit (1-13403 bp). Grey dotted lines indicate boundaries of rDNA coding unit elements (left to right: 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS). **(B)** A stacked plot of all individual reads mapping to rDNA (182,658 reads). Plot is aligned to a schematic of rDNA coding unit (1-13403 bp). Vertical dotted lines indicate boundaries of rDNA coding unit elements (left to right: 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS). **(C)** Stacked plot of all reads completely spanning ITS elements (22 reads). Reads are defined as such if they map across the entirety of ITS¹ or ITS², +1 bp beyond, on both 5' and 3' ends. **(D)** Stacked plot of all reads completely spanning ITS¹ and ITS² (4 reads). Reads are defined as such if they map across the entirety of ITS¹ and ITS², +1 bp beyond, on both 5' and 3' ends. **(E)** DRS summary statistics

All rRNA reads are mapped to the published consensus sequence (Accession No. BK000964.3).

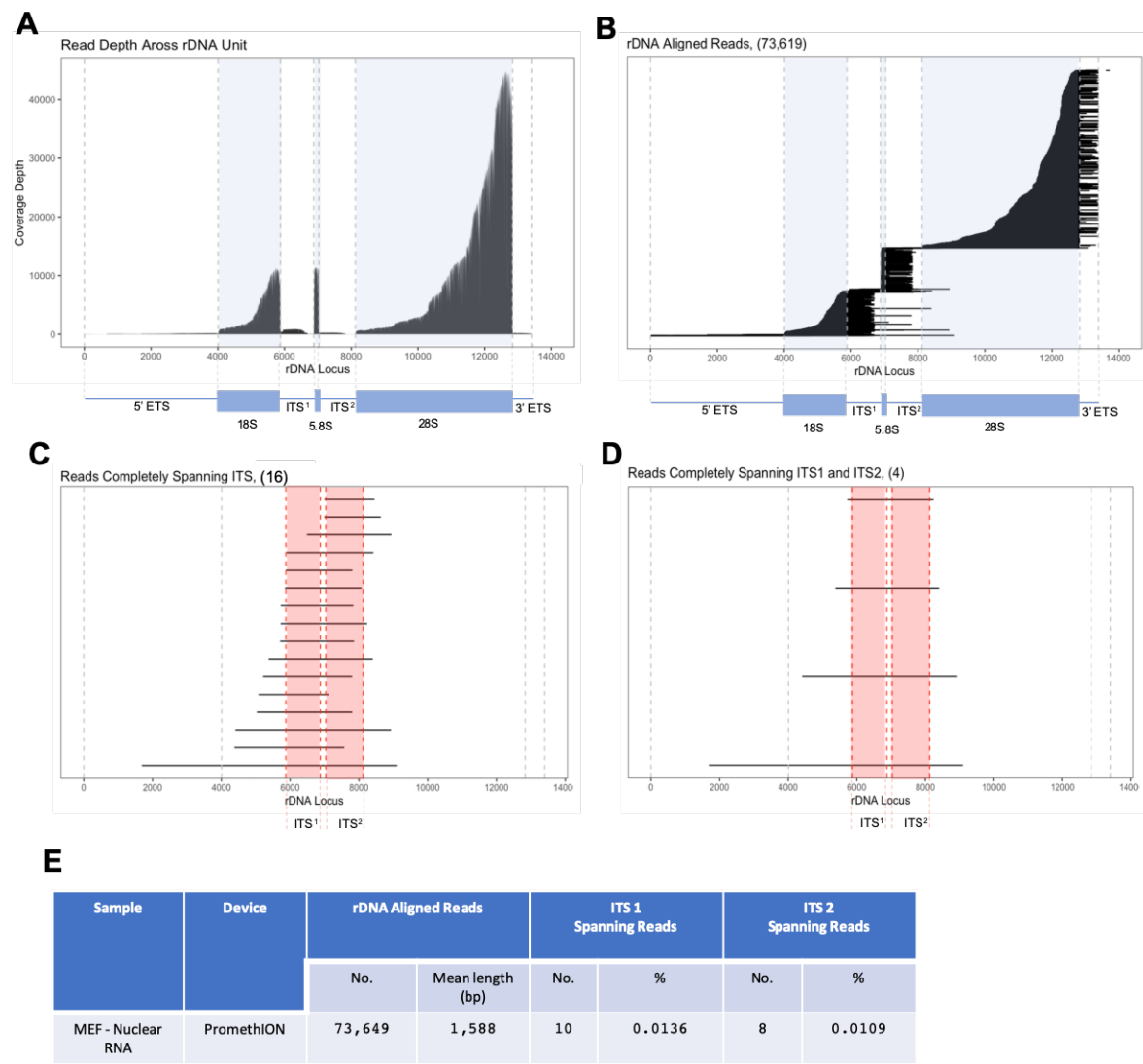


Figure 4.14 Nanopore DRS of MEF nuclear RNA. (A) Read coverage depth across rDNA unit presented for pilot 1(red), 2(blue) and 3(purple). Coverage depth plot is aligned to a schematic of rDNA coding unit (1-13403 bp). Grey dotted lines indicate boundaries of rDNA coding unit elements (left to right: 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS). (B) A stacked plot of all individual reads mapping to rDNA (73,619 reads). Plot is aligned to a schematic of rDNA coding unit (1-13403 bp). Vertical dotted lines indicate boundaries of rDNA coding unit elements (left to right: 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS). (C) Stacked plot of all reads completely spanning ITS elements (16 reads). Reads are defined as such if they map across the entirety of ITS¹ or ITS², +1 bp beyond, on both 5' and 3' ends. (D) Stacked plot of all reads completely spanning ITS¹ and ITS² (4 reads). Reads are defined as such if they map across the entirety of ITS¹ and ITS², +1 bp beyond, on both 5' and 3' ends. (E) DRS summary statistics

All rRNA reads are mapped to the published consensus sequence (Accession No. BK000964.3).

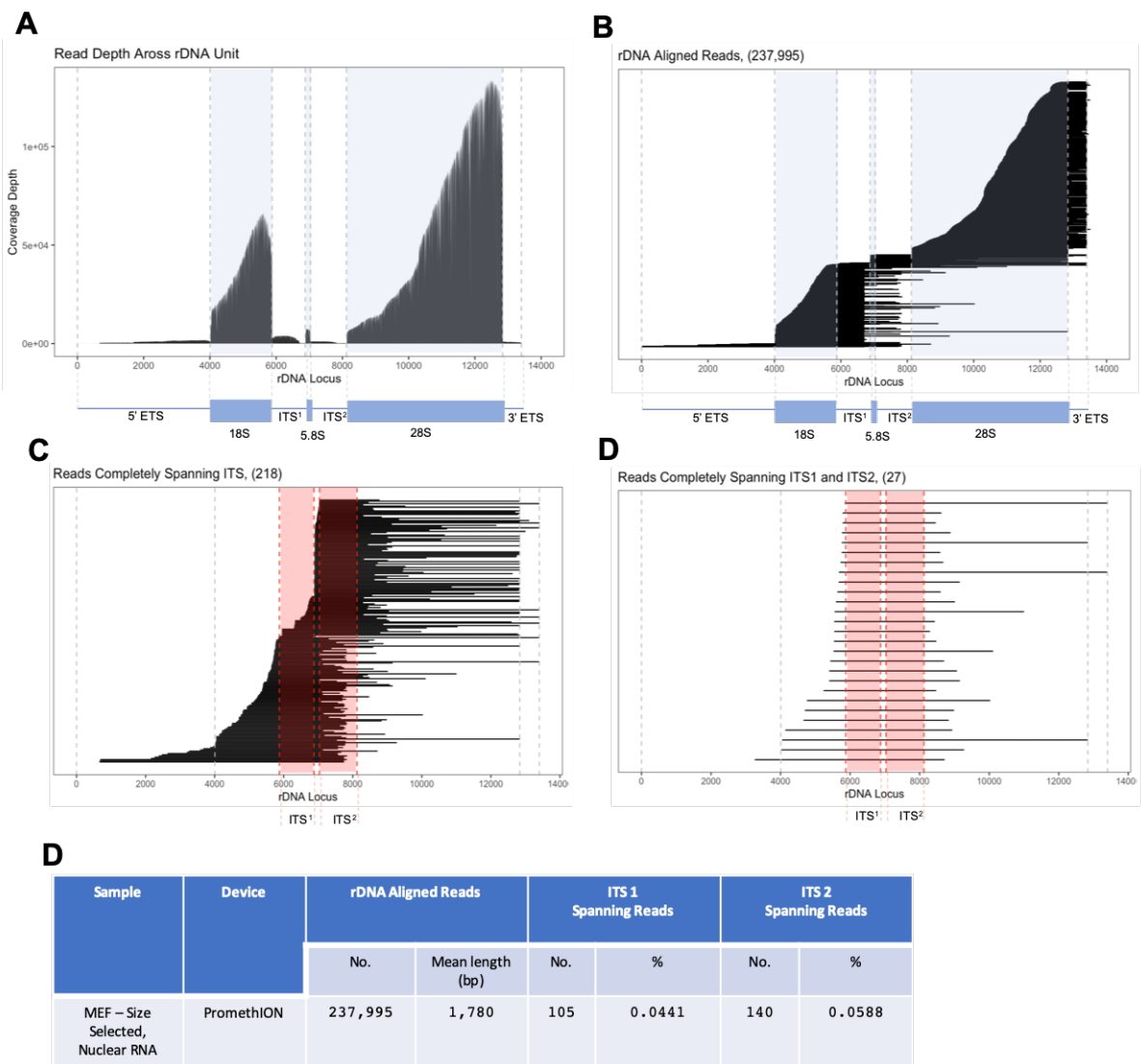


Figure 4.15 Nanopore DRS of MEF, size selected, nuclear RNA. (A) Read coverage depth across rDNA unit presented for pilot 1(red), 2(blue) and 3(purple). Coverage depth plot is aligned to a schematic of rDNA coding unit (1-13403 bp). Grey dotted lines indicate boundaries of rDNA coding unit elements, (left to right- 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS). (B) A stacked plot of all individual reads mapping to rDNA (237,995 reads). Plot is aligned to a schematic of rDNA coding unit (1-13403 bp). Vertical dotted lines indicate boundaries of rDNA coding unit elements (left to right: 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS). (C) Stacked plot of all reads completely spanning ITS elements (218 reads). Reads are defined as such if they map across the entirety of ITS¹ or ITS², +1 bp beyond, on both 5' and 3' ends. (D) Stacked plot of all reads completely spanning ITS¹ and ITS² (27 reads). Reads are defined as such if they map across the entirety of ITS¹ and ITS², +1 bp beyond, on both 5' and 3' ends. (E) DRS summary statistics
All rRNA reads are mapped to the published consensus sequence (Accession No. BK000964.3).

For the DRS runs described above, the results suggest that sample pre-processing protocol optimisation, as outlined for ONT cDNA-Seq, translated positively to ONT DRS sequencing of MEF RNA. Generally, DRS outperformed cDNA-seq in several metrics, including rRNA read length, transcribed spacer capture and maximum read length. Comparing size-selected, nuclear RNA preparations for cDNA sequencing and DRS, the mean read length increased from 962 bp to 1780 bp, whilst the longest captured rRNA read measured 7182 bp and 8968 bp for each respective sequencing approach.

Additionally, a 3.5-fold increase in reads completely covering ITS¹ with ITS² coverage between comparable between the two.

4.3.6 *In vitro* 5' capping of rRNA increases the 5' coverage of Nanopore DRS reads

Assessing the overall coverage of the rDNA locus for both cDNA and direct RNA sequencing, a distinct increase in coverage is noted at the 3' end. This is most evident when examining coverage of both the 18S and 28S coding subunits, owing to their larger size relative to the 5.8S coding subunit. The 3' bias is considered inherent to DRS, due to sequence reads being generated 3'→5'. However, considering 3' bias is also observed for the cDNA sequencing runs discussed above, read generation directionality may not be the only contributing factor. The reason for read 3' bias thus remains unclear. Nevertheless, possible causes could include the artificial fragmentation of RNA molecules during library preparation, or the natural activity of 3' exonucleases. Whatever the cause, the loss of coverage at the 5' end hinders the maximal read length and poses an issue in capturing full-length pre-rRNA transcripts. Unlike messenger RNAs, rRNAs lack a 5' cap structure, which serves to protect transcripts from premature degradation. Considering this, it was hypothesised that the *in vitro* addition of a 5' cap could act to stabilise rRNAs and protect from potential 5' degradation, increasing the length of pre-rRNA molecules captured. In a bid to improve the stability of extracted rRNA, MEF nuclear RNA was subjected to the *in vitro* addition of a 7-methylguanylate 5' cap. RNA was then poly(A) tailed and oligo (dT) bead enriched, after which it was used as input for ONT DRS library preparation.

PromethION DRS of MEF *in vitro* 5' capped, nuclear RNA, yielded 311,343 rRNA reads, with a mean read length of 1,540 bp. A comparison of the coverage depth across the rDNA unit is presented in **Figure 4.16A** alongside that of 'uncapped' MEF nuclear RNA, used as a baseline to assess the impact of *in vitro* 5' capping on rDNA coverage. Coverage depth is presented as a % of total rRNA reads to account for dissimilarities in the overall read number between the two sequencing runs. An overall 3' bias was observed for both samples. Even so, a clear increase in 5' coverage in the case of 5' capped nuclear RNA was seen. Assessment of *in vitro* 5' capping on pre-rRNA capture, produced 97 reads completely spanning ITS¹ (0.0312 % of total rRNA reads), and 95 reads completely spanning ITS² (0.0305 % of total rRNA reads) (**Figure 4.16B**). This translates to a ~2.9-fold increase in complete ITS¹ coverage and a ~2.8-fold increase in ITS² coverage when compared to uncapped nuclear RNA. A slight reduction in mean read length was noted (1,588 to 1,540 bp), likely due to the increased manipulation of RNA during the 5' capping process. Considering the finding from these analyses, *in vitro* 5' capping was deemed a potentially useful tool in increasing the capture of pre-rRNA processing intermediate, and was therefore introduced as an additional step in every subsequent DRS sequencing preparation.

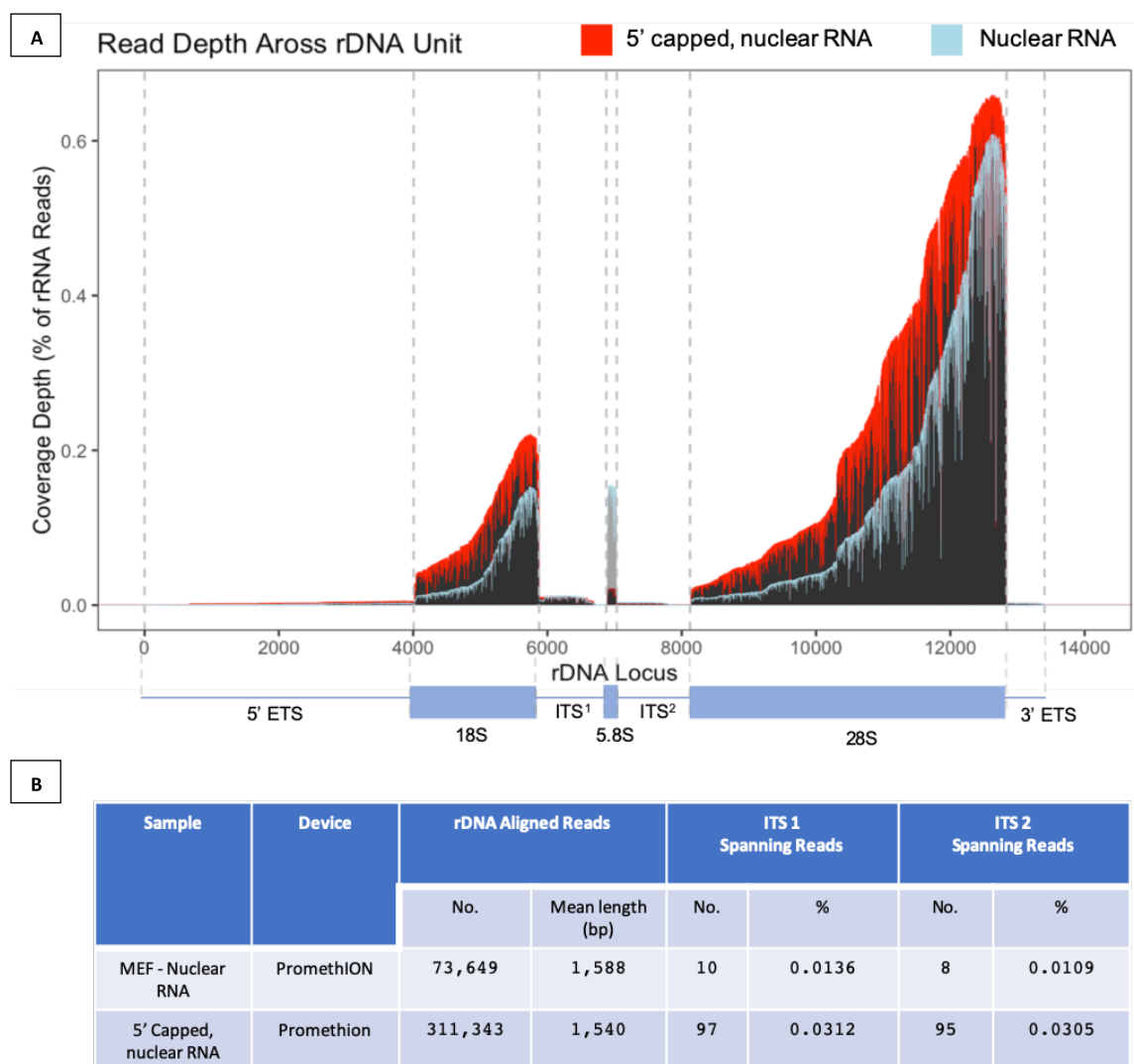


Figure 4.16 Nanopore DRS of MEF, in vitro 5' capped nuclear RNA. (A) Read coverage depth across rDNA unit presented for MEF nuclear RNA (blue) and MEF in vitro 5' capped nuclear RNA. Coverage depth (y-axis) is presented as the % of total rRNA reads, accounting for differences in read output between the two sequencing runs. Plot is aligned to a schematic of rDNA coding unit (1-13403 bp). Grey dotted lines indicate boundaries of rDNA coding unit elements (left to right: 5'ETS, 18S, ITS¹, 5.8S, ITS², 28S, 3'ETS) **(B)** DRS summary statistics.

In summary, the various steps taken to increase the capture of pre-rRNA processing intermediates have led to the development of an improved sample pre-processing protocol for ONT RNA-Seq, outlined in **Figure 4.17**. Libraries prepared with one or more of the outlined developments have demonstrated general improvements in mean read length, as well as the capture of rRNA transcripts spanning across transcribed spacer elements and mapping to multiple coding subunits. Therefore, the protocol development outlined here may serve to facilitate the study of rRNA modifications across multiple rRNA coding subunits, in a haplotype-specific context.

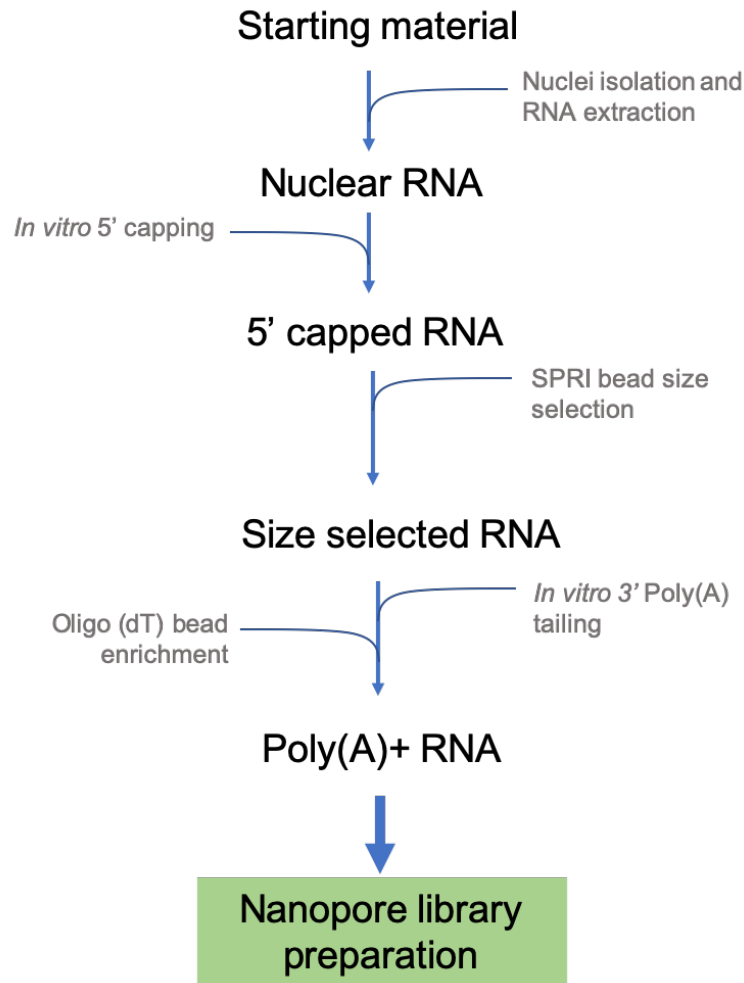


Figure 4.17 Optimised sample pre-processing protocol for nanopore sequencing of ribosomal RNA. All steps assessed to improve the capture of pre-rRNA are introduced to the sample pre-preparation protocol.

4.3.7 ONT DRS data set generation

To explore rRNA haplotype expression and modification profiles across different cell types, tissue and developmental stages, a range of ONT DRS data sets were generated with ONT MinION and PromethION devices (**Table 4.1**). Data sets were generated for cultured cell types, including MEFs and MESNC. Embryoid bodies (EBs) differentiated from a culture of MESNCs were also sequenced to evaluate rRNA haplotypes within a developmental context. The validation of EB development was confirmed through the immunofluorescent visualisation of germ layer specific protein markers GATA-4 (endoderm), SMA (mesoderm) and β III-Tubulin (ectoderm), presented in **Figure 4.18**. Adhered EB's were stained individually for each marker, with cell populations expressing markers for each individual germ layer positively identified alongside cells absent of markers, indicative of

multilineage EB germ layer differentiation. Additionally, sequencing data was generated “A”, “B” and “C”. Liver tissue was sequenced for each individual, with the additional sequencing of pancreatic tissue from “B” and kidney tissue from “C”. The sequencing data generated in this study was produced using both MinION and PromethION devices, using a combination of new and previously used flow cells. This is reflected in the dramatic differences in sequencing output for certain runs described here. Additionally, multiple sequencing runs were conducted for certain samples, to maximise total read output and facilitate downstream analyses.

Device	Sample	Sequencing run #	rRNA read output	Total rRNA reads
MinION	MEF	1 *	125,678	636,685
		2	86,210	
		3 *	265,699	
		4 *	159,098	
	MESC	1	101,875	285,703
		2	15,402	
		3	119,293	
		4	49,133	
	EB	1 *	201,466	341,834
		2 *	140,368	
	Liver (A)	1 *	194,429	194,429
PromethION	MEF	1	311,343	835,228
		2	73,619	
		3	182,658	
		4	237,995	
		5	29,613	
	MESC	1	110,118	131,189
		2	21,071	
	Liver (B)	1	12,238	171,609
		2	25,000	
		3	134,371	
	Pancreas (B)	1	47,813	134,056
		2	86,243	
	Liver (C)	1	25,603	75,459
		2	42,201	
		3	7,655	
	Kidney (C)	1	14,975	14,975

Table 4.1 ONT DRS data sets generated on MinION and PromethION devices. Sample type and number of sequencing runs specified. Total rRNA read output for each sequencing run is presented alongside a total for each of the different samples, per device. (*) denotes runs conducted on new flow cells at optimal sequencing capacity (available pores > 75% of total capacity). All other runs were conducted on pre-used/ washed flow cells of varying sequencing capacity.

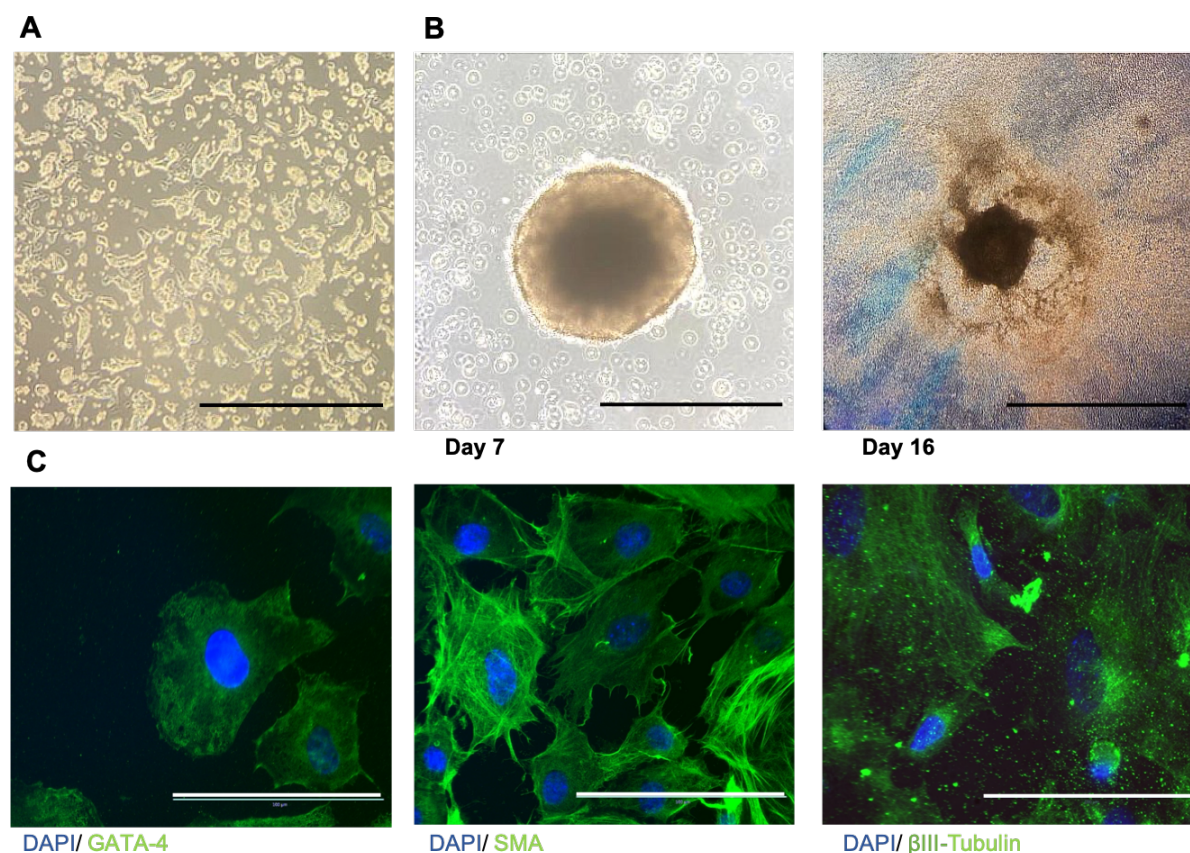


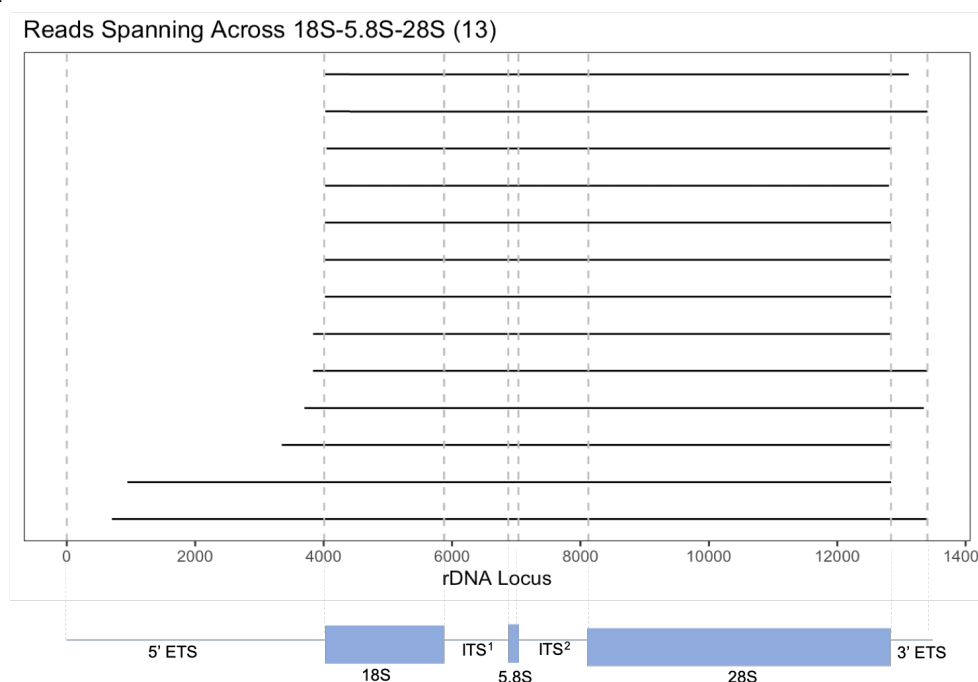
Figure 4.18 EB formation and differentiation validation (A) Feeder free MESCs cultured in 2i media; pluripotency was validated via qPCR assessment of pluripotency markers NANOG, OCT4 and SOX2. (B) MESC seeded in EB differentiation media aggregate into distinct spheroid masses after 7 days of culture in super low adherence plates (left). EB transferred to gelatinised coverslip, and differentiated for a total of 16 days (right). (C) immunofluorescent staining of differentiated EB with antibodies targeting germ layer specific markers GATA4 (endoderm), SMA (mesoderm) and βIII-tubulin (ectoderm). Counterstained with DAPI
Scale bar= 50 μm (A-B), 20 μm (C)

4.3.8 ONT DRS allows for the capture of near full-length rRNA primary transcripts

The sample pre-preparation development outlined in this chapter, allowed for the capture of considerably large pre-rRNA transcripts, spanning multiple coding subunits. PromethION sequenced MEF DRS data from 4 sequencing runs was combined to produce a data set with 835,228 rRNA reads. An assessment of read length revealed the longest single read measuring 12,895 bp, spanning across all 3 coding subunits and the majority of the 5'ETS. From this data set, 13 reads in total were found to span completely across all three coding subunits (**Figure 4.19A**). Similarly, MinION sequenced MESC DRS data from 4 sequencing runs was combined to produce a total of 285,703 rRNA reads. Size assessment revealed the longest single read to measure 13,378 bp, just 25 bp less than the full-length

primary transcript (13,403 bp). In total 77 individual reads were found to span completely across the 3 coding subunits (**Figure 4.19B**).

A



B

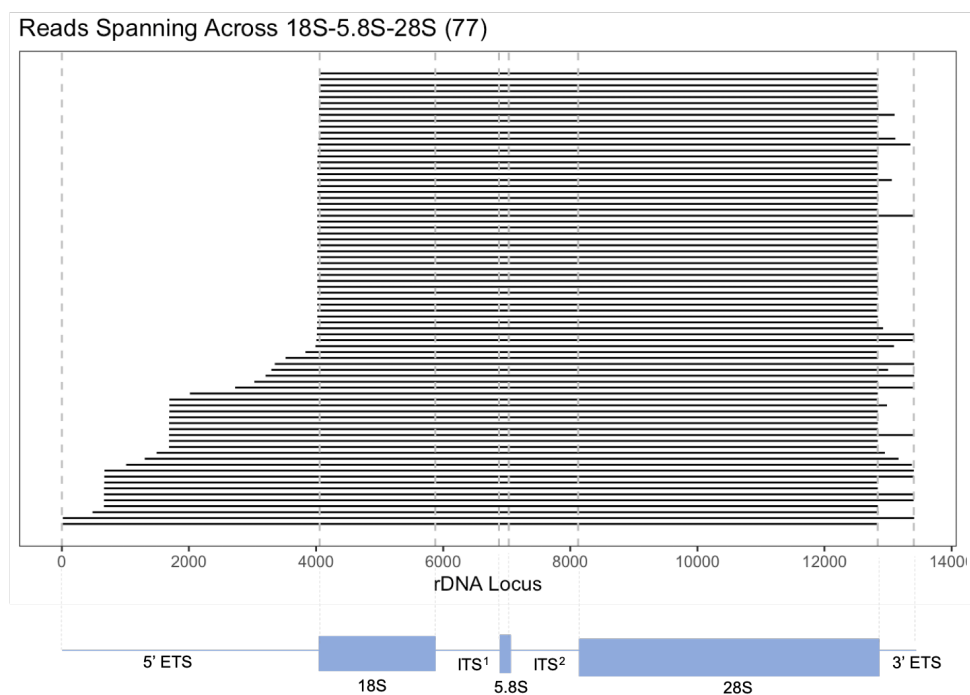


Figure 4.19 Longest pre-rRNA reads captured with ONT DRS. (A) 13 reads span completely across all three coding subunits, reads are aligned to a schematic of rDNA coding unit. Data is presented from a combination of 4 PromethION sequenced MEF DRS data sets (835,228 rRNA reads). **(B)** 77 reads span completely across all three coding subunits, reads are aligned to a schematic of rDNA coding unit. Data is presented from a combination of 4 MinION sequenced MESC DRS data sets (285,703 rRNA reads).

4.3.9 Expression of rRNA haplotypes is cell type-specific

The expression of rRNA haplotypes “ATA” and “ATG” was assessed in MEFs, MESCs and MESC-derived EBs. Haplotypes were distinguished by assessing SNPs at positions 6007, 6777, 6832 and 12736. Reads expressing the “ATA” haplotype were identified if they presented with guanine ‘G’ at position 6007 or an adenine ‘A’ at position 6832. Likewise, reads expressing the “ATG” haplotype were distinguished if they presented with ‘A’ at position 6777 or ‘G’ at position 12736. DRS data for each sample type was filtered to extract reads for which haplotype-specific positions (6007, 6777, 6832, and 12736) had been accurately base-called, i.e., the individual read had been confidently assigned a nucleotide at the specified positions. Across all datasets and positions assessed, accurately base-called reads made up 40-50 % of the total reads covering a position. Assessing position-specific coverage, it emerged that each of the 4 positions was unequally represented. Coverage of each position across 15 different DRS runs is presented in **Figure 4.20**. The coverage of position 12,736 (positioned at the far 3’ end of 28s rRNA) was disproportionately higher in all sequencing runs presented, owing to the intrinsic 3’ bias of DRS. Contrastingly, coverage of positions (6007, 6777, 6832), which occur within a transcribed spacer elements (ITS¹) were substantially lower with position 6777 coverage being the least. Considering this, a minimum threshold for position coverage was set to 10 reads, so that a data set was disregarded if any position had less than 10 base called reads assigned to it.

Haplotype expression in MEFs, MESCs and EBs was assessed using data from 5, 5 and 2 sequencing runs respectively, with data generated on both the MinION and PromethION devices. Reads were assessed for the nucleotide composition at the specified position and classified according to a haplotype or as ‘other’, individually, for each sequencing run considered here. Haplotype expression was assessed by considering the fraction of reads expressing a haplotype-specific SNP as a percentage of the total base called read for each position (**Figure 4.21**).

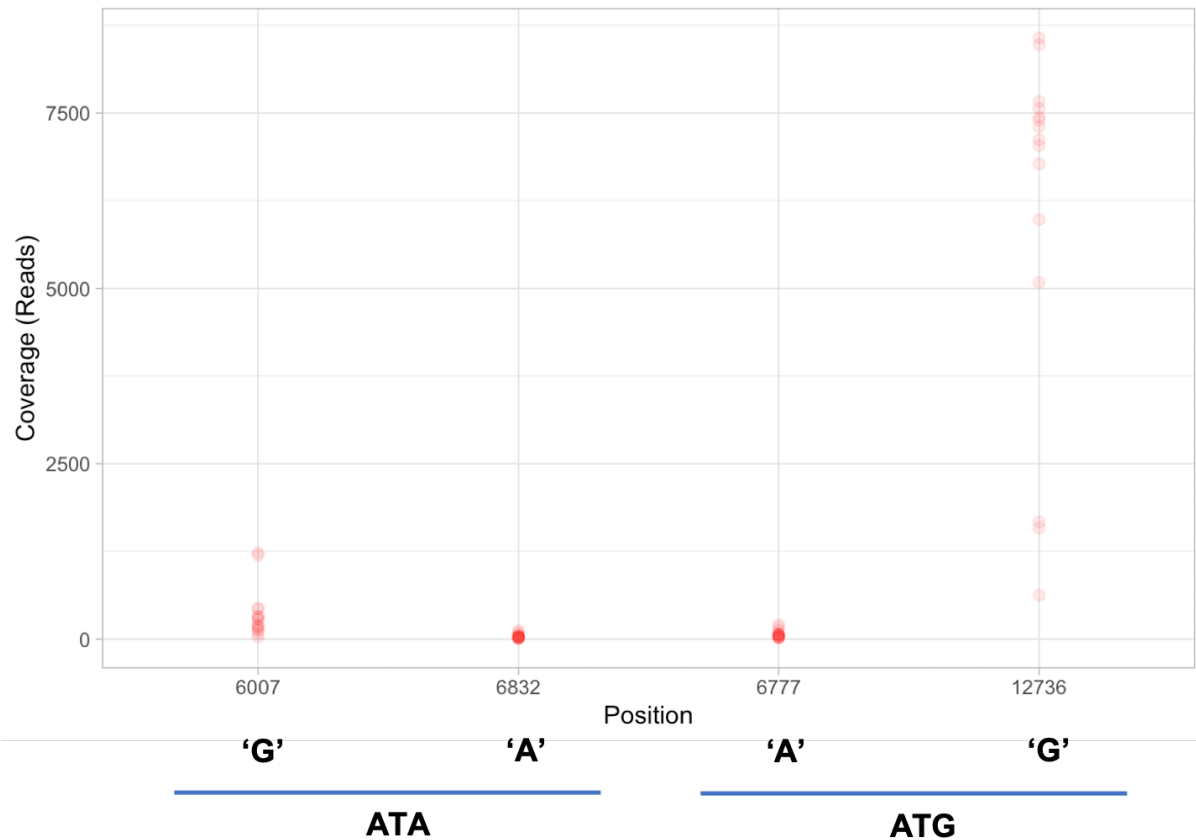


Figure 4.20 Coverage of haplotype specific SNPs is unequal. Haplotype specific positions (x-axis) plotted against the total coverage at each position. Data are presented from 15 DRS data sets (individual red marker).

Figure 4.21 presents the comparison of haplotype expression for the 3 cell types considered here. Expression of the ATA haplotype was significantly greater in MESC when compared with MEFs ($p < 0.01$), when considering SNPs G-6007 and A-6832, with a mean difference in expression of +12% and +22%, respectively. Conversely, expression of the “ATG” haplotype was significantly higher in MEFs compared to MESC ($p < 0.01$), when considering SNP G-12736, with a mean difference in expression of +13%. “ATG” SNP A-6777, displayed no significant difference between the two cell types, likely resulting from the reduced coverage at this position compounded by possibly small differences in expression between the cell types. Haplotype expression in EBs, when considering the level of “ATA” SNPs (G-6007 and A-6832) was greater than that observed for MEFs but lower than that observed for MESC., whilst EB expression of “ATG” SNP A-6777 was greater than that observed for MESC and lower than that observed for MEFs. EB haplotype expression was not statistically assessed due to the limited sample size.

Considering each SNP independently across all datasets and cell types, the positions with the greatest mean coverage were positions 12,736 followed by position 6007. Coverage of positions 6832 and 6777 was substantially lower, with 6777 coverage being the least. Coverage differences are reflected in the range of % SNP expression between data sets of each type, with positions 6007 and 12736 displaying the least amount of overall deviation across data sets for any given SNP. In contrast, for position 6777, all data sets from all three cell types display a considerable range. Since haplotype-specific SNPs are intrinsically linked, the observations made for positions with the highest coverage, i.e. 6007 and 12,736 were used as a determinant of rRNA allele-specific expression. Considering this, the mean expression of the ATA and ATG haplotype in MEFs was considerably different, with ATA expression at ~3% and ATG expression at ~29%. The mean expression of both haplotypes in MESCs was comparable, with a <1% difference when considering positions 6007 and 12736.

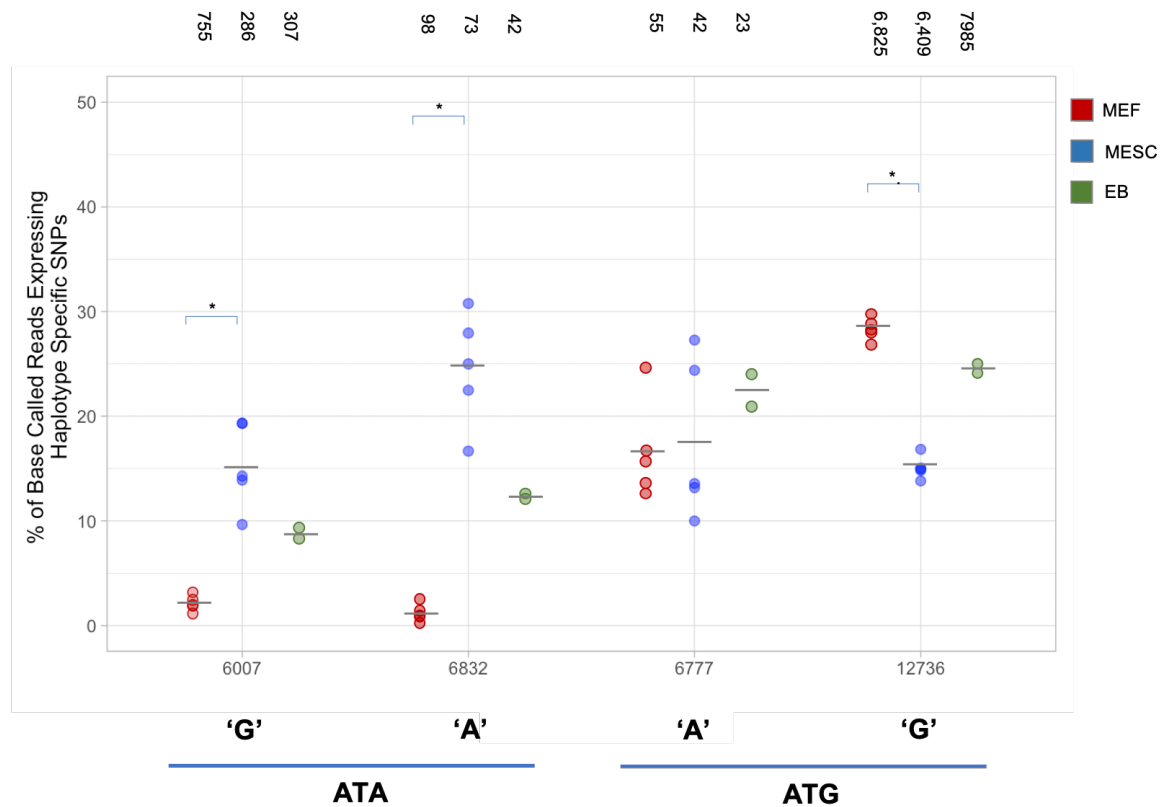


Figure 4.21 rDNA haplotype-identifying alleles are differentially expressed. Plot generated from analysis of Nanopore DRS data, presenting the expression of haplotype-identifying alleles in MEF, MESCs and EBs. Haplotypes, defining SNPs and their respective positions (x-axis) are plotted against the percentage of basecalled reads with the haplotype specific allele (y-axis) for each sample type: MEFs (red, n=5), MESCs (blue, n=6), and EBs (green, n=2). Above each set of data points is the mean number of base called reads assessed per position for each sample type. Each coloured point represents the expression determined from a single sequencing run, whilst horizontal bars indicate mean expression across all sequencing runs per condition. Expression of “ATA” haplotype specific SNPs G-6007 and A-6832 is significantly higher in MESC relative to expression in MEFs ($p < 0.01$, Wilcoxon rank sum test). Expression of “ATG” specific SNP G-12736 is significantly higher in MEFs than in MESCS ($p < 0.01$, Wilcoxon rank sum test). No significant difference is observed in the cell specific expression of “ATG” haplotype SNP-6777. Statistical significance of differential haplotype expression could not be determined for EBs due to the limited sample size.

4.3.10 Analysing rRNA modifications with Nanocompore

Nanocompore, an RNA modification calling tool for Nanopore DRS data sets, was employed to compare the modification profiles of rRNA across the various sample types studied here. In brief, Nanocompore works by considering the current intensity and dwell times of individual read mapping to a specific transcript. Reads are assessed per k-mer (a string of consecutive bases of length k , 5 in this particular case), and a comparison of these parameters in two experimental test conditions is used to predict sites of differential RNA modification. Nanocompore is designed to compare two distinct conditions, a control condition in which a particular RNA modification is completely absent or expressed at considerably lower levels relative to the other, and a test condition in which modification levels are being assessed. Here, however, Nanocompore was used in an attempt to identify potential sites at which rRNA is differentially modified across different samples, rather than deduce positions of RNA modification within a particular condition.

To this end, a baseline of modification-calling ‘noise’ was first established, in which two data sets for which no difference was expected were analysed with Nanocompore. EB DRS data sets, generated from the same RNA sample (accounting for any biological variation), were pre-processed as outlined. Potential modification sites were assessed across the pre-rRNA transcript, for all coding unit elements. Nanocompore uses a bivariate classification method based on 2 components, Gaussian mixture model (GMM) clustering followed by a logistic regression test (logit) to determine if there is a significant difference in the distribution of reads between the two conditions. False rate discovery assessment (FDR) of the logit p-values allowed for the adjustment and refinement of data used in downstream analyses.

Figure 4.22 presents the modification sites flagged across the length of the pre-rRNA transcript, plotted against the $-\log_{10}(\text{FDR})$ of the p-value for a comparison of 2 EB data sets, 1 data set assigned per condition (201,466 compared to 140,368 reads). Using a $-\log_{10}(\text{FDR})$ threshold of 2 (recommended by Leger et al., 2021), and ($\text{FDR} < 0.01$), the analysis revealed potential modification sites within both the 18S and 5.8S transcripts. In this comparison, a k-mer starting at position 1772 of the 18S rRNA was flagged with the greatest confidence, with a $-\log_{10}(\text{FDR})$ of 4.89. This was taken as a cut-off for ‘genuine’ sites of modification with all further analyses conducted using a $-\log_{10}(\text{FDR})$ minimum threshold of +5, equating to an FDR of < 0.00001 .

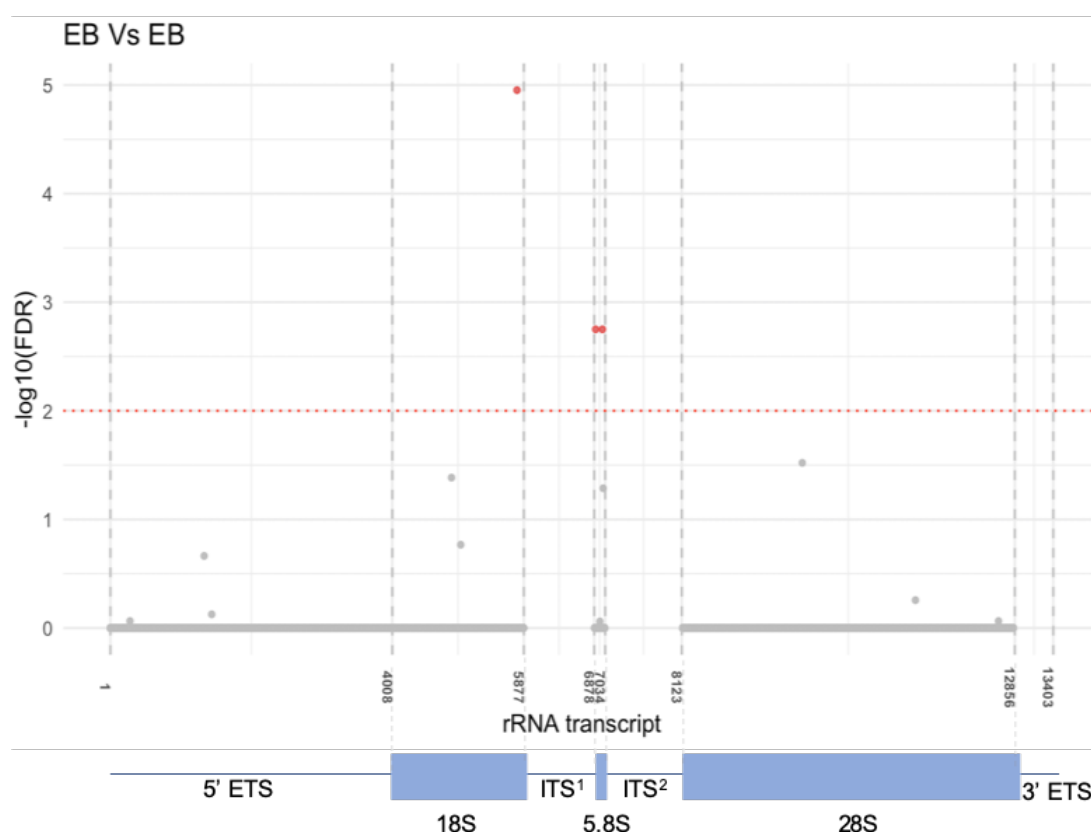


Figure 4.22 Establishing baseline Nanopore modification-calling ‘noise’. A 1:1 Nanopore analysis from two EB DRS data sets. Modification-called positions are plotted according to their position within the pre-rRNA transcript (x-axis) against their $-\log_{10}(\text{FDR})$. Grey points represent individual k-mers with a significance value >0.01 ($-\log_{10}(\text{FDR}) < 2$). All k-mers with $\text{FDR} < 0.01$ ($-\log_{10}(\text{FDR}) > 2$) are presented in red, with a horizontal line indicating the significance threshold. Plot is aligned to a schematic of pre-rRNA transcript. Terminal positions indicate start and end of the 47S rRNA, with secondary positions indicating starting and end-positions of rRNA coding subunits 18S, 5.8S and 28S.

4.3.11 Differential modification profiles are observed between MEF and MESC rRNA

To evaluate the differential modification of rRNA in cell types representative of different developmental stages and cell specialisation lineages, a cross-comparison of DRS data sets from a series of distinct cell/tissue types was conducted. Specifically, in increasing order of development, an rRNA modification comparison of MESC, MESC derived EBs, MEFs and liver tissue DRS data sets was conducted. A Nanopore comparison of MinION-sequenced MESC and EB data sets (4:2) was conducted with an overall comparison of 295,703 and 341,834 reads respectively, the results of which are presented in **Figure 4.23**. All positions flagged via single metric analyses (FDR corrected P-values) as potential sites of interest across the 4 rRNA species examined (pre-rRNA, 18S, 5.8S and 28S) are presented in **Figure 4.23A**, with colored positions identified as being differentially modified above the set level of significance. Results from this comparison suggest the greatest number of flagged positions occur within the 18S and 28S, with comparatively less positions flagged within pre-

and 5.8S rRNA species. For first pass analyses, pre- and 5.8S rRNA species were excluded from further consideration. **Figure 4.23B-C (left panels)** display the positions flagged from single metric analyses for 18S and 28S mature rRNA transcripts, for which Nanocompore analysis identified 78 and 66 positions above the threshold of significance ($-\log_{10}(\text{FDR}) > 5$) respectively. Shark fin plots presented in **Figure 4.23B-C (right panels)** present the analyses from double metric analyses, with FDR corrected P-values plotted against the absolute value of the Nanocompore logistic regression log odds ratio (GMM logit method), with positions of greatest significance considered to fall furthest from the shark fin data cluster. The 10 positions of greatest significance were considered for further analysis and compared to known RNA modification sites in human rRNA (Taoka *et al.*, 2018). Flagged positions corresponding to known sites in human rRNA were designated as potential sites of differential modification if human modifications occurred within +2 bp of the flagged site. The comparison of MESC's and EB's data sets yielded 4 potential sites of differential modification within the 18S, identified as positions 1) U-1441, possibly corresponding to U-1442 in human 18S, a substrate for methylation (Um), 2) G-876, possibly corresponding to G-867 in human 18S, a substrate for methylation (Gm), 3) U-1327, possibly corresponding to U-1328 in human 18S, a substrate for methylation (Um), and 4) G-1328, possibly corresponding to G-1328 in human 18S, a substrate for methylation (Gm). Within the 28S transcript, 3 potential sites were identified: 1) U-4502, possibly corresponding to U4502 in human 28S, a substrate for pseudouridylation (Ψ), 2) U-4323, possibly corresponding to U-4323 in human 28S, a substrate for pseudouridylation (Ψ) and 3) G1509, possibly corresponding to G-1509 in human 28S, a substrate for methylation (Gm). These potential sites of differential modification between MEFs and EBs are displayed in **Figure 4.23D**.

A Nanocompore comparison of MinION-sequenced MEF and MESC data sets (2:4) was also conducted, based on an overall comparison of 284,776 and 295,703 reads respectively, the results of which are presented in **Figure 4.24**. All positions flagged via single metric analyses (FDR corrected P-value) as potential sites of interest across the 4 rRNA species examined (pre-, 18S, 5.8S and 28S) are presented in **Figure 4.24A**, with colored positions identified as being differentially modified above the set level of significance. Results from this comparison suggest the greatest number of flagged positions occur within the 18S and 28S, with comparatively less positions flagged within pre- and 5.8S rRNA species. For first pass analyses, pre- and 5.8S rRNA species were excluded from further consideration. **Figure 4.24B-C (left panels)** present the positions flagged from single metric analyses for each mature rRNA transcript (18S and 28S), for which analysis identified 330 and 420 positions respectively, which were above the threshold of significance ($-\log_{10}(\text{FDR}) > 5$). Shark fin plots presented in **Figure 4.24B-C (right panel)** present the analyses from double metric analyses, with FDR corrected P-values plotted against the absolute value of the Nanocompore logistic regression log odds ratio (GMM logit method), with positions of greatest interest considered to fall furthest from the shark fin

data cluster. For both 18S and 28S mature transcripts, a 3' bias in flagged positions was observed, likely owing to the relative increase in sequencing coverage at the 3' end inherent to Nanopore DRS. Flagged positions corresponding to known sites in human rRNA were designated as potential sites of differential modification if human rRNA modifications occurred within +2 bp of the flagged site. Within the 18S transcript, 3 potential sites were identified: 1) U-1441, possibly corresponding to U-1442 in human 18S, a substrate for methylation (Um), 2) U-1625, possibly corresponding to U-1625 in human 18S, a substrate for pseudouridylation (Ψ), and 3) A-1850, possibly corresponding to A-1850 in human 18S, a substrate for di-methylation (M^6_2A). Within the 28S transcript, 3 potential sites were identified, position 1) U-4502, possibly corresponding to U-4502 in human 28S, a substrate for pseudouridylation (Ψ), 2) C-4507, possibly corresponding to C-4506 in human 28S a substrate for methylation (Cm), and 3) A-3808, possibly corresponding to A-3809 in human 28S a substrate for methylation (Am). These potential sites of differential modification between MEFs and MESCs are displayed in **Figure 4.24D**.

Similarly, a Nanocompore comparison of MinION-sequenced MEF and EB data sets (2:2) was conducted based on an overall comparison of 284,776 and 341,834 reads respectively, the results of which are presented in **Figure 4.25**. All positions flagged via single metric analyses (FDR corrected P-values) as potential sites of interest across the 4 rRNA species examined (pre-rRNA, 18S, 5.8S and 28S) are presented in **Figure 4.25A**, with colored positions identified as being differentially modified above the set level of significance. Results from this comparison suggest the greatest number of flagged positions occur within the 18S and 28S, with comparatively less positions flagged within pre- and 5.8S rRNA species. For first pass analyses, pre- and 5.8S rRNA species were excluded from further consideration. **Figure 4.25B-C (left panels)** display the positions flagged from single metric analyses for each mature rRNA transcript (18S and 28S), for which analysis identified 9 and 10 positions above the threshold of significance ($-\log_{10}(FDR) > 5$) respectively. Shark fin plots presented in **Figure 4.25B-C (right panels)** present the analyses from double metric analyses, with FDR corrected P-values plotted against the absolute value of the Nanocompore logistic regression log odds ratio (GMM logit method), with positions of greatest significance considered to fall furthest from the shark fin data cluster. In comparison to MEF vs MESC, a comparison of MEF and EB data sets yielded substantially fewer positions of potential interest, though significant positions remain concentrated at the 3' ends of the mature 18S and 28S rRNA transcripts, as well as the 5' end of pre-rRNA 5'ETS. Flagged positions were compared to known sites in human rRNA and were designated as potential sites of differential modification if human rRNA modification occurred within +2 bp of the flagged site. A comparison of MEF and EB data sets yielded 3 potential sites of differential modification within the 18S, identified as positions 1) U-1441, possibly corresponding to U-1442 in human 18S, a substrate for methylation (Um), 2) A-1850, possibly corresponding to A-1850 in human 18S, a substrate for di-methylation (M^6_2A), and 3) U-1174, possibly corresponding to U-1174 in human 18S,

a substrate for pseudouridylation (Ψ). Within the 28S transcript, 3 potential sites were identified: 1) U-4323, possibly corresponding to U-4323 in human 28S, a substrate for pseudouridylation (Ψ), 2) G-1509, possibly corresponding to G-1509 in human 28S, a substrate for methylation (Gm) and 3) U-4501, corresponding to U-4502 in human, a substrate for pseudouridylation (Ψ). These potential sites of differential modification between MEFs and EBs are displayed in **Figure 4.25D**.

Finally, a comparison of MinION-sequenced MESC and liver data sets (3:1) was conducted, based on an overall comparison of 183,828 and 194,429 reads, respectively, the results of which are presented in **Figure 4.26**. All positions flagged via single metric analyses (FDR corrected P-values) as potential sites of interest across the 4 rRNA species examined (pre-, 18S, 5.8S and 28S) are presented in **Figure 4.26A**, with colored positions identified as being differentially modified above the set level of significance. Results from this comparison suggest the greatest number of flagged positions occur within the 18S and 28S, with comparatively less positions flagged within pre- and 5.8S rRNA species. For first pass analyses, pre- and 5.8S rRNA species were excluded from further consideration. **Figure 4.26B-C (left panels)** display the positions flagged from single metric analyses for each mature rRNA transcript (18S and 28S), for which analysis identified 798 and 1052 positions respectively, which were above the threshold of significance ($-\log_{10}(\text{FDR}) > 5$). Shark fin plots presented in **Figure 4.26B-C (right panels)** present the analyses from double metric analyses, with FDR corrected P-values plotted against the absolute value of the Nanopore logistic regression log odds ratio (GMM logit method), with positions of greatest interest considered to fall furthest from the shark fin data cluster. Flagged positions were compared to known sites in human rRNA and designated as potential sites of differential modification if human rRNA modification occurred within +2 bp of the flagged site. This comparison yielded 2 potential sites of differential modification within the 18S, identified as positions 1) U-1441, possibly corresponding to U-1442 in human 18S, a substrate for methylation (Um), and 2) G-867, possibly corresponding to G-867 in human 18S, a substrate for methylation (Gm). Within the 28S transcript, 3 potential sites were identified: 1) G-1509, possibly corresponding to G-1509 in human 28S, a substrate for methylation (Gm), 2) A-3808, possibly corresponding to A-3809 in human 28S, a substrate for methylation (Am), and 3) U-4524, possibly corresponding to U-4522 in human 28S, a substrate for pseudouridylation (Ψ). These potential sites of differential modification between MESC and EBs are displayed in **Figure 4.26D**.

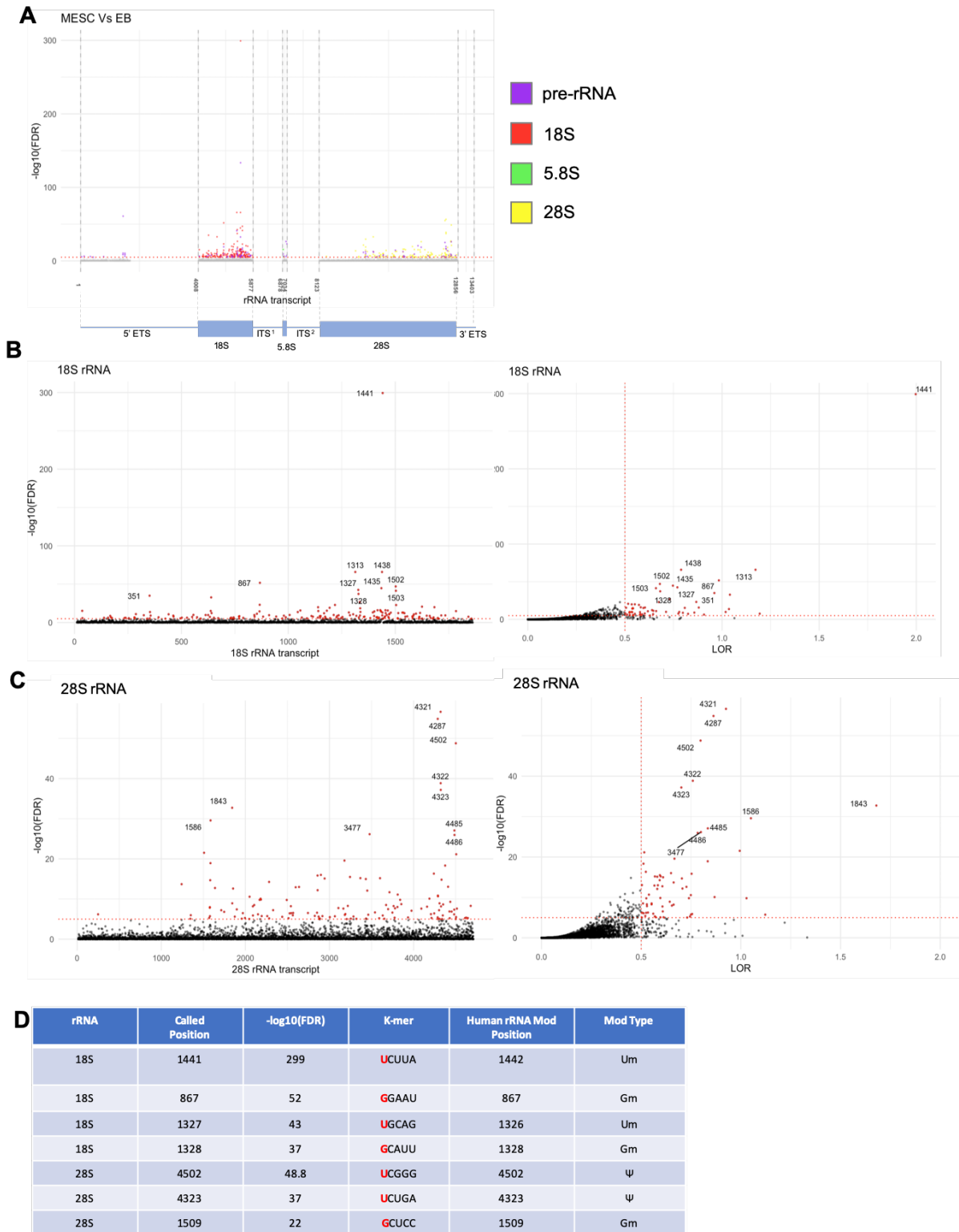


Figure 4.23 Nanopore differential modification-calling in MESC and EB DRS datasets (A) Called positions from 4:2 MESC vs EB data sets, flagged position within the pre-rRNA transcript (x-axis) plotted against the $-\log_{10}(\text{FDR})$. Coloured points represent significant k-mers ($-\log_{10}(\text{FDR}) > 5$). Plot is aligned to a schematic of pre-rRNA transcript. Terminal positions indicate start and end of the 47S rRNA, with secondary positions indicating starting and end-positions of rRNA coding subunits 18S, 5.8S and 28S. (B) (left) Positions flagged across the length of the mature 18S rRNA, with significant positions ($-\log_{10}(\text{FDR}) > 5$) in red. (right) Shark fin plot for positions across the mature 18S rRNA showing the absolute value of the Nanopore logistic regression log odds ratio (GMM logit method) (x-axis) plotted against its $-\log_{10}(\text{FDR})$ value (y-axis) for each flagged k-mer. Significant positions ($-\log_{10}(\text{FDR}) > 5$ & $\text{LOR} > 0.5$) are indicated in red, with top 10 k-mers labelled with their corresponding positions. (C) (left) Positions flagged across the length of the mature 28S rRNA, with significant positions ($-\log_{10}(\text{FDR}) > 5$) in red. (right) Shark fin plot for positions across the mature 28S rRNA showing the absolute value of the Nanopore logistic regression log odds ratio (GMM logit method) (x-axis) plotted against its $-\log_{10}(\text{FDR})$ value (y-axis) for each flagged k-mer. Significant positions ($-\log_{10}(\text{FDR}) > 5$) are indicated in red, with top 10 k-mers labelled with their corresponding positions. (D) Sites of differential modification and potential equivalents in human rRNA, considering only top 10 significant positions.

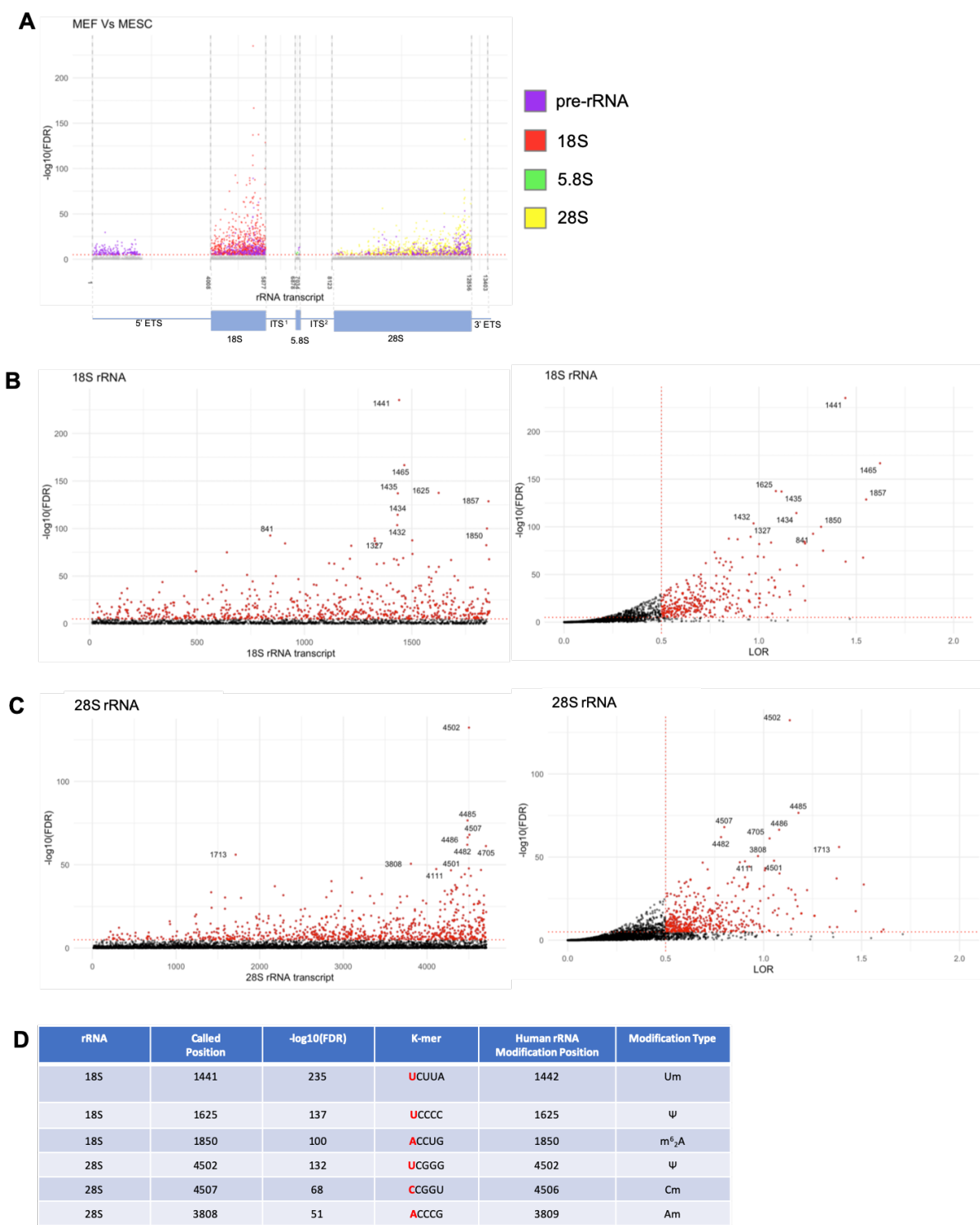
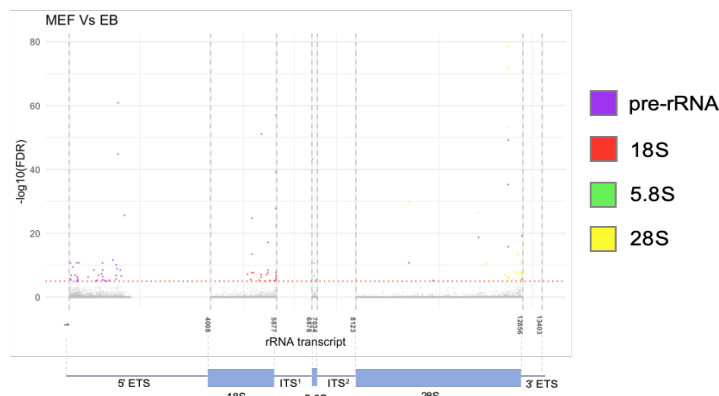
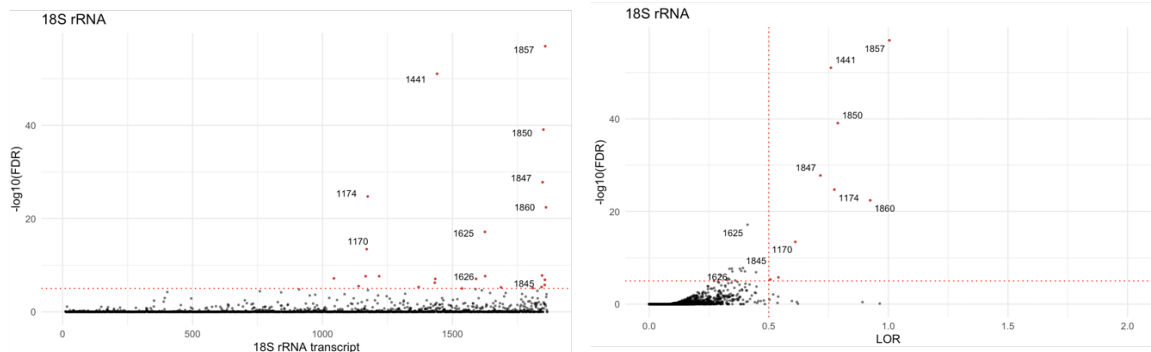


Figure 4.24 Nanopore differential modification-calling in MEF and MESC DRS datasets (A) Called positions from 2:4 MEF vs MESC data sets, flagged position within the pre-rRNA transcript (x-axis) plotted against the $-\log_{10}(\text{FDR})$. Coloured points represent significant k-mers ($-\log_{10}(\text{FDR}) > 5$). Plot is aligned to a schematic of pre-rRNA transcript. Terminal positions indicate start and end of the 47S rRNA, with secondary positions indicating starting and end-positions of rRNA coding subunits 18S, 5.8S and 28S. **(B)** (left) Positions flagged across the length of the mature 18S rRNA, with significant positions ($-\log_{10}(\text{FDR}) > 5$) in red. (right) Shark fin plot for positions across the mature 18S rRNA showing the absolute value of the Nanopore logistic regression log odds ratio (GMM logit method) (x-axis) plotted against its $-\log_{10}(\text{FDR})$ value (y-axis) for each flagged k-mer. Significant positions ($-\log_{10}(\text{FDR}) > 5$ & $\text{LOR} > 0.5$) are indicated in red, with top 10 k-mers labelled with their corresponding positions. **(C)** (left) Positions flagged across the length of the mature 28S rRNA, with significant positions ($-\log_{10}(\text{FDR}) > 5$) in red. (right) Shark fin plot for positions across the mature 28S rRNA showing the absolute value of the Nanopore logistic regression log odds ratio (GMM logit method) (x-axis) plotted against its $-\log_{10}(\text{FDR})$ value (y-axis) for each flagged k-mer. Significant positions ($-\log_{10}(\text{FDR}) > 5$) are indicated in red, with top 10 k-mers labelled with their corresponding positions. **(D)** Sites of differential modification and potential equivalents in human rRNA, considering only top 10 significant positions.

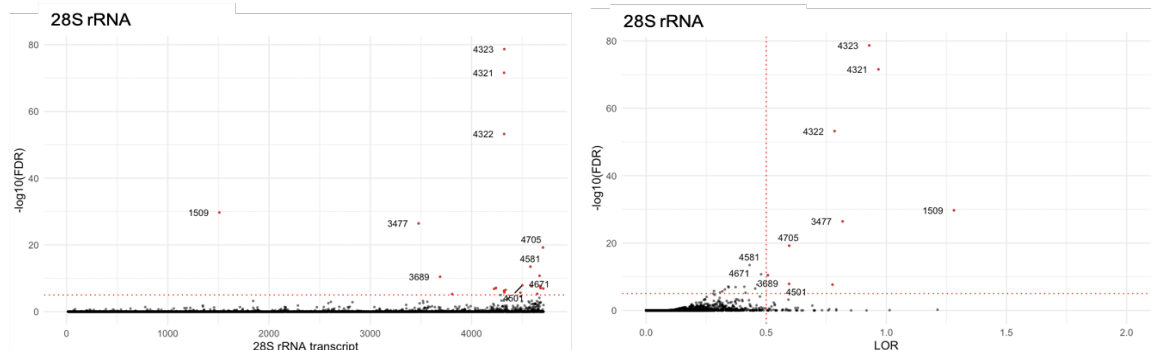
A



B



C



D

rRNA	Called Position	$-\log_{10}(\text{FDR})$	K-mer	Human rRNA Mod Position	Mod Type
18S	1441	52	UCUUA	1442	Um
18S	1850	39	ACCUG	1850	m ⁶ A
18S	1174	25	UGGUU	1174	ψ
28S	4323	77	UCUGA	4323	ψ
28S	1509	30	GCUCG	1509	Gm
28S	4501	8	UCGGG	4502	ψ

Figure 4.25 Nanopore differential modification calling in MEF and EB DRS datasets (A) Called positions from 2:2 MEF vs EB data sets, flagged position within the pre-rRNA transcript (x-axis) plotted against the $-\log_{10}(\text{FDR})$. Coloured points represent significant k-mers ($-\log_{10}(\text{FDR}) > 5$). Plot is aligned to a schematic of pre-rRNA transcript. Terminal positions indicate start and end of the 47S rRNA, with secondary positions indicating starting and end-positions of rRNA coding subunits 18S, 5.8S and 28S. **(B)** (left) Positions flagged across the length of the mature 18S rRNA, with significant positions ($-\log_{10}(\text{FDR}) > 5$) in red. (right) Shark fin plot for positions across the mature 18S rRNA showing the absolute value of the Nanopore logistic regression log odds ratio (GMM logit method) (x-axis) plotted against its $-\log_{10}(\text{FDR})$ value (y-axis) for each flagged k-mer. Significant positions ($-\log_{10}(\text{FDR}) > 5$ & $\text{LOR} > 0.5$) are indicated in red, with top 10 k-mers labelled with their corresponding positions. **(C)** (left) Positions flagged across the length of the mature 28S rRNA, with significant positions ($-\log_{10}(\text{FDR}) > 5$) in red. (right) Shark fin plot for positions across the mature 28S rRNA showing the absolute value of the Nanopore logistic regression log odds ratio (GMM logit method) (x-axis) plotted against its $-\log_{10}(\text{FDR})$ value (y-axis) for each flagged k-mer. Significant positions ($-\log_{10}(\text{FDR}) > 5$) are indicated in red, with top 10 k-mers labelled with their corresponding positions. **(D)** Sites of differential modification and potential equivalents in human rRNA, considering only top 10 significant positions.

Long-Read Sequencing Analysis of Ribosomal RNA Modifications

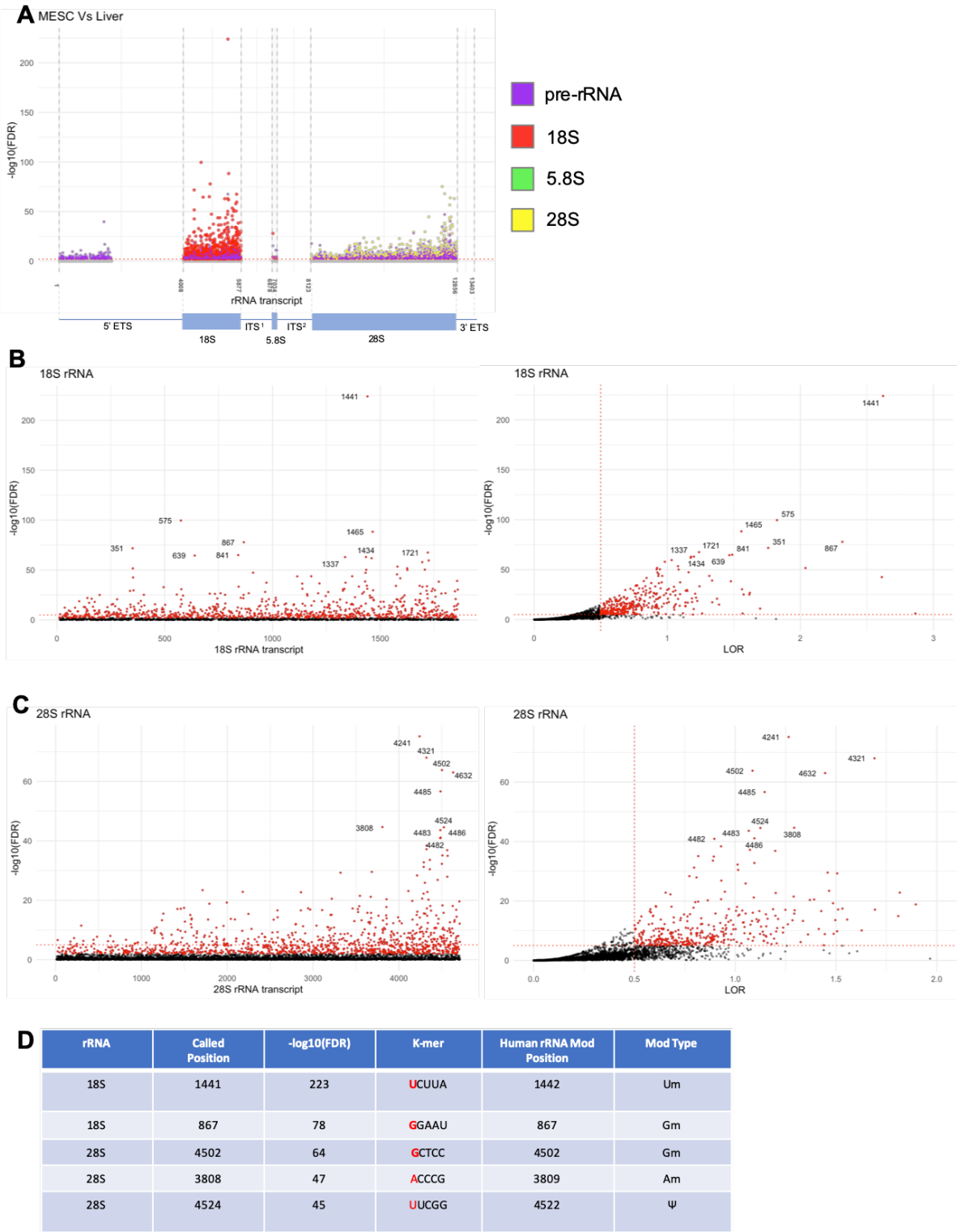


Figure 4.26 Nanocompare differential modification calling in MESC and liver DRS datasets (A) Called positions from 2:2 MEF vs EB data sets, flagged position within the pre-rRNA transcript (x-axis) plotted against the $-\log_{10}(\text{FDR})$. Coloured points represent significant k-mers ($-\log_{10}(\text{FDR}) > 5$). Plot is aligned to a schematic of pre-rRNA transcript. Terminal positions indicate start and end of the 47S rRNA, with secondary positions indicating starting and end-positions of rRNA coding subunits 18S, 5.8S and 28S. (B) (left) Positions flagged across the length of the mature 18S rRNA, with significant positions ($-\log_{10}(\text{FDR}) > 5$) in red. (right) Shark fin plot for positions across the mature 18S rRNA showing the absolute value of the Nanocompare logistic regression log odds ratio (GMM logit method) (x-axis) plotted against its $-\log_{10}(\text{FDR})$ value (y-axis) for each flagged k-mer. Significant positions ($-\log_{10}(\text{FDR}) > 5$ & $\text{LOR} > 0.5$) are indicated in red, with top 10 k-mers labelled with their corresponding positions. (C) (left) Positions flagged across the length of the mature 28S rRNA, with significant positions ($-\log_{10}(\text{FDR}) > 5$) in red. (right) Shark fin plot for positions across the mature 28S rRNA showing the absolute value of the Nanocompare logistic regression log odds ratio (GMM logit method) (x-axis) plotted against its $-\log_{10}(\text{FDR})$ value (y-axis) for each flagged k-mer. Significant positions ($-\log_{10}(\text{FDR}) > 5$) are indicated in red, with top 10 k-mers labelled with their corresponding positions. (D) Sites of differential modification and potential equivalents in human rRNA, considering only top 10 significant positions.

4.4 Discussion

4.4.1 Nanopore RNA sequencing of ribosomal RNA

The development of a sample pre-processing protocol has proven to routinely increase the capture of pre-rRNA transcripts, otherwise severely underrepresented in ‘standard’ sequencing sample preparations. The improvements outlined here have served to specifically increase the yield of reads spanning across transcribed spacers, with a ~7-fold increase in the complete coverage of both ITS¹ and ITS² elements. This allowed for the capture of pre-rRNA molecules spanning across all 3 coding subunits in both MEF and MESC DRS data sets. However, even with the extensive protocol development, pre-rRNA reads made up a minority of total rRNA reads, with the majority of reads still mapping to mature rRNAs. The considerable cost of nanopore sequencing necessitates the consideration of additional methods to maximise yields of useful pre-rRNA transcripts. One way of achieving this may be to employ sequence capture methods to isolate specific transcripts from total RNA preparations. A study by Smith *et al.* (2019) demonstrated the effective isolation of 16S rRNA from crude cell lysates in preparation for nanopore direct RNA sequencing. By exploiting the biotin-streptavidin interaction, hybridisation of sequence-specific biotinylated DNA probes to 16S rRNA transcripts, followed by separation via streptavidin-conjugated magnetic bead isolation, allowed for a 5-fold increase in 16S mapped reads compared to libraries without bead enrichment (Smith *et al.*, 2019). Sequence capture methods have been shown to assist in the targeted sequencing of specific loci for high throughput NGS (Albert *et al.*, 2007; Gnirke *et al.*, 2009). These methods rely on the subsequent PCR amplification of the captured sequence, which acts to off-sets the often, low target yields (Anderman *et al.*, 2020). This principle could, in theory, be applied to supplement the capture of pre-rRNA transcripts by firstly targeting transcribed spacer elements with intact cleavage sites, followed by a secondary enrichment targeting coding subunit sequences to ensure the capture of only pre-rRNA. However, considering the varying efficiency of such approaches, coupled with the scarcity of pre-rRNA molecules, and the inability to PCR amplify targets, this may serve to contribute slightly, rather than revolutionise pre-rRNA capture.

Though, rRNA makes up approximately 80% of the total cellular RNA (Warner, 1999), processing of the pre-rRNA transcript occurs swiftly, with studies indicating processing can begin co-transcriptionally, with the removal of a 650 bp terminal sequence of the 5’ETS (Braglia, Kawauchi and Proudfoot, 2011). Studies suggest that the half-life of murine rRNA primary transcript is approximately 1-2 minutes, with secondary cleavage events occurring rapidly after this, leading to the exponential decay of the precursor molecule (Lazdins, Delannoy and Sollner-Webb, 1997; Popov *et al.*, 2013).

Considering this, sequence-specific enrichment may still not yield enough full-length transcripts for nanopore sequencing, and so the inhibition of the rRNA processing pathway may be a critical consideration for future studies. In this study, two compounds, Flavopiridol, and 5-FU were evaluated. These chemotherapeutic compounds are shown to inhibit the early and late stages of rRNA processing respectively. Cell cycle progression, used as a determinant of drug effect, indicated no effect of flavopiridol in any cell line within the concentrations tested. However, a considerable increase in S-phase arrest was observed in MEFs treated with 100 μ M 5-FU. Additionally, the treatment of MEFs with 25 μ M 5-FU yielded favourable sequencing results, increasing the abundance of pre-rRNA. This was specifically indicated by an increase in reads spanning ITS¹, with a \sim 2.5-fold increase in coverage. As 5-FU is a pyrimidine analogue, it is readily incorporated into actively transcribed RNA in the place of uracil. Its incorporation is thought to inhibit rRNA processing by reducing the susceptibility of rRNA to the ribonucleolytic activity of the Exosome subunit *Rrp6* (Silverstein, De Valdivia and Visa, 2011). However, the incorporation of 5-FU fundamentally alters the chemical properties of the transcript, ultimately affecting how RNA molecules interact with sequencing nanopores (Xu et al., 2020). Nanopore base-calling algorithms do not currently possess the ability to distinguish 5-FU, and its incorporation is likely to lead to transcript-wide base-calling errors, hindering the accurate deduction of the nucleotide sequence as well as RNA modifications detection (Amarasinghe et al., 2020; Xue et al., 2020). Due to the potential for disruption of downstream analysis, its use in this study was discontinued. A study by Burger et.al, (2010), has explored the impact of a range of non-nucleotide analogues on rRNA processing inhibition, which function via alternative pathways to that of 5-FU (Burger et al., 2010). Roscovitine, a broad range cyclin dependant kinase (CDK) inhibitor (Meijer and Raymond, 2003), was shown to almost entirely abolish 32S and mature transcript generation, without impacting total primary transcript levels (Burger et al., 2010). Considering the wide scope of CDK action (Bach, Blondel and Meijer, 2006) it is unclear how treatments affecting such large-scale cellular processes will impact the epitranscriptomic profile of rRNA and should be a key consideration for any future work. Overall, the study of full-length pre-RNA transcripts may be invaluable in discerning the epitranscriptome profiles of specific rRNA alleles, shedding light on the biological relevance of specific alleles but also the role of RNA modifications in specific environmental contexts.

Alternatively, sequencing efforts could simply be increased. The output for the majority of the sequencing runs reported in this chapter is not representative of the expected yields from comparable sequencing kits and flow cells. This is due to many of the sequencing datasets, for both MinION and PromethION devices, having been generated on flow cells which had been used previously, and were therefore of suboptimal capacity. ONT DRS sequencing on a single MinION flow cell is expected to

yield, on average, 1 Million full-length reads per flow cell. Considering the results from this study, just under 300,000 rRNA reads from MESC data sets, yielded 77 pre-rRNA reads completely spanning all 3 coding subunits. If using a flow cell of maximum capacity, with additional enrichment methods, this number can be expected to rise considerably. This would no doubt, provide greater coverage of key transcripts and allow for much deeper analyses. Even though there is considerable room for improvement, the length of rRNA reads captured in this study is longer than any reported in the literature for Nanopore DRS data sets (Smith *et al.*, 2019; Jain *et al.*, 2021; Stephenson *et al.*, 2022). The protocol development outlined here, in conjunction with additional enrichment methods, and increased sequencing may serve to bolster the capture of pre-rRNA transcripts spanning across all three coding subunits. This will be critical in determining, for the first time, rRNA modification profiles across the length of all three coding subunits within a single molecule.

4.4.2 Cell-specific expression of ribosomal RNA alleles

Nanopore DRS datasets from three developmentally distinct cell populations, MEFs, MESCs and EBs were assessed to determine rRNA allele expression within a developmental context. By considering the occurrence of specific SNPs at positions 6777 and 12736, the expression of the ATA and ATG haplotypes was determined. In MEFs, the mean expression of the ATA haplotype defining SNP was considerably lower than that of the ATG haplotype defining SNP, quantified as 4% and 28% respectively. The expression of specific rRNA haplotype defining SNPs in MEFs, reported here, agrees with a previous study exploring the epigenetic profile of specific rRNA variants. Specifically, the reduced expression of the ATA haplotype in MEFs correlates with finding by Algarra *et al.*, (2022) which demonstrated differential DNA methylation of the ATA and ATG haplotypes (Rodriguez-Algarra *et al.*, 2022). The study showed, that the MEF ATA haplotype displays significant methylation across the length of the rDNA coding unit ($\geq 60\%$), whilst the ATG haplotype remains largely unmethylated. Additionally, methylation patterns across both haplotypes were seen to negatively correlate with the expression of rRNA alleles in both, C57BL/6J kidney and muscle tissue. In this study, the expression of ATA and ATG haplotype defining SNPs was markedly less unbalanced in MESC datasets, with the mean expression of haplotype defining SNP at each position being within 2% of one another. Proportional expression of both haplotypes may indicate similar levels of epigenetic markers such as DNA methylation across both haplotypes and may be reflective of the cell population's pluripotent state. MESC data sets also exhibited the greatest degree of variation when considering any given SNP. As each library was prepared from a separate culture of cells, fluctuations in cell culture conditions may have contributed to the varying levels of haplotype expression. MESCs are particularly susceptible to changes in the culture environments, with slight reductions in the potency of differentiation inhibitors

resulting in rapid differentiation (Tamm, Galitó and Annerén, 2013). The differentiation process is considered to correlate with an increase in gene methylation, perhaps leading to the observed outcomes (Huang *et al.*, 2015). Though cell populations were screened to ensure pluripotency after cultures were established, it cannot be stated with certainty that cell culture conditions were unequivocally identical at all points of RNA collection. To prevent this, improvements in cell culturing stringency may prove beneficial.

In these analyses, haplotype-specific positions 6832 and 6777 were not thoroughly considered due to the reduced coverage observed. These positions occur within ITS¹, the spacer element between 18S and 5.8S rRNA (Michot *et al.*, 1989). Due to the intense endo- and ex-nucleolytic activity occurring within ITS¹ during pre-rRNA processing (Preti *et al.*, 2013), levels of this region are greatly reduced in comparison to sequences corresponding to coding subunits (e.g., 12736). Additionally, although positioned in relative proximity (<100 bp), these positions are not equally represented, likely owing to their proximity to nuclease target sites (Wang, Anikin and Pestov, 2014). In future experiments, enrichment of reads spanning ITS¹ may facilitate the study of these positions and allow for a more complete assessment of rRNA haplotype expression.

4.4.3 The cell-specific differential modification of rRNA

The comparative rRNA epitranscriptome analyses presented in this chapter has identified potential sites of differential modification across sample representative of varying development stages, from embryonic stem cells to organ tissue. The position identified in these analyses are by no means exhaustive, only having considered the top 10 positions of greatest significance. Even so, the positions identified may represent potential candidates for differential RNA modifications between the sample types compared. Position A-1850-18S is one site which may be of particular interest and the focus of future work. This site presents in multiple comparisons conducted. Specifically, its potential modification levels are seen to be significantly different when comparing MEFs to both EBs and MESCs. Comparable to site A-1850 in human 18s rRNA (Yang *et al.*, 2016), it is a known m⁶₂A modification substrate and lies adjacent to another m⁶₂A site at position 1851 (Natchiar *et al.*, 2017). Both these positions occur within a critical functional domain of the ribosome, the decoding centre (Zorbas *et al.*, 2015), a region vital for ensuring the fidelity of the codon-anticodon interaction, along with mRNA translation and translocation (Sergiev *et al.*, 1998). Modifications within this functional domain are highly conserved and are considered to be required for ribosome assembly (Zorbas *et al.*, 2015). In yeast the equivalent positions are A1781 and A1782 (m6A) (Conrad *et al.*, 1998), and are shown to be

in direct contact with the ribosomal protein eL41, forming a bridge between the large subunit to the decoding centre in the small subunit. This interaction is thought to mediate long-range structural information between the subunits, with eL41 acting as a pivot for small subunit rotation during the translation process (Ben-Shem *et al.*, 2011). It has been suggested that this may be a mechanism by which these conserved RNA modifications directly couple with ribosomal proteins to impact translation efficiency (Sharma and Lafontaine, 2015). Considering the highly conserved nature of modification at this site across evolution (Zorbas *et al.*, 2015) it may be of particular interest to explore the potential reasons for the observed differential modification levels, within a cell-specific and developmental context.

Position U-4502-28S also repeatedly occurs as a potential site of differential modification in the sample types compared in this study. Due to how nanopore sequence signals are detected (on a k-mer level, rather than a single nucleotide) (Wang, Yang and Wang, 2014), this position may be equivalent to 4500 or 4502 in human 28S. These are known sites for the trimethylation of uracil or its conversion into pseudouridine respectively (Hughes and Maden, 1978; Ofengand, 2002; Taoka *et al.*, 2018). Both these positions occur very close to, or within the peptidyl transfer centre (PTC) on helix H92, which during translation is positioned close to the 3'-CCA binding region of an A-site tRNA (Cheng *et al.*, 2017). Differential modification at this site may cause alterations to tRNA binding and as a result, impact the efficiency of protein translation. Site 4500-28S is considered to be equivalent to 2923-25S in yeast which is similarly positioned within the PTC, and occurs amongst a string of conserved pseudouridines (Henras *et al.*, 2015). Diminished pseudouridylation at the PTC has been shown to impact, yeast growth, and specifically cause deficiencies in tRNA binding and protein translational whilst altering ribosome structure (King *et al.*, 2003).

Interestingly, position U-1441-18S, consistently presents as the most significantly different position with a known human rRNA modification analogue, across all comparisons. The equivalent position in human 18S rRNA is likely to be U-1442, which is a known substrate for the methylation of uracil. Though the literature does not provide any evidence for this specific site as one that is crucial for, or directly involved in ribosome function, it does occur within the 3' major domain of the 18S rRNA. X-ray crystallographic structure investigations have shown that helices in this domain, interact with the 18S 5' major- and central domains to form the 'mRNA binding tunnel' (Cheng *et al.*, 2017) a region encompassing both mRNA and tRNA recognition sites. RNA modifications have been shown to exert both local and distal effects, either directly impacting the immediate environment by altering the nucleotide's binding capacity, or by causing indirect changes to ribosome conformation through interactions with ribosomal proteins with changes expressing at distance to the site of modification

(Sharma *et al.*, 2015; Ben-Shem *et al.*, 2011). This may be a means by which differences in 1441-18S modification, impact function within the mRNA binding tunnel, possibly translating to changes in the ribosomal translational efficiency.

The level of analysis conducted here does not permit for the deduction of the condition-position relationship, i.e., it is not possible to deduce what cell line a certain position is more or less modified in. Additionally, time did not permit the study of RNA modifications at the single-molecule level and deduction of haplotype-specific profiles. For future experiments, the use of appropriate control RNA will be necessary to determine cell-specific positions of interest. This may be achieved simply through the use of *in vitro* transcribed rRNA as recommended for Nanocompore analyses (Leger *et al.*, 2021). Due to its synthetic nature, we can be confident that *in vitro* transcribed RNA will be devoid of all modifications that naturally occurring rRNA would accumulate during maturation. By utilizing this as an effective 'control' sample, a baseline of 'un-modified' RNA state nanopore signal signature could be established. This control data set could then simply be compared against relevant 'test' samples abundant in biologically significant RNA modifications using Nanocompore comparison software. As a result, sites of modification could confidently be called and quantified in each biological sample independently after which, comparison of biologically test sample with one another could possibly yield insights into the differential levels of specific modifications between samples. Additionally, it may be possible to compare samples of interest to simulated data sets, which utilise computationally generated k-mer current signals for a defined sequence, as demonstrated by Leger *et al.*, 2021. However, due to the great computational burden of this approach, coupled with the exceeding difficulty of accurately recreating k-mer-specific current signal noise *in silico*, this is not considered an ideal approach. Though limited in scope, the analyses from these data sets provide information on potential sites within the 18S and 28S rRNA which may be differentially modified within a cell-specific context, providing a platform for further study in which the biological relevance of these RNA modifications can be determined.

4.5 Conclusions

To conclude, the ONT DRS pre-sequencing protocol development outlined in this chapter has led to the capture of pre-rRNA molecules longer than any reported in published studies. Additionally, the sites of differential modification predicted here may serve to shed light on cell-specific functions of certain RNA modifications, opening up avenues to pursue further functional characterisation of these modifications in different cell types. Further optimisation through sequence-specific capture methods or drug-induced inhibition of cellular pathways may likely lead to the further enrichment of pre-rRNA in sequencing libraries. Considering the scarcity of short-lived pre-rRNAs, this may be a necessary

Long-Read Sequencing Analysis of Ribosomal RNA Modifications
consideration for future work. Nevertheless, the work outlined here may provide a foundation from which to deduce RNA modification profiles across multiple coding subunits within a single molecule. Additionally, it may serve as a framework to further explore haplotype-specific profiles across cell types and different environmental contexts.

5 Discussion and Conclusions

5.1 Research summaries

This thesis had 2 inter-related aims, which it addressed over two research chapters. An In-depth discussion of the research is provided in the discussion sections within each chapter. Here, I will discuss and conclude the main findings of each chapter with respect to the outlined aims. This will be accompanied by a consideration of issues faced during the pursuit of achieving the research goals, as well as an outline of future experiments which may be conducted to build on the work undertaken in this study.

5.1.1 Aim 1

To establish a methodology in which molecular combing could be used in conjunction with SNP-specific probes to

- 1. Capture entire rDNA clusters at the single-molecule level**
- 2. Characterise the large-scale arrangement of rDNA SNP alleles within an entire cluster.**
- 3. To explore the epigenetic profiles of entire rDNA clusters in a SNP specific context**

The work in chapter 3 partially fulfilled the first aim of this thesis. The molecular combing methodology was established to obtain individual DNA molecules spanning ~ 6 Mbp. Considering the size of a single rDNA unit, molecules of this length should permit the direct visualisation of rDNA clusters containing upwards of 100 rDNA repeats. The protocol outlined in Kaykov et al., (2016), reported fibres up to 12 Mbp in length, with an average length of ~ 2 Mb, and was used as a starting point for establishing the method. This methodology had not been previously applied by our lab, nor was it a routine procedure for any other research team at our institute, and I was the first to establish and optimise its use. Therefore, considerable optimisation was necessary to establish a working protocol. Efforts were made to reach out to Dr A. Kaykov and other experts in the field, however, these discussions, unfortunately, did not amount to much success. The numerous technical challenges this method presents with, especially when pursuing such long molecules, compounded with the lack of expertise to draw from, meant that establishing and optimising the methodology took a considerable amount of time. Even so, considerably long molecules were captured, with the mean length of fibres being 2-3 Mbp. Initial attempts were made at probing combed DNA with DNA FISH probes, to visualise and

evaluate the possible capture of entire clusters, however, due to the lack of success with fibre-FISH, this remains to be confirmed. Even so, the length assessment of combed DNA fibres, coupled with the fact that rDNA clusters exhibit large deviations in size, may suggest, that fibres captured in this study are of adequate length to partially, if not completely span entire rDNA clusters.

Furthermore, chapter 3 outlines the progress made towards characterising the large-scale genomic architecture of rDNA clusters, specifically, the arrangement of rDNA promoter variants. In particular, progress was made towards generating dCas9 SNP-specific probes targeting rDNA promoter variants, along with the generation of control cell lines in which to test the allele targeting capabilities and SNP specificity of probes. The aim was to probe combed DNA with SNP-specific probes and characterise the arrangement of rDNA promoter alleles within an entire rDNA cluster. However, due to the unforeseen changes in global circumstances, this aspect of the project was put on pause and remains incomplete. Additionally, the epigenetic assessment of rDNA clusters in an allele-specific context, remains unevaluated, since the completion of this goal was dependent on the fulfilment of the aforementioned goals. Considering the promising progress made, it is unfortunate to have stopped so prematurely in the pursuit of these aims. Regardless, the work conducted may provide the foundation from which to explore the genetic and epigenetic landscape of rDNA at the cluster level, gaining insights into the large-scale arrangement of rDNA like never before.

5.1.2 Aim 2

To establish the Nanopore DRS methodology to

- 1. Sequence ribosomal RNA with particular focus on sequencing full-length pre-rRNA primary transcript molecules**
- 2. Evaluate rRNA modifications profiles within a developmental context**
- 3. Dissect rRNA allele-specific modification profiles at the single-molecule level**

The work outlined in chapter 4 partially fulfilled these aims. The Nanopore DRS methodology was successfully applied to sequencing rRNA, however, from initial attempts it was clear that the rRNA primary transcript was especially elusive. Not only this but even pre-rRNA processing intermediates (transcripts mapping to more than one coding unit), were captured in exceedingly small quantities. Extensive development of the pre-sequencing sample preparation protocol allowed for a significant increase in the capture of pre-rRNA molecules, with a nearly 7-fold increase observed. However, even with this, pre-rRNA reads made up <1% of total rRNA reads, necessitating the need for further enrichment of these short-lived transcripts in future experiments. Even so, analysis of MEF and MESCC data sets revealed the capture of near full-length primary transcripts, with many more reads mapping

to all three coding subunits. It is important to note that, the capture of full-length primary transcripts is not critical for the assessment of RNA modifications profiles across multiple coding subunit sequences, as reads spanning across the three coding units are likely to suffice. The rRNA read lengths reported in this study exceed any in the published literature (Smith *et al.*, 2019; Jain *et al.*, 2021; Stephenson *et al.*, 2022), and may provide novel insights into rRNA modification dynamics at scales beyond just those of mature transcripts.

Additionally, several DRS data sets were generated for the assessment of RNA modification differences within a developmental context. The work discussed in chapter 4, outlines the use of Nanocompore, in uncovering many differentially modified bases between cell-specific data sets. Several sites predicted to be differentially modified, are known sites of modification in both human and yeast rRNA. Additionally, some of the predicted sites are key regulators of ribosome function, specifically involved in critical processes such as mRNA and tRNA recognition and binding. Considering the predicated differential modification observed between cell types studies here, it will be of value to further assess the role of these specific modifications in a developmental context. Unfortunately, due to the extensive time required for generating data sets with sufficient coverage of pre-rRNA, a thorough evaluation of cell-specific modification profiles remains incomplete. Specifically, it could not be determined, to which cell types each predicated position belonged, only that these sites were differentially modified between conditions. This was due to the lack of an appropriate comparison control. In future work, will be necessary to first, determine the relationship between predicated positions and the cell types tested, by comparing to a data set devoid of RNA modifications. Next, to then evaluate the degree to which each predicted position is modified in the specific cell types tested. Also, it is important to note that sites of differential modification were predicted from ensembles of rRNA transcripts and not via the comparison of individual molecules. Within the permitted time frame, it was not possible to begin to evaluate rRNA modification profiles at the single-molecule level. This will be necessary for future work, to thoroughly examine rRNA profiles across multiple coding subunits within an individual transcript, as well as assess any differential modification between different rRNA alleles.

5.2 Research Challenges

Though the work in this study has demonstrated progress in developing the methodology needed to fulfil the outlined research aims, a number of goals have gone unachieved, and some key research questions remain unanswered. Research progress was significantly hindered by the COVID-19 pandemic. This was expressed as any number of obstacles, from complete lockdown to restricted lab

access and a severe hindrance to the supply of key reagents, consumables and samples for study. Planned experiments were abandoned, whilst those already underway were put on hold, due to the inaccessibility of resources, and facilities as well as the overwhelming uncertainty about future circumstances. Regarding the challenges faced during specific experimental procedures, there may be potential issues arising from the use of directly labelled dCas9 probes to visualise genomic loci. Direct labelling of probes with a string of 3 fluorophores (GFP or RFP), as intended in this study, could make effective visualisation difficult, owing to the limitations of available microscopy tools. On a combed DNA substrate where target molecules are stretched at great linear distances, it may be necessary to use additional signal amplification methods such as FRACTAL, from which fluorescent signals can be amplified through successive rounds of labelling (Cho *et al.*, 2020). Concerning the capture of pre-rRNA molecules, only 2 compounds were tested for their ability to inhibit rRNA processing, within a small concentration range. A large-scale multi-compound analysis could prove useful in identifying compounds which effectively inhibit rRNA processing within a target cell line, and be invaluable in increasing pre-rRNA yields. Additionally, stringent statistical testing of haplotype defining SNP expression in different cell types will no doubt benefit from increased sample sizes. Specifically, this study was unable to determine with statistical significance, the expression levels using only 2 EB DRS data sets. Regarding modification calling, a single computational tool was utilised. Validating these observations with more established methods will allow for a more confident identification of potential sites of differential modification.

5.3 Future experiments and directions

Future experiments could be conducted to achieve the unfulfilled research goals discussed above, and see the application of the methodologies explored, to elucidate the large-scale genetic, and epigenetic architecture of rDNA. Specifically, the capture of entire rDNA clusters with molecular combing, in conjunction with FISH (fibre-FISH) (Ersfeld *et al.*, 2014), could allow for the study of rDNA clusters on a scale never before seen. This methodology has the potential for studying large-scale organisations of rDNA loci, surpassing the capabilities of any currently available, imaging or sequencing-based genetic profiling methods. This combined methodology could be applied to determine the copy number composition of rDNA clusters on specific chromosomes and the variations in different cell types and tissues. Also, molecular combing may be applied to characterising, not only the basal composition of clusters but also copy number expansion and retraction in response to a range of environmental stresses, in a chromosome-specific manner. The SNP-specific probing of combed rDNA clusters could prove essential for thoroughly dissecting rDNA loci with respect to their variant composition, shedding light on the arrangement, relation and interplay of genetic variants, within and

between clusters. Additionally, combining molecular combing with single-molecule methylation analysis (methyl-combing) (Németh *et al.*, 2014), could yield insights into the rDNA epigenetic response at the cluster level, potentially allowing for the identification of environmentally sensitive loci and their response to different environmental cues.

The direct sequencing of ribosomal RNA with Nanopore technology has permitted the capture of individual transcripts spanning across multiple coding subunits providing a substrate with which to assess the RNA modification profiles across all coding subunits at the single-molecule level. Considering the low abundance of these molecules, it will be necessary for future work, to further enrich and increase yields to allow for a deeper analysis. The sites of differential modification predicted in this study will need to be further examined, with particular focus on the position-condition relationship, i.e., what positions are more or less modified in specific sample types. Whilst sites of differential modification have been predicted in this study, analysis of DRS data was only conducted with a single computational tool. Nanocompore, though novel and shown to be robust, is in its infancy, and can only serve to predict sites indirectly from fluctuations in current intensities (Leger *et al.*, 2021). Besides examining Nanocompore predicted sites with other computational tools such as Epinano (Lieu *et al.*, 2017); or CHIUE (Mateos *et al.*, 2022), differential modification at predicted sites, will need to be confirmed with more established methods. Conventional methods of RNA modification detection, such as those based on immunoprecipitation (MeRIP-Seq, i/miCLIP) (Dominissini *et al.*, 2012; Grozhik *et al.*, 2017) or those exploiting chemically induced alteration of RT-Profiles ((Li *et al.*, 2015; Zaringhalam and Papavasiliou, 2016)) in conjunction with deep sequencing may be used to validate Nanocompore predicted sites. Further work could be undertaken to assess the function of specific rRNA modifications across cell types and tissue. Since the large majority of rRNA modifications are induced by site-specific enzymes and snoRNAs (Sloan *et al.*, 2016), systemic inhibition of specific modification-inducing elements may yield insights into the biological significance of the differential modification predicted in this study.

5.4 Conclusion

This thesis has examined two themes: ribosomal DNA, and ribosomal RNA, using long molecule analysis. Chapter 3 outlines the progress made towards establishing a methodology which would potentially facilitate the large-scale study of the rDNA genetic and epigenetic landscape, on the level of entire clusters, at the single-molecule level. Chapter 4, presents the protocol development allowing

for the capture of near full-length rRNA primary transcripts, in addition to identifying sites of differential modification which may be key contributors to cell specificity in a developmental context. The research outlined here lays the foundation for future work with which to thoroughly dissect ribosomal DNA and its dynamic response to environmental cues, whilst providing a means with which to explore the epitranscriptome across the entirety of a single rRNA transcript. Studies aligned with such pursuits will no doubt take our understanding of this largely uncharacterised region of the genome to greater heights.

References

- Agrawal, S. and Ganley, A. R. D. (2018) 'The conservation landscape of the human ribosomal RNA gene repeats', *PLoS ONE*, 13(12). doi: 10.1371/journal.pone.0207531.
- Agris, P. F., Sierzputowska-Gracz, H. and Smith, C. (1986) 'Transfer RNA Contains Sites of Localized Positive Charge: Carbon NMR Studies of [¹³C] Methyl-Enriched *Escherichia coli* and Yeast tRNA^{Phe}', *Biochemistry*, 25(18). doi: 10.1021/bi00366a022.
- Albert, T. J. *et al.* (2007) 'Direct selection of human genomic loci by microarray hybridization', *Nature Methods*, 4(11). doi: 10.1038/nmeth1111.
- Allemand, J. F. *et al.* (1997) 'pH-dependent specific binding and combing of DNA', *Biophysical Journal*, 73(4). doi: 10.1016/S0006-3495(97)78236-5.
- Amarasinghe, S. L. *et al.* (2020) 'Opportunities and challenges in long-read sequencing data analysis', *Genome Biology*. doi: 10.1186/s13059-020-1935-5.
- Anreiter, I. *et al.* (2021) 'New Twists in Detecting mRNA Modification Dynamics', *Trends in Biotechnology*. doi: 10.1016/j.tibtech.2020.06.002.
- Ansel, K. M. *et al.* (2008) 'Mouse Eri1 interacts with the ribosome and catalyzes 5.8S rRNA processing', *Nature Structural and Molecular Biology*, 15(5). doi: 10.1038/nsmb.1417.
- Armistead, J. *et al.* (2009) 'Mutation of a Gene Essential for Ribosome Biogenesis, EMG1, Causes Bowen-Conradi Syndrome', *American Journal of Human Genetics*, 84(6). doi: 10.1016/j.ajhg.2009.04.017.
- Bach, S., Blondel, M. and Meijer, L. (2006) 'Evaluation of CDK inhibitor selectivity: From affinity chromatography to yeast genetics', in *Inhibitors of Cyclin-Dependent Kinases as Anti-Tumor Agents*. doi: 10.1201/9781420005400.
- Baliga, N. S. *et al.* (2004) 'Genome sequence of *Haloarcula marismortui*: A halophilic archaeon from the Dead Sea', *Genome Research*, 14(11). doi: 10.1101/gr.2700304.
- Bansal, V. and Boucher, C. (2019) 'Sequencing Technologies and Analyses: Where Have We Been and Where Are We Going?', *iScience*. doi: 10.1016/j.isci.2019.06.035.
- Barker, D. J. P. *et al.* (1989) 'WEIGHT IN INFANCY AND DEATH FROM ISCHAEMIC HEART DISEASE', *The Lancet*, 334(8663). doi: 10.1016/S0140-6736(89)90710-1.
- Barker, D. J. P. *et al.* (1993) 'Fetal nutrition and cardiovascular disease in adult life', *The Lancet*, 341(8850). doi: 10.1016/0140-6736(93)91224-A.
- Barker, D. J. P. *et al.* (2009) 'Growth and chronic disease: Findings in the Helsinki Birth Cohort', *Annals of Human Biology*, 36(5). doi: 10.1080/03014460902980295.
- Barker, D. J. P. and Osmond, C. (1986) 'Diet and coronary heart disease in England and Wales during

- and after the second world war', *Journal of Epidemiology and Community Health*, 40(1), pp. 37–44. doi: 10.1136/jech.40.1.37.
- Barker, D. J. P. and Thornburg, K. L. (2013) 'The obstetric origins of health for a lifetime', *Clinical Obstetrics and Gynecology*. doi: 10.1097/GRF.0b013e31829cb9ca.
- Baßler, J. and Hurt, E. (2019) 'Eukaryotic Ribosome Assembly', *Annual Review of Biochemistry*, 88. doi: 10.1146/annurev-biochem-013118-110817.
- Basu, A. *et al.* (2011) 'Requirement of rRNA Methylation for 80S Ribosome Assembly on a Cohort of Cellular Internal Ribosome Entry Sites', *Molecular and Cellular Biology*, 31(22). doi: 10.1128/mcb.05804-11.
- Baudin-baillieu, A. *et al.* (2009) 'Nucleotide modifications in three functionally important regions of the *Saccharomyces cerevisiae* ribosome affect translation accuracy', *Nucleic Acids Research*, 37(22). doi: 10.1093/nar/gkp816.
- Baxter-Roshek, J. L., Petrov, A. N. and Dinman, J. D. (2007) 'Optimization of ribosome structure and function by rRNA base modification', *PLoS ONE*, 2(1). doi: 10.1371/journal.pone.0000174.
- Ben-Shem, A. *et al.* (2011) 'The structure of the eukaryotic ribosome at 3.0 Å resolution', *Science*, 334(6062). doi: 10.1126/science.1212642.
- Bensimon, A. *et al.* (1994) 'Alignment and sensitive detection of DNA by a moving interface', *Science*, 265(5181). doi: 10.1126/science.7522347.
- Beringer, M. (2008) 'Modulating the activity of the peptidyl transferase center of the ribosome', *RNA*. doi: 10.1261/rna.980308.
- Berres, M. E. *et al.* (2017) 'Transcriptome profiling identifies ribosome biogenesis as a target of alcohol teratogenicity and vulnerability during early embryogenesis', *PLoS ONE*, 12(1). doi: 10.1371/journal.pone.0169351.
- Birkedal, U. *et al.* (2015) 'Profiling of ribose methylations in RNA by high-throughput sequencing', *Angewandte Chemie - International Edition*, 54(2). doi: 10.1002/anie.201408362.
- Black, D. L. (2003) 'Mechanisms of alternative pre-messenger RNA splicing', *Annual Review of Biochemistry*. doi: 10.1146/annurev.biochem.72.121801.161720.
- Blasiak, J. *et al.* (2020) 'The aging stress response and its implication for amd pathogenesis', *International Journal of Molecular Sciences*. doi: 10.3390/ijms21228840.
- Blin, M. *et al.* (2021) 'DNA molecular combing-based replication fork directionality profiling', *Nucleic Acids Research*, 49(12). doi: 10.1093/nar/gkab219.
- Boccaletto, P. *et al.* (2018) 'MODOMICS: A database of RNA modification pathways. 2017 update', *Nucleic Acids Research*, 46(D1). doi: 10.1093/nar/gkx1030.
- Boo, S. H. and Kim, Y. K. (2020) 'The emerging role of RNA modifications in the regulation of mRNA

- stability', *Experimental and Molecular Medicine*. doi: 10.1038/s12276-020-0407-z.
- Braglia, P., Kawauchi, J. and Proudfoot, N. J. (2011) 'Co-transcriptional RNA cleavage provides a failsafe termination mechanism for yeast RNA polymerase i', *Nucleic Acids Research*, 39(4). doi: 10.1093/nar/gkq894.
- Branton, D. *et al.* (2009) 'Nihms-89972.Pdf', *Nature Biotechnology*, 26(10), pp. 1146–1153. doi: 10.1038/nbt.1495.The.
- Buchhaupt, M. *et al.* (2014) 'Partial methylation at Am100 in 18S rRNA of Baker's yeast reveals ribosome heterogeneity on the level of eukaryotic rRNA modification', *PLoS ONE*, 9(2). doi: 10.1371/journal.pone.0089640.
- Burger, K. *et al.* (2010) 'Chemotherapeutic drugs inhibit ribosome biogenesis at various levels', *Journal of Biological Chemistry*, 285(16). doi: 10.1074/jbc.M109.074211.
- Caburet, S. *et al.* (2005) 'Human ribosomal RNA gene arrays display a broad range of palindromic structures', *Genome Research*, 15(8). doi: 10.1101/gr.3970105.
- Cantara, W. A. *et al.* (2011) 'The RNA modification database, RNAMDB: 2011 update', *Nucleic Acids Research*, 39(SUPPL. 1). doi: 10.1093/nar/gkq1028.
- Carone, B. R. *et al.* (2010) 'Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals', *Cell*, 143(7). doi: 10.1016/j.cell.2010.12.008.
- Carron, C. *et al.* (2011) 'Analysis of two human pre-ribosomal factors, bystin and hTsr1, highlights differences in evolution of ribosome biogenesis between yeast and mammals', *Nucleic Acids Research*, 39(1). doi: 10.1093/nar/gkq734.
- Cattanach, B. M. (1966) 'The location of Cattanach's translocation in the X-chromosome linkage map of the mouse', *Genetical Research*, 8(2). doi: 10.1017/S0016672300010107.
- Chanou, A. and Hamperl, S. (2021) 'Single-Molecule Techniques to Study Chromatin', *Frontiers in Cell and Developmental Biology*. doi: 10.3389/fcell.2021.699771.
- Chen, B. *et al.* (2013) 'Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system', *Cell*, 155(7). doi: 10.1016/j.cell.2013.12.001.
- Chen, X. and Zhang, J. (2016) 'The Genomic Landscape of Position Effects on Protein Expression Level and Noise in Yeast', *Cell Systems*, 2(5). doi: 10.1016/j.cels.2016.03.009.
- Cheng, J. *et al.* (2017) '3.2-Å-resolution structure of the 90S preribosome before A1 pre-rRNA cleavage', *Nature Structural and Molecular Biology*, 24(11). doi: 10.1038/nsmb.3476.
- Cho, Y. *et al.* (2020) 'FRACTAL: Signal amplification of immunofluorescence: Via cyclic staining of target molecules', *Nanoscale*, 12(46). doi: 10.1039/d0nr05800a.
- Chudoba, I. *et al.* (2004) 'mBAND: A high resolution multicolor banding technique for the detection of complex intrachromosomal aberrations', in *Cytogenetic and Genome Research*. doi:

10.1159/000077521.

Conconi, A. *et al.* (1989) 'Two different chromatin structures coexist in ribosomal RNA genes throughout the cell cycle', *Cell*, 57(5), pp. 753–761. doi: 10.1016/0092-8674(89)90790-3.

Conrad, J. *et al.* (1998) 'The rluC gene of Escherichia coli codes for a pseudouridine synthase that is solely responsible for synthesis of pseudouridine at positions 955, 2504, and 2580 in 23 S ribosomal RNA', *Journal of Biological Chemistry*, 273(29). doi: 10.1074/jbc.273.29.18562.

Cui, C., Shu, W. and Li, P. (2016) 'Fluorescence in situ hybridization: Cell-based genetic diagnostic and research applications', *Frontiers in Cell and Developmental Biology*. doi: 10.3389/fcell.2016.00089.

Davis, D. R. (1995) 'Stabilization of RNA stacking by pseudouridine', *Nucleic Acids Research*, 23(24). doi: 10.1093/nar/23.24.5020.

Decatur, W. A. and Fournier, M. J. (2002) 'rRNA modifications and ribosome function', *Trends in Biochemical Sciences*. doi: 10.1016/S0968-0004(02)02109-6.

Deng, W. *et al.* (2015) 'CASFiSH: CRISPR/Cas9-mediated in situ labeling of genomic loci in fixed cells', *Proceedings of the National Academy of Sciences of the United States of America*, 112(38). doi: 10.1073/pnas.1515692112.

Dev, V. G. *et al.* (1977) 'Nucleolus organizers in Mus musculus subspecies and in the rag mouse cell line', *Genetics*, 86(2). doi: 10.1093/genetics/86.2.389.

Docherty, S. J. *et al.* (2012) 'A genetic association study of DNA methylation levels in the DRD4 gene region finds associations with nearby SNPs', *Behavioral and Brain Functions*. Behavioral and Brain Functions, 8(1), p. 1. doi: 10.1186/1744-9081-8-31.

Doll, A. and Grzeschik, K. H. (2001) 'Characterization of two novel genes, WBSCR20 and WBSCR22, deleted in Williams-Beuren syndrome', *Cytogenetics and Cell Genetics*, 95(1–2). doi: 10.1159/000057012.

Dominissini, D. *et al.* (2012) 'Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq', *Nature*, 485(7397). doi: 10.1038/nature11112.

Eichler, D. C. and Craig, N. (1994) 'Processing of Eukaryotic Ribosomal RNA', *Progress in Nucleic Acid Research and Molecular Biology*, 49(C). doi: 10.1016/S0079-6603(08)60051-3.

Eickbush, T. H. and Eickbush, D. G. (2007) 'Finely orchestrated movements: Evolution of the ribosomal RNA genes', *Genetics*. doi: 10.1534/genetics.107.071399.

Ekamper, P. *et al.* (2017) 'War-related excess mortality in The Netherlands, 1944–45: New estimates of famine- and non-famine-related deaths from national death records', *Historical Methods*, 50(2). doi: 10.1080/01615440.2017.1285260.

Eriksen, K. G. *et al.* (2017) 'Influence of intergenerational in utero parental energy and nutrient restriction on offspring growth in rural Gambia', *FASEB Journal*, 31(11), pp. 4928–4934. doi:

10.1096/fj.201700017R.

Ersfeld, K. (2004) 'Fiber-FISH: fluorescence in situ hybridization on stretched DNA.', *Methods in molecular biology (Clifton, N.J.)*, 270. doi: 10.1385/1-59259-793-9:395.

Fortier, L., Ponton, D. and Gilbert, M. (1995) 'The match/mismatch hypothesis and the feeding success of fish larvae in ice-covered southeastern Hudson Bay', *Marine Ecology Progress Series*, 120(1–3). doi: 10.3354/meps120011.

Frommer, M. *et al.* (1992) 'A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands', *Proceedings of the National Academy of Sciences of the United States of America*, 89(5). doi: 10.1073/pnas.89.5.1827.

Furlan, M. *et al.* (2021) 'Computational methods for RNA modification detection from nanopore direct RNA sequencing data', *RNA Biology*. doi: 10.1080/15476286.2021.1978215.

Gagnon-Kugler, T. *et al.* (2009) 'Loss of Human Ribosomal Gene CpG Methylation Enhances Cryptic RNA Polymerase II Transcription and Disrupts Ribosomal RNA Processing', *Molecular Cell*, 35(4). doi: 10.1016/j.molcel.2009.07.008.

Ganot, P., Bortolin, M. L. and Kiss, T. (1997) 'Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs', *Cell*, 89(5). doi: 10.1016/S0092-8674(00)80263-9.

Gao, Y. *et al.* (2021) 'Quantitative profiling of N 6-methyladenosine at single-base resolution in stem-differentiating xylem of *Populus trichocarpa* using Nanopore direct RNA sequencing', *Genome Biology*, 22(1). doi: 10.1186/s13059-020-02241-7.

Garalde, D. R. *et al.* (2018) 'Highly parallel direct RN A sequencing on an array of nanopores', *Nature Methods*. Nature Publishing Group, 15(3), pp. 201–206. doi: 10.1038/nmeth.4577.

Gemmell, N. J. (2021) 'Repetitive DNA: genomic dark matter matters', *Nature Reviews Genetics*, 22(6). doi: 10.1038/s41576-021-00354-8.

Genuth, N. R. and Barna, M. (2018) 'The Discovery of Ribosome Heterogeneity and Its Implications for Gene Regulation and Organismal Life', *Molecular Cell*. doi: 10.1016/j.molcel.2018.07.018.

Geula, S. *et al.* (2015) 'm6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation', *Science*, 347(6225). doi: 10.1126/science.1261417.

Ghoshal, K. *et al.* (2004) 'Role of Human Ribosomal RNA (rRNA) Promoter Methylation and of Methyl-CpG-binding Protein MBD2 in the Suppression of rRNA Gene Expression', *Journal of Biological Chemistry*, 279(8). doi: 10.1074/jbc.M309393200.

Gibbons, J. G. *et al.* (2015) 'Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes', *Proceedings of the National Academy of Sciences of the United States of America*, 112(8), pp. 2485–2490. doi: 10.1073/pnas.1416878112.

Gigova, A. *et al.* (2014) 'A cluster of methylations in the domain IV of 25S rRNA is required for

- ribosome stability', *RNA*, 20(10). doi: 10.1261/rna.043398.113.
- Gleditzsch, D. *et al.* (2019) 'PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures', *RNA Biology*. doi: 10.1080/15476286.2018.1504546.
- Gluckman, P. D. and Hanson, M. A. (2004) 'Developmental origins of disease paradigm: A mechanistic and evolutionary perspective', *Pediatric Research*. doi: 10.1203/01.PDR.0000135998.08025.FB.
- Gnirke, A. *et al.* (2009) 'Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing', *Nature Biotechnology*, 27(2). doi: 10.1038/nbt.1523.
- Godfrey, K. M. *et al.* (2007) 'Epigenetic mechanisms and the mismatch concept of the developmental origins of health and disease', *Pediatric Research*. doi: 10.1203/pdr.0b013e318045bedb.
- Gonzalez, I. L. and Sylvester, J. E. (1995) 'Complete sequence of the 43-kb human ribosomal DNA repeat: Analysis of the intergenic spacer', *Genomics*, 27(2). doi: 10.1006/geno.1995.1049.
- Goodpasture, C. and Bloom, S. E. (1975) 'Visualization of nucleolar organizer regions in mammalian chromosomes using silver staining', *Chromosoma*, 53(1). doi: 10.1007/BF00329389.
- Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) 'Coming of age: Ten years of next-generation sequencing technologies', *Nature Reviews Genetics*. Nature Publishing Group, 17(6), pp. 333–351. doi: 10.1038/nrg.2016.49.
- Gopanenko, A. V. *et al.* (2021) 'Knockdown of the ribosomal protein el38 in hek293 cells changes the translational efficiency of specific genes', *International Journal of Molecular Sciences*, 22(9). doi: 10.3390/ijms22094531.
- Grandi, P. *et al.* (2002) '90S pre-ribosomes include the 35S pre-rRNA, the U3 snoRNP, and 40S subunit processing factors but predominantly lack 60S synthesis factors', *Molecular Cell*, 10(1). doi: 10.1016/S1097-2765(02)00579-8.
- Grewal, S. I. S. and Elgin, S. C. R. (2007) 'Transcription and RNA interference in the formation of heterochromatin', *Nature*. doi: 10.1038/nature05914.
- Grozdanov, P., Georgiev, O. and Karagyozev, L. (2003) 'Complete sequence of the 45-kb mouse ribosomal DNA repeat: Analysis of the intergenic spacer', *Genomics*, 82(6), pp. 637–643. doi: 10.1016/S0888-7543(03)00199-X.
- Grozhiik, A. V. *et al.* (2017) 'Mapping m6A at individual-nucleotide resolution using crosslinking and immunoprecipitation (MiCLIP)', in *Methods in Molecular Biology*. doi: 10.1007/978-1-4939-6807-7_5.
- Grummt, I. and Ladurner, A. G. (2008) 'A Metabolic Throttle Regulates the Epigenetic State of rDNA', *Cell*. doi: 10.1016/j.cell.2008.04.026.
- Grummt, I. and Längst, G. (2013) 'Epigenetic control of RNA polymerase I transcription in

- mammalian cells', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. doi: 10.1016/j.bbagr.2012.10.004.
- Grummt, I. and Pikaard, C. S. (2003) 'Epigenetic silencing of RNA polymerase I transcription', *Nature Reviews Molecular Cell Biology*. doi: 10.1038/nrm1171.
- Guhaniyogi, J. and Brewer, G. (2001) 'Regulation of mRNA stability in mammalian cells', *Gene*. doi: 10.1016/S0378-1119(01)00350-X.
- Hall, D. B., Wade, J. T. and Struhl, K. (2006) 'An HMG Protein, Hmo1, Associates with Promoters of Many Ribosomal Protein Genes and throughout the rRNA Gene Locus in *Saccharomyces cerevisiae*', *Molecular and Cellular Biology*, 26(9). doi: 10.1128/mcb.26.9.3672-3679.2006.
- Harewood, L. and Fraser, P. (2014) 'The impact of chromosomal rearrangements on regulation of gene expression', *Human Molecular Genetics*, 23(R1). doi: 10.1093/hmg/ddu278.
- Hebras, J. *et al.* (2020) 'Developmental changes of rRNA ribose methylations in the mouse', *RNA Biology*, 17(1). doi: 10.1080/15476286.2019.1670598.
- Heiskanen, M. *et al.* (1995) 'Visual mapping by fiber-FISH', *Genomics*, 30(1), pp. 31–36. doi: 10.1006/geno.1995.0005.
- Helm, M. (2006) 'Post-transcriptional nucleotide modification and alternative folding of RNA', *Nucleic Acids Research*. doi: 10.1093/nar/gkj471.
- Helm, M., Lyko, F. and Motorin, Y. (2019) 'Limited antibody specificity compromises epitranscriptomic analyses', *Nature Communications*. doi: 10.1038/s41467-019-13684-3.
- Henderson, A. S., Warburton, D. and Atwood, K. C. (1972) 'Location of ribosomal DNA in the human chromosome complement.', *Proceedings of the National Academy of Sciences of the United States of America*, 69(11), pp. 3394–3398. doi: 10.1073/pnas.69.11.3394.
- Hendra, C. *et al.* (2021) 'Detection of m6A from direct RNA sequencing using a Multiple Instance Learning framework', *bioRxiv*.
- Heng, H. H. Q. and Tsui, L. C. (1993) 'Modes of DAPI banding and simultaneous in situ hybridization', *Chromosoma*, 102(5). doi: 10.1007/BF00661275.
- Henikoff, S., Jackson, J. M. and Talbert, P. B. (1995) 'Distance and pairing effects on the brown(Dominant) heterochromatic element in *Drosophila*', *Genetics*, 140(3). doi: 10.1093/genetics/140.3.1007.
- Henras, A. K. *et al.* (2015) 'An overview of pre-ribosomal RNA processing in eukaryotes', *Wiley Interdisciplinary Reviews: RNA*, 6(2), pp. 225–242. doi: 10.1002/wrna.1269.
- Hiranniramol, K., Chen, Y. and Wang, X. (2020) 'CRISPR/Cas9 Guide RNA Design Rules for Predicting Activity', in *Methods in Molecular Biology*. doi: 10.1007/978-1-0716-0290-4_19.
- Hirsch, C. and Schildknecht, S. (2019) 'In vitro research reproducibility: Keeping up high standards',

- Frontiers in Pharmacology*. doi: 10.3389/fphar.2019.01484.
- Holland, M. L. *et al.* (2016) 'Early-life nutrition modulates the epigenetic state of specific rDNA genetic variants in mice', *Science*. American Association for the Advancement of Science, p. aaf7040.
- Huang, G. *et al.* (2015) 'Molecular basis of embryonic stem cell self-renewal: From signaling pathways to pluripotency network', *Cellular and Molecular Life Sciences*. doi: 10.1007/s00018-015-1833-2.
- Hughes, D. G. and Maden, E. H. (1978) 'The pseudouridine contents of the ribosomal ribonucleic acids of three vertebrate species. Numerical correspondence between pseudouridine residues and 2'-O-methyl groups is not always conserved', *Biochemical Journal*, 171(3). doi: 10.1042/bj1710781.
- Jenjaroenpun, P. *et al.* (2021) 'Decoding the epitranscriptional landscape from native RNA sequences', *Nucleic Acids Research*, 49(2). doi: 10.1093/nar/gkaa620.
- Jhanwar, S. C., Prensky, W. and Chaganti, R. S. K. (1981) 'Localization and metabolic activity of ribosomal genes in Chinese hamster meiotic and mitotic chromosomes', *Cytogenetics and Cell Genetics*, 30(1). doi: 10.1159/000131586.
- Jin, H. *et al.* (2019) 'N6-methyladenosine modification of ITGA6 mRNA promotes the development and progression of bladder cancer', *EBioMedicine*, 47. doi: 10.1016/j.ebiom.2019.07.068.
- Jonkhout, N. *et al.* (2017) 'The RNA modification landscape in human disease', *RNA*. doi: 10.1261/rna.063503.117.
- Kahl, V. F. S. *et al.* (2020) 'Telomere Length Measurement by Molecular Combing', *Frontiers in Cell and Developmental Biology*, 8. doi: 10.3389/fcell.2020.00493.
- Kawai, G. *et al.* (1992) 'Conformational Rigidity of Specific Pyrimidine Residues in tRNA Arises from Posttranscriptional Modifications That Enhance Steric Interaction between the Base and the 2'-Hydroxyl Group', *Biochemistry*, 31(4). doi: 10.1021/bi00119a012.
- Kaykov, A. *et al.* (2016) 'Molecular Combing of Single DNA Molecules on the 10 Megabase Scale', *Scientific Reports*. Nature Publishing Group, 6, pp. 1–9. doi: 10.1038/srep19636.
- Kellner, S. *et al.* (2014) 'Absolute and relative quantification of RNA modifications via biosynthetic isotopomers', *Nucleic Acids Research*, 42(18). doi: 10.1093/nar/gku733.
- Kent, T., Lapik, Y. R. and Pestov, D. G. (2009) 'The 5' external transcribed spacer in mouse ribosomal RNA contains two cleavage sites', *RNA*, 15(1). doi: 10.1261/rna.1384709.
- Kerkel, K. *et al.* (2008) 'Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation', *Nature Genetics*, 40(7), pp. 904–908. doi: 10.1038/ng.174.
- Kim, N. and Jinks-Robertson, S. (2012) 'Transcription as a source of genome instability', *Nature Reviews Genetics*. doi: 10.1038/nrg3152.

- King, T. H. *et al.* (2003) 'Ribosome structure and activity are altered in cells lacking snoRNPs that form pseudouridines in the peptidyl transferase center', *Molecular Cell*, 11(2). doi: 10.1016/S1097-2765(03)00040-6.
- Kiss-László, Z. *et al.* (1996) 'Site-specific ribose methylation of preribosomal RNA: A novel function for small nucleolar RNAs', *Cell*, 85(7). doi: 10.1016/S0092-8674(00)81308-2.
- Kleinjan, D. J. and Van Heyningen, V. (1998) 'Position effect in human genetic disease', *Human Molecular Genetics*. doi: 10.1093/hmg/7.10.1611.
- Klinge, S. and Woolford, J. L. (2019) 'Ribosome assembly coming into focus', *Nature Reviews Molecular Cell Biology*. doi: 10.1038/s41580-018-0078-y.
- Kobayashi, T. *et al.* (2004) 'SIR2 regulates recombination between different rDNA repeats, but not recombination within individual rRNA genes in yeast', *Cell*, 117(4). doi: 10.1016/S0092-8674(04)00414-3.
- Kobayashi, T. (2011) 'How does genome instability affect lifespan?: Roles of rDNA and telomeres T Kobayashi rDNA, telomeres and aging', *Genes to Cells*. doi: 10.1111/j.1365-2443.2011.01519.x.
- Kopp, E., Mayr, B. and Schleger, W. (1988) 'Ribosomal RNA expression in a mammalian hybrid, the hinny', *Chromosoma*, 96(6). doi: 10.1007/BF00303037.
- Koš, M. and Tollervey, D. (2010) 'Yeast Pre-rRNA Processing and Modification Occur Cotranscriptionally', *Molecular Cell*, 37(6). doi: 10.1016/j.molcel.2010.02.024.
- Kovářová, P. *et al.* (2007) 'Chromosome analysis and sorting in *Vicia sativa* using flow cytometry', *Biologia Plantarum*, 51(1). doi: 10.1007/s10535-007-0009-9.
- Krogh, N. *et al.* (2020) 'Profiling of ribose methylations in ribosomal RNA from diffuse large B-cell lymphoma patients for evaluation of ribosomes as drug targets', *NAR Cancer*, 2(4). doi: 10.1093/narcan/zcaa035.
- Kuderna, L. F. K. *et al.* (2020) 'Flow Sorting Enrichment and Nanopore Sequencing of Chromosome 1 From a Chinese Individual', *Frontiers in Genetics*, 10. doi: 10.3389/fgene.2019.01315.
- Kumar, S. and Mohapatra, T. (2021) 'Deciphering Epitranscriptome: Modification of mRNA Bases Provides a New Perspective for Post-transcriptional Regulation of Gene Expression', *Frontiers in Cell and Developmental Biology*. doi: 10.3389/fcell.2021.628415.
- Kurihara, Y. *et al.* (1994) 'Chromosomal locations of Ag-NORs and clusters of ribosomal DNA in laboratory strains of mice', *Mammalian Genome*, 5(4). doi: 10.1007/BF00360550.
- Labit, H. *et al.* (2008) 'A simple and optimized method of producing silanized surfaces for FISH and replication mapping on combed DNA fibers', *BioTechniques*, 45(6). doi: 10.2144/000113002.
- Lafontaine, D. L. J. and Tollervey, D. (2001) 'The function and synthesis of ribosomes', *Nature Reviews Molecular Cell Biology*, 2(7), pp. 514–520. doi: 10.1038/35080045.

- Lazdins, I. B., Delannoy, M. and Sollner-Webb, B. (1997) 'Analysis of nucleolar transcription and processing domains and pre-rRNA movements by in situ hybridization', *Chromosoma*, 105(7–8). doi: 10.1007/BF02510485.
- Lebofsky, R. and Bensimon, A. (2003) 'Single DNA molecule analysis: Applications of molecular combing', *Briefings in Functional Genomics and Proteomics*, 1(4), pp. 385–396. doi: 10.1093/bfpg/1.4.385.
- Lee, T. M. and Zucker, I. (1988) 'Vole infant development is influenced perinatally by maternal photoperiodic history', *American Journal of Physiology - Regulatory Integrative and Comparative Physiology*, 255(5). doi: 10.1152/ajpregu.1988.255.5.r831.
- Leger, A. et al. (2021) 'RNA modifications detection by comparative Nanopore direct RNA sequencing', *Nature Communications*, 12(1). doi: 10.1038/s41467-021-27393-3.
- Lewis, J. D. and Izaurrealde, E. (1997) 'The role of the cap structure in RNA processing and nuclear export', *European Journal of Biochemistry*. doi: 10.1111/j.1432-1033.1997.00461.x.
- Li, X. et al. (2015) 'Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome', *Nature Chemical Biology*, 11(8). doi: 10.1038/nchembio.1836.
- Linder, B. et al. (2015) 'Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome', *Nature Methods*, 12(8). doi: 10.1038/nmeth.3453.
- Linder, B. and Jaffrey, S. R. (2019) 'Discovering and mapping the modified nucleotides that comprise the epitranscriptome of mrna', *Cold Spring Harbor Perspectives in Biology*, 11(6). doi: 10.1101/cshperspect.a032201.
- Liu, H., Begik, O. and Novoa, E. M. (2021) 'EpiNano: Detection of m6A RNA Modifications Using Oxford Nanopore Direct RNA Sequencing', in *Methods in Molecular Biology*. doi: 10.1007/978-1-0716-1374-0_3.
- Liu, Y. et al. (2015) 'In vitro CRISPR/cas9 system for efficient targeted DNA editing', *mBio*, 6(6). doi: 10.1128/mBio.01714-15.
- Liu, Y., Wang, P. and Dou, S. (2007) 'Single-molecule studies of DNA by molecular combing', *Progress in Natural Science*, 17(5). doi: 10.1080/10020070708541028.
- Logsdon, G. A., Vollger, M. R. and Eichler, E. E. (2020) 'Long-read human genome sequencing and its applications', *Nature Reviews Genetics*. doi: 10.1038/s41576-020-0236-x.
- Lynch, C., Chan, C. S. and Drake, A. J. (2017) 'Early life programming and the risk of non-alcoholic fatty liver disease', *Journal of Developmental Origins of Health and Disease*. doi: 10.1017/S2040174416000805.
- Maass, P. G. et al. (2018) 'Spatiotemporal allele organization by allele-specific CRISPR live-cell imaging (SNP-CLING)', *Nature Structural and Molecular Biology*, 25(2). doi: 10.1038/s41594-017-

0015-3.

Machnicka, M. A. *et al.* (2013) 'MODOMICS: A database of RNA modification pathways - 2013 update', *Nucleic Acids Research*, 41(D1). doi: 10.1093/nar/gks1007.

Madugalle, S. U. *et al.* (2020) 'RNA N6-Methyladenosine and the Regulation of RNA Localization and Function in the Brain', *Trends in Neurosciences*. doi: 10.1016/j.tins.2020.09.005.

Mallajosyula, S. S. and Pati, S. K. (2007) 'Effect of protonation on the electronic properties of DNA base pairs: Applications for molecular electronics', *Journal of Physical Chemistry B*, 111(40). doi: 10.1021/jp076063m.

Mao, Y. *et al.* (2019) 'm6A in mRNA coding regions promotes translation via the RNA helicase-containing YTHDC2', *Nature Communications*, 10(1). doi: 10.1038/s41467-019-13317-9.

Matsuda, Y. *et al.* (1994) 'Chromosomal mapping of mouse 5S rRNA genes by direct R-banding fluorescence in situ hybridization', *Cytogenetics and Cell Genetics*, 66(4). doi: 10.1159/000133704.

Mauro, V. P. and Edelman, G. M. (2002) 'The ribosome filter hypothesis', *Proceedings of the National Academy of Sciences of the United States of America*, 99(19). doi: 10.1073/pnas.192442499.

Mcgrath, J. J. *et al.* (2013) 'Where GWAS and epidemiology meet: Opportunities for the simultaneous study of genetic and environmental risk factors in schizophrenia', *Schizophrenia Bulletin*, 39(5). doi: 10.1093/schbul/sbt108.

Meijer, L. and Raymond, E. (2003) 'Roscovitine and other purines as kinase inhibitors. From starfish oocytes to clinical trials', *Accounts of Chemical Research*. doi: 10.1021/ar0201198.

Meyer, K. D. *et al.* (2012) 'Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons', *Cell*, 149(7). doi: 10.1016/j.cell.2012.05.003.

Michalet, X. *et al.* (1997) 'Dynamic molecular combing: Stretching the whole human genome for high-resolution studies', *Science*, 277(5331), pp. 1518–1523. doi: 10.1126/science.277.5331.1518.

Micura, R. *et al.* (2001) 'Methylation of the nucleobases in RNA oligonucleotides mediates duplex-hairpin conversion', *Nucleic Acids Research*, 29(19). doi: 10.1093/nar/29.19.3997.

Mikheikin, A. *et al.* (2017) 'DNA nanomapping using CRISPR-Cas9 as a programmable nanoparticle', *Nature Communications*, 8(1). doi: 10.1038/s41467-017-01891-9.

Moss, T. *et al.* (2007) 'A housekeeper with power of attorney: The rRNA genes in ribosome biogenesis', *Cellular and Molecular Life Sciences*. doi: 10.1007/s00018-006-6278-1.

Muller, H. J. (1930) 'Types of visible variations induced by X-rays in *Drosophila*', *Journal of Genetics*, 22(3). doi: 10.1007/BF02984195.

Natchiar, S. K. *et al.* (2017) 'Visualization of chemical modifications in the human 80S ribosome structure', *Nature*, 551(7681). doi: 10.1038/nature24482.

Nazari, Z. E. and Gurevich, L. (2013) 'Controlled deposition and combing of DNA across

- lithographically defined patterns on silicon', *Beilstein Journal of Nanotechnology*, 4(1). doi: 10.3762/bjnano.4.8.
- Nederhof, E. and Schmidt, M. V. (2012) 'Mismatch or cumulative stress: Toward an integrated hypothesis of programming effects', *Physiology and Behavior*. doi: 10.1016/j.physbeh.2011.12.008.
- Németh, A. (2014) 'Methyl-combing: Single-Molecule analysis of dna methylation on stretched DNA fibers', in *Methods in Molecular Biology*. doi: 10.1007/978-1-62703-706-8_18.
- Novoa, E. M., Mason, C. E. and Mattick, J. S. (2017) 'Charting the unknown epitranscriptome', *Nature Reviews Molecular Cell Biology*. doi: 10.1038/nrm.2017.49.
- Nurk, S. *et al.* (2022) 'The complete sequence of a human genome', *Science*, 376(6588). doi: 10.1126/science.abj6987.
- Ober, C. and Vercelli, D. (2011) 'Gene-environment interactions in human disease: Nuisance or opportunity?', *Trends in Genetics*. doi: 10.1016/j.tig.2010.12.004.
- Ofengand, J. (2002) 'Ribosomal RNA pseudouridines and pseudouridine synthases', in *FEBS Letters*. doi: 10.1016/S0014-5793(02)02305-0.
- Ogle, J. M. *et al.* (2001) 'Recognition of cognate transfer RNA by the 30S ribosomal subunit', *Science*, 292(5518). doi: 10.1126/science.1060612.
- Ottman, R. (1996) 'Gene-environment interaction: Definitions and study designs', *Preventive Medicine*. doi: 10.1006/pmed.1996.0117.
- Parker, M. T. *et al.* (2020) 'Nanopore direct RNA sequencing maps the complexity of arabidopsis mRNA processing and m6A modification', *eLife*, 9. doi: 10.7554/eLife.49658.
- Payne, A. *et al.* (2019) 'Bulkvis: A graphical viewer for Oxford nanopore bulk FAST5 files', *Bioinformatics*, 35(13). doi: 10.1093/bioinformatics/bty841.
- Polacek, N. and Mankin, A. S. (2005) 'The ribosomal peptidyl transferase center: Structure, function, evolution, inhibition', *Critical Reviews in Biochemistry and Molecular Biology*. doi: 10.1080/10409230500326334.
- Popov, A. *et al.* (2013) 'Duration of the first steps of the human rRNA processing', *Nucleus (United States)*, 4(2). doi: 10.4161/nucl.23985.
- Pratanwanich, P. N. *et al.* (2021) 'Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore', *Nature Biotechnology*, 39(11). doi: 10.1038/s41587-021-00949-w.
- Preti, M. *et al.* (2013) 'Gradual processing of the ITS1 from the nucleolus to the cytoplasm during synthesis of the human 18S rRNA', *Nucleic Acids Research*, 41(8). doi: 10.1093/nar/gkt160.
- Qu, L. hu, Nicoloso, M. and Bachellerie, J. pierre (1991) 'A sequence dimorphism in a conserved domain of human 28S rRNA. Uneven distribution of variant genes among individuals. differential

- expression in hela cells', *Nucleic Acids Research*, 19(5). doi: 10.1093/nar/19.5.1015.
- Ramagopal, S. (1990) 'Induction of cell-specific ribosomal proteins in aggregation-competent nonmorphogenetic *Dictyostelium discoideum*', *Biochemistry and Cell Biology*, 68(11). doi: 10.1139/o90-190.
- Ramaswami, G. *et al.* (2013) 'Identifying RNA editing sites using RNA sequencing data alone', *Nature Methods*, 10(2). doi: 10.1038/nmeth.2330.
- Rang, F. J., Kloosterman, W. P. and de Ridder, J. (2018) 'From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy', *Genome Biology*. doi: 10.1186/s13059-018-1462-9.
- Remacle, C., Bieswal, F. and Reusens, B. (2004) 'Programming of obesity and cardiovascular disease', *International Journal of Obesity*, 28. doi: 10.1038/sj.ijo.0802800.
- Reynolds, L. P. *et al.* (2010) 'Developmental programming: the concept, large animal models, and the key role of uteroplacental vascular development.', *Journal of animal science*. doi: 10.2527/jas.2009-2359.
- Rice, F. *et al.* (2010) 'The links between prenatal stress and offspring development and psychopathology: Disentangling environmental and inherited influences', *Psychological Medicine*, 40(2). doi: 10.1017/S0033291709005911.
- Richard, G.-F., Kerrest, A. and Dujon, B. (2008) 'Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes', *Microbiology and Molecular Biology Reviews*, 72(4), pp. 686–727. doi: 10.1128/mmbr.00011-08.
- Rodriguez-Algarra, F. *et al.* (2022) 'Genetic variation at mouse and human ribosomal DNA influences associated epigenetic states', *Genome Biology*, 23(1). doi: 10.1186/s13059-022-02617-x.
- Roseboom, T. J. *et al.* (2000) 'Coronary heart disease after prenatal exposure to the Dutch famine, 1944-45', *Heart*, 84(6), pp. 595–598. doi: 10.1136/heart.84.6.595.
- Roseboom, T., de Rooij, S. and Painter, R. (2006) 'The Dutch famine and its long-term consequences for adult health', *Early Human Development*, 82(8), pp. 485–491. doi: 10.1016/j.earlhumdev.2006.07.001.
- RUSSELL, L. B. and BANGHAM, J. W. (1961) 'Variegated-type position effects in the mouse.', *Genetics*, 46. doi: 10.1093/genetics/46.5.509.
- Sanij, E. *et al.* (2008) 'UBF levels determine the number of active ribosomal RNA genes in mammals', *Journal of Cell Biology*, 183(7), pp. 1259–1274. doi: 10.1083/jcb.200805146.
- Santoro, R. and Grummt, I. (2001) 'Molecular mechanisms mediating methylation-dependent silencing of ribosomal gene transcription', *Molecular Cell*, 8(3). doi: 10.1016/S1097-2765(01)00317-3.

- Sato, Y., Fujiwara, T. and Kimura, H. (2017) 'Expression and function of different guanine-plus-cytosine content 16S rRNA genes in *Haloarcula hispanica* at different temperatures', *Frontiers in Microbiology*, 8(MAR). doi: 10.3389/fmicb.2017.00482.
- Schilling, E., El Chartouni, C. and Rehli, M. (2009) 'Allele-specific DNA methylation in mouse strains is mainly determined by cis-acting sequences', *Genome Research*, 19(11). doi: 10.1101/gr.095562.109.
- Schmeing, T. M., Moore, P. B. and Steitz, T. A. (2003) 'Structures of deacylated tRNA mimics bound to the E site of the large ribosomal subunit', *RNA*, 9(11). doi: 10.1261/rna.5120503.
- Schossere, M. *et al.* (2015) 'Methylation of ribosomal RNA by NSUN5 is a conserved mechanism modulating organismal lifespan', *Nature Communications*, 6. doi: 10.1038/ncomms7158.
- Schulz, L. C. (2010) 'The Dutch hunger winter and the developmental origins of health and disease', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1012911107.
- Schwartz, S. *et al.* (2014) 'Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA', *Cell*, 159(1). doi: 10.1016/j.cell.2014.08.028.
- Schwartz, S. and Motorin, Y. (2017) 'Next-generation sequencing technologies for detection of modified nucleotides in RNAs', *RNA Biology*. doi: 10.1080/15476286.2016.1251543.
- Sharma, S. and Lafontaine, D. L. J. (2015) "'View From A Bridge": A New Perspective on Eukaryotic rRNA Base Modification', *Trends in Biochemical Sciences*. doi: 10.1016/j.tibs.2015.07.008.
- Shatkin, A. J. (1976) 'Capping of eucaryotic mRNAs', *Cell*. doi: 10.1016/0092-8674(76)90128-8.
- Shiao, Y. H. *et al.* (2005) 'Allele-specific germ cell epimutation in the spacer promoter of the 45S ribosomal RNA gene after Cr(III) exposure', *Toxicology and Applied Pharmacology*, 205(3). doi: 10.1016/j.taap.2004.10.017.
- Shiao, Y. H. *et al.* (2011) 'Ontogeny-Driven rDNA rearrangement, methylation, and transcription, and paternal influence', *PLoS ONE*, 6(7). doi: 10.1371/journal.pone.0022266.
- Silverstein, R. A., De Valdivia, E. G. and Visa, N. (2011) 'The incorporation of 5-fluorouracil into RNA affects the ribonucleolytic activity of the exosome subunit Rrp6', *Molecular Cancer Research*, 9(3). doi: 10.1158/1541-7786.MCR-10-0084.
- Sims, J., Schlögelhofer, P. and Kurzbauer, M. T. (2021) 'From Microscopy to Nanoscopy: Defining an Arabidopsis thaliana Meiotic Atlas at the Nanometer Scale', *Frontiers in Plant Science*. doi: 10.3389/fpls.2021.672914.
- Sinclair, D. A. and Guarente, L. (1997) 'Extrachromosomal rDNA circles - A cause of aging in yeast', *Cell*, 91(7). doi: 10.1016/S0092-8674(00)80493-6.
- Slatko, B. E., Gardner, A. F. and Ausubel, F. M. (2018) 'Overview of Next-Generation Sequencing Technologies', *Current Protocols in Molecular Biology*, 122(1). doi: 10.1002/cpmb.59.

- Sloan, K. E. *et al.* (2017) 'Tuning the ribosome: The influence of rRNA modification on eukaryotic ribosome biogenesis and function', *RNA Biology*. doi: 10.1080/15476286.2016.1259781.
- Smith, A. M. *et al.* (2019) 'Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing', *PLoS ONE*, 14(5), pp. 1–15. doi: 10.1371/journal.pone.0216709.
- Speicher, M. R., Ballard, S. G. and Ward, D. C. (1996) 'Karyotyping human chromosomes by combinatorial multi-fluor FISH', *Nature Genetics*, 12(4). doi: 10.1038/ng0496-368.
- Spielmann, M., Lupiáñez, D. G. and Mundlos, S. (2018) 'Structural variation in the 3D genome', *Nature Reviews Genetics*. doi: 10.1038/s41576-018-0007-0.
- Srivastava, A. K. and Schlessinger, D. (1991) 'Structure and organization of ribosomal DNA', *Biochimie*, 73(6). doi: 10.1016/0300-9084(91)90042-Y.
- Stefanovsky, V. Y. *et al.* (2001) 'An immediate response of ribosomal transcription to growth factor stimulation in mammals is mediated by ERK phosphorylation of UBF', *Molecular Cell*, 8(5), pp. 1063–1073. doi: 10.1016/S1097-2765(01)00384-7.
- Steffensen, D. M., Duffey, P. and Prenskey, W. (1974) 'Localisation of 5S ribosomal RNA genes on human chromosome 1', *Nature*, 252(5485). doi: 10.1038/252741a0.
- Stephenson, W. *et al.* (2022) 'Direct detection of RNA modifications and structure using single-molecule nanopore sequencing', *Cell Genomics*, 2(2). doi: 10.1016/j.xgen.2022.100097.
- Stoiber, M. H. *et al.* (2016) 'De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing', *bioRxiv*.
- Stults, D. M. *et al.* (2008) 'Genomic architecture and inheritance of human ribosomal RNA gene clusters', *Genome Research*, 18(1), pp. 13–18. doi: 10.1101/gr.6858507.
- Sturtevant, A. H. (1925) 'THE EFFECTS OF UNEQUAL CROSSING OVER AT THE BAR LOCUS IN DROSOPHILA', *Genetics*, 10(2). doi: 10.1093/genetics/10.2.117.
- Sutton, G. M., Centanni, A. V. and Butler, A. A. (2010) 'Protein malnutrition during pregnancy in C57BL/6J mice results in offspring with altered circadian physiology before obesity', *Endocrinology*, 151(4). doi: 10.1210/en.2009-1133.
- Talbert, P. B. and Henikoff, S. (2006) 'Spreading of silent chromatin: Inaction at a distance', *Nature Reviews Genetics*. doi: 10.1038/nrg1920.
- Tamm, C., Galitó, S. P. and Annerén, C. (2013) 'A comparative study of protocols for mouse embryonic stem cell culturing', *PLoS ONE*, 8(12). doi: 10.1371/journal.pone.0081156.
- Taoka, M. *et al.* (2009) 'An analytical platform for mass spectrometry-based identification and chemical analysis of RNA in ribonucleoprotein complexes', *Nucleic Acids Research*, 37(21). doi: 10.1093/nar/gkp732.
- Taoka, M. *et al.* (2018) 'Landscape of the complete RNA chemical modifications in the human 80S

- ribosome', *Nucleic Acids Research*, 46(18). doi: 10.1093/nar/gky811.
- Terenin, I. M. *et al.* (2005) 'A Cross-Kingdom Internal Ribosome Entry Site Reveals a Simplified Mode of Internal Ribosome Entry', *Molecular and Cellular Biology*, 25(17). doi: 10.1128/mcb.25.17.7879-7888.2005.
- Tollervey, D. *et al.* (1993) 'Temperature-sensitive mutations demonstrate roles for yeast fibrillarin in pre-rRNA processing, pre-rRNA methylation, and ribosome assembly', *Cell*, 72(3). doi: 10.1016/0092-8674(93)90120-F.
- Treangen, T. J. and Salzberg, S. L. (2012) 'Repetitive DNA and next-generation sequencing: Computational challenges and solutions', *Nature Reviews Genetics*. doi: 10.1038/nrg3117.
- Tschochner, H. and Hurt, E. (2003) 'Pre-ribosomes on the road from the nucleolus to the cytoplasm', *Trends in Cell Biology*. doi: 10.1016/S0962-8924(03)00054-0.
- Tseng, H. *et al.* (2008) 'Mouse ribosomal RNA genes contain multiple differentially regulated variants', *PLoS ONE*, 3(3). doi: 10.1371/journal.pone.0001843.
- Valdez, B. C. *et al.* (2004) 'The Treacher Collins syndrome (TCOF1) gene product is involved in ribosomal DNA gene transcription by interacting with upstream binding factor', *Proceedings of the National Academy of Sciences of the United States of America*, 101(29), pp. 10709–10714. doi: 10.1073/pnas.0402492101.
- Wakimoto, B. T. and Hearn, M. G. (1990) 'The effects of chromosome rearrangements on the expression of heterochromatic genes in chromosome 2L of *Drosophila melanogaster*', *Genetics*, 125(1). doi: 10.1093/genetics/125.1.141.
- Walt, D. R. (2013) 'Optical methods for single molecule detection and analysis', *Analytical Chemistry*, 85(3). doi: 10.1021/ac3027178.
- Wang, M., Anikin, L. and Pestov, D. G. (2014) 'Two orthogonal cleavages separate subunit RNAs in mouse ribosome biogenesis', *Nucleic Acids Research*, 42(17). doi: 10.1093/nar/gku787.
- Wang, M. and Lemos, B. (2017) 'Ribosomal DNA copy number amplification and loss in human cancers is linked to tumor genetic context, nucleolus activity, and proliferation', *PLoS Genetics*, 13(9), pp. 1–24. doi: 10.1371/journal.pgen.1006994.
- Wang, M. and Pestov, D. G. (2011) '5'-end surveillance by Xrn2 acts as a shared mechanism for mammalian pre-rRNA maturation and decay', *Nucleic Acids Research*, 39(5). doi: 10.1093/nar/gkq1050.
- Wang, Y., Yang, Q. and Wang, Z. (2014) 'The evolution of nanopore sequencing', *Frontiers in Genetics*, 5(DEC). doi: 10.3389/fgene.2014.00449.
- Wang, Yunhao *et al.* (2021) 'Nanopore sequencing technology, bioinformatics and applications', *Nature Biotechnology*. doi: 10.1038/s41587-021-01108-x.

- Warner, J. R. (1999) 'The economics of ribosome biosynthesis in yeast', *Trends in Biochemical Sciences*. doi: 10.1016/S0968-0004(99)01460-7.
- Watkins, N. J. and Bohnsack, M. T. (2012) 'The box C/D and H/ACA snoRNPs: Key players in the modification, processing and the dynamic folding of ribosomal RNA', *Wiley Interdisciplinary Reviews: RNA*. doi: 10.1002/wrna.117.
- Weiss, L. C., Leimann, J. and Tollrian, R. (2015) 'Predator-induced defences in *Daphnia longicephala*: Location of kairomone receptors and timeline of sensitive phases to trait formation', *Journal of Experimental Biology*, 218(18). doi: 10.1242/jeb.124552.
- Wellauer, P. K. and Dawid, I. B. (1977) 'The structural organization of ribosomal DNA in *Drosophila melanogaster*', *Cell*. Elsevier, 10(2), pp. 193–212.
- Woodall, S. M. *et al.* (1996) 'Chronic maternal undernutrition in the rat leads to delayed postnatal growth and elevated blood pressure of offspring', *Pediatric Research*, 40(3). doi: 10.1203/00006450-199609000-00012.
- Xu, B. *et al.* (2017) 'Ribosomal DNA copy number loss and sequence variation in cancer', *PLoS Genetics*, 13(6), pp. 1–25. doi: 10.1371/journal.pgen.1006771.
- Xue, L. *et al.* (2020) 'Solid-state nanopore sensors', *Nature Reviews Materials*. doi: 10.1038/s41578-020-0229-6.
- Xue, S. and Barna, M. (2012) 'Specialized ribosomes: A new frontier in gene regulation and organismal biology', *Nature Reviews Molecular Cell Biology*. doi: 10.1038/nrm3359.
- Yang, J. *et al.* (2016) 'Mapping of complete set of ribose and base modifications of yeast rRNA by RP-HPLC and mung bean nuclease assay', *PLoS ONE*, 11(12). doi: 10.1371/journal.pone.0168873.
- Yokota, H. *et al.* (1997) 'A new method for straightening DNA molecules for optical restriction mapping', *Nucleic Acids Research*, 25(5). doi: 10.1093/nar/25.5.1064.
- Yoon, A. *et al.* (2006) 'Impaired control of IRES-mediated translation in X-linked dyskeratosis congenita', *Science*, 312(5775). doi: 10.1126/science.1123835.
- Yoshimura, J. *et al.* (2019) 'Recompleting the *Caenorhabditis elegans* genome', *Genome Research*, 29(6). doi: 10.1101/gr.244830.118.
- Yusupov, M. M. *et al.* (2001) 'Crystal structure of the ribosome at 5.5 Å resolution', *Science*, 292(5518). doi: 10.1126/science.1060089.
- Zaringhalam, M. and Papavasiliou, F. N. (2016) 'Pseudouridylation meets next-generation sequencing', *Methods*. doi: 10.1016/j.ymeth.2016.03.001.
- Zebarjadian, Y. *et al.* (1999) 'Point Mutations in Yeast CBF5 Can Abolish In Vivo Pseudouridylation of rRNA', *Molecular and Cellular Biology*, 19(11). doi: 10.1128/mcb.19.11.7461.
- Zentner, G. E. *et al.* (2011) 'Integrative genomic analysis of human ribosomal DNA', *Nucleic Acids*

Research, 39(12). doi: 10.1093/nar/gkq1326.

Zhang, D. *et al.* (2018) 'CRISPR-bind: a simple, custom CRISPR/dCas9-mediated labeling of genomic DNA for mapping in nanochannel arrays', *bioRxiv*.

Zimmer, C. *et al.* (2017) 'Transgenerational transmission of a stress-coping phenotype programmed by early-life stress in the Japanese quail', *Scientific Reports*, 7. doi: 10.1038/srep46125.

Zorbas, C. *et al.* (2015) 'The human 18S rRNA base methyltransferases DIMT1L and WBSCR22-TRMT112 but not rRNA modification are required for ribosome biogenesis', *Molecular Biology of the Cell*, 26(11). doi: 10.1091/mbc.E15-02-0073.