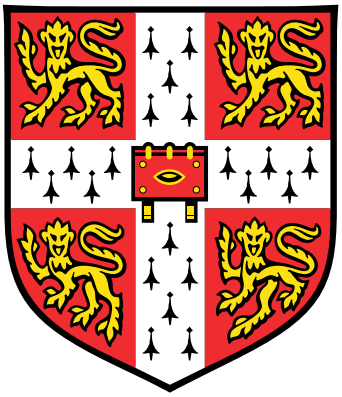
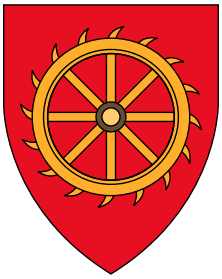


Computational Analysis of Transcriptional Regulation



Jack Michael Monahan

European Molecular Biology Laboratory,
European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

St. Catharine's College October 2019

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.
This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Acknowledgements and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation contains fewer than 60,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Jack Michael Monahan
October 2019

Summary

Computational Analysis of Transcriptional Regulation
by Jack Michael Monahan

It is doubtful Friedrich Miescher appreciated how groundbreaking and transformative his isolation of 'nuclein' in 1869 would prove. Eukaryotic gene expression is a noisy process that is subject to multiple layers of regulation. Key features of this are the three-dimensional (3D) chromatin organisation of eukaryotic genomes and the post-transcriptional control of RNA fates. Eukaryotic nuclear DNA is tightly packaged as chromatin that is further folded into higher order structures. The 3D folding of eukaryotic genome lends itself to the formation of interactions between otherwise distant regions of the genome. These interactions modulate transcription. I investigated the impact of human papillomavirus (HPV) 16 integration on host chromatin organisation and transcription using the W12 model for early cervical carcinogenesis with a novel Chromosome Conformation Capture (3C) method that specifically enriches for interactions involving viral integrants. Integration occurs without disrupting host 3D chromatin structure but alters the expression of many neighbouring host genes.
The advent of reliable protocols for performing single-cell RNA sequencing (scRNA-seq) has revealed that transcriptional noise is widespread and a biologically important feature in many populations of mammalian cells. Ageing is associated with the progressive decline in biological function. It has recently been described that aged somatic tissues have greater cell-to-cell transcriptional variability. Ageing is also associated with a decline in male fertility. Some have attributed this to the clonal expansion of selfish spermatogonial lineages. To address this, I explored the effect of ageing on the transcriptomes of sorted populations of mouse undifferentiated spermatogonia using bulk and single-cell RNA sequencing (RNA-seq). While subtle changes in mean gene expression are detectable, it was apparent that ageing, unlike in somatic tissues, leads to a decline in cell-to-cell transcriptional variability. This may reflect the phenomenon of selfish spermatogonial selection.

Finally, I explored the role of an RNA post-transcriptional modification (RPTM), N^6-methyladenosine (m^6A), in buffering transcriptional noise. Maternally-supplied YTHDF2 is essential for degradation of m^6A-modified transcripts during the maternal-to-zygotic transition (MZT) early in mammalian embryogenesis. YTHDF2 targets increase in abundance in its absence. Using scRNA-seq data generated from control and maternal conditional knock-out mouse zygotes I show that many of these targets exhibit greater cell-to-cell transcriptional heterogeneity in the absence of YTHDF2-mediated degradation. Suggesting that YTHDF2 has a additional function in buffering transcriptional noise.

I would like to dedicate this thesis to my parents Catherine and John.

Acknowledgements

First and foremost, I would like to thank my PhD supervisor Dr. Anton Enright for having me in his lab. It was a risk taking on a fellow Irishman but I think it has worked out well for the both of us. I am very grateful for all the support and guidance he has given me over the last four years. I am indebted to Adrien Léger for his all insight and input. Our discussions helped solve many of the problems I encountered during my PhD.

I would like to thank all the past members of the Enright Lab for their friendship and support: Matthew Davis, Tommaso Leonardi, Dimitrios Vitsios, Stijn van Dongen and Elsa Kentepozidou. The lab pub and bouldering sessions really helped maintain my sanity during my PhD.

All my work was done in collaboration with others. I would particularly like to thank Prof. Dónal O'Carroll in Edinburgh and the members of his lab Ivalya Ivanova and Marcos Morgan for their support and assistance in interpreting data. I would like to thank Nick Coleman, Ian Groves, Cinzia Scarpini and Emma Knight for their contributions to my early PhD experience. I would like to thank the members of my thesis advisory committee, John Marioni, Eric Miska and Marco Marcia for their guidance and advice. I would like to thank the EMBL International PhD programme for funding my research and EMBL-EBI for providing such a nurturing environment. I would like to extend my warmest gratitude to Prof. Ewan Birney for letting me join his lab after Anton moved on from EBI.

Importantly, I appreciate all the support that my friends in Dublin and Cambridge gave me. I would like to thank my friends from Dublin Colm Ferrari, Aoife Cleary, Sean Flynn, Robin Burns and Lara Cassidy for being there when I needed them. Nils, thank you for introducing me to the wonders of bouldering and the Blue Moon. Dani, thanks for all the interesting chats about science and the world; even if I don't necessarily remember them all. I'd like to thank Hannah, Omar, Steve and Fynn for all the Cambridge memories and experiences.

Additionally, I wish to thank my examiners Bertie Göttgens and Simon Moxon for a very enjoyable viva. Thank you very much for the very worthwhile discussion and for all the

advice you gave me.
I would like to express my gratitude to my parents, Catherine and John, for keeping me grounded and never letting me get too full of myself.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Abbreviations and Acronyms

GFRα1^{pos}	CD9^{pos} GFRα1-GFP^{pos} MIWI2-tdTom^{neg} c-Kit^{neg}
GFRα1^{pos} MIWI2^{pos}	CD9^{pos} GFRα1-GFP^{pos} MIWI2-tdTom^{pos} c-Kit^{neg}
Piwil4	piwi-like RNA-mediated gene silencing 4
***Ythdf2*^{CKO}**	Ythdf2$^{HA\text{-}Fl/HA\text{-}Fl}$; Vav-iCre
***Ythdf2*^{CTRL}**	*Ythdf2*$^{+/+}$ *Zp3Cre Tg*$^{+}$
***Ythdf2*^{WT}**	Ythdf2$^{HA\text{-}Fl/HA\text{-}Fl}$
***Ythdf2*^{mCKO}**	*Ythdf2*$^{HA\text{-}Fl/HA\text{-}Fl}$ *Zp3Cre Tg*$^{+}$
2D	two-dimensional
3C	Chromosome Conformation Capture
3D	three-dimensional
4C	Circularised Chromosome Conformation Capture
5mC	5-Methylcytosine
A$_{al}$	A Aligned
A$_{d}$	A Dark
A$_{pr}$	A Paired
A$_{p}$	A Pale
A$_{s}$	A Single
AGO	Argonaute
AML	acute myeloid leukaemia
APA	alternative polyadenylation
APC	antigen-presenting cell
APOT	amplification of papillomavirus oncogene transcripts

ATAC-seq assay for transposase-accessible chromatin using sequencing

BASiCS Bayesian analysis of single-cell sequencing data
BM basement membrane
bp base-pair

capture-seq Capture-sequencing
cDNA complementary DNA
CFS common fragile site
CGI CpG island
CHi-C Capture Hi-C
ChiP-seq Chromatin Immunoprecipitation sequencing
chr chromosome
CIN cervical intraepithelial neoplasia
cKO conditional knockout
CLIP UV-crosslinking and immunoprecipitation
CNV copy number variation
COF cofactor
CPM counts per million
CPSF Cleavage and polyadenylation specificity factor
CRE cis-regulatory element
CTCF CCCTC-binding factor
CTCFL CCCTC-binding factor like
CTD carboxyl-terminal domain

DC dendritic cell
DDR DNA damage response
DMRT1 *doublesex* and *mab-3*-related transcription factor 1
DMRTB *doublesex* and *mab-3*-related transcription factor B1
DMSO dimethyl sulfoxide
DNA Deoxyribonucleic acid
DNA-seq DNA sequencing
DNMT DNA methyltransferase

dNTP deoxyribonucleoside triphosphate
DSB double strand break
dsDNA double-stranded DNA
dsRNA double-stranded RNA

E embryonic day
EFDR expected false discovery rate
EGFR epidermal growth factor receptor
EHT endothelial-to-haematopoetic transition
EJC exon junction complex
ENA European Nucleotide Archive
ENCODE Encylcopedia of DNA Elements
ERCC External RNA Control Consortium
eRNA enhancer RNA
ERV endogenous retrovirus
EST expressed sequence tag

FACS fluorescence-activated cell sorting
FISH fluorescence *in situ* hybridisation
FOXJ3 forkhead box protein J3 isoform X2

Gb gigabase
gDNA genomic DNA
GDNF glial cell-derived neurotrophic factor
GFP green fluorescent protein
GFRα1 GDNF family receptor alpha-1
GO gene ontology
GOF gain-of-function
GSEA Gene set enrichment analysis
GTF general transcription factor
GV germinal vesicle

HDAC histone deacetylase

HDI Human Development Index
Hi-C high-throughput chromosome confirmation capture
HiCUP Hi-C User Pipeline
HNRNP heterogeneous nuclear ribonucleoprotein
HNSCC head and neck squamous cell carcinoma
HOX homeobox
HPV human papillomavirus
HR homologous recombination
HR-HPV high-risk HPV
HSC haematopoietic stem cell
HSPC haematopoietic stem and progenitor cell

IAP intracisternal-A particle
IP immunoprecipitation
IVF *in vitro* fertilisation
IVT *in vitro* transcribed

kb kilobase
KD knock-down
KO knockout

LAD lamina-associated domain
LCR long control region
LECA last eukaryotic common ancestor
LFC $\log_2$(fold-change)
lncRNA long non-coding RNA
LOF loss-of-function

m^{1}A N^1-methyladenosine
m^{6}A N^6-methyladenosine
m^{6}Am N^6,2′-O-dimethyladenosine
m^{6}AMD m^{6}A-mediated decay
m^{7}G N^7-methylguanosine

MAC	membrane attack complex
MACS	Magnetic Activated Cell Sorting
MAD	median absolute deviation
MAPQ	mapping quality
Mb	megabase
MCMC	Markov chain Monte Carlo
meRIP-seq	methylated RNA immunoprecipitation sequencing
mESC	human embryonic stem cell
mESC	mouse embryonic stem cell
MFE	minimum free energy
miRMD	miRNA-mediated decay
miRNA	microRNA
MIWI2^{pos}	CD9^{pos} GFRα1-GFP^{neg} MIWI2-tdTom^{pos} c-Kit^{neg}
MMEJ	microhomology-mediated end joining
mRNA	messenger RNA
MZT	maternal-to-zygotic transition
NB	negative binomial
ncRNA	non-coding RNA
NCx	normal cervical tissue
NEB	nuclear envelope breakdown
NGN3	Neurogenin 3
NGS	next-generation sequencing
NHEK	normal human epidermal keratinocytes
NMD	nonsense-mediated mRNA decay
nt	nucleotide
NTP	ribonucleoside triphosphate
ONS	UK Office for National Statistics
ONT	Oxford Nanopore Technologies
ORF	open reading frame

P-body processing body
P10 10 days post birth
PABPC1 Polyadenylate-binding cytoplasmic protein 1
PABPN1 Polyadenylate-binding nuclear protein 1
PacBio Pacific Biosciences
PAP polyadenylate polymerase
PAS polyadenylation signal
PCA Principal component analysis
PCR polymerase chain reaction
PGC primordial germ cell
PIC pre-initiation complex
piRNA Piwi-interacting RNA
PIWI P-element Induced WImpy testis
poly(A) polyadenylate
pRb retinoblastoma protein
pre-LSC pre-leukaemic stem cell

QC quality control
qPCR quantitative polymerase chain reaction

RET rearranged during transfection
RIP-seq RNA-immunoprecipitation sequencing
RISC RNA-induced silencing complex
RNA Ribonucleic acid
RNA-seq RNA sequencing
RNAi RNA interference
RNAP RNA polymerase
RNAP II RNA polymerase II
RNP ribonucleoprotein
RPTM RNA post-transcriptional modification
rRNA ribosomal RNA
RS-PCR restriction-site PCR
RT reverse transcription

RTase	reverse transcriptase
SAGE	serial analysis of gene expression
SC	stem cell
SCC	squamous cell carcinoma
SCRiBL	Sequence Capture of Regions Interacting with Bait Loci
scRNA-seq	single-cell RNA sequencing
SIL	squamous intraepithelial lesion
SINE	short interspersed element
siRNA	short hairpin RNA
siRNA	small interfering RNA
SMRT-seq	Single Molecule, Real-Time sequencing
SNP	single nucleotide polymorphism
SOX15	SRY-Box 15
SSC	spermatogonial stem cell
ssRNA	single-stranded RNA
STAT1	Signal transducer and activator of transcription 1
t-SNE	t-Distributed Stochastic Neighbour Embedding
TAC	transit-amplifying cell
TAD	topologically associating domain
TCA	citric acid
TCGA	The Cancer Genome Atlas
TEs	transposable element
TF	transcription factor
tRNA	transfer RNA
TSS	transcriptional start site
TUTase	terminal uridyl transferase
UMI	unique molecular identifier
URR	upstream regulatory region
UTR	untranslated region

v4C virtual 4C
VST variance-stabilising transformation

WT wild-type

YTH YT521-B homology
YTHDF2 YTH m^6A RNA binding protein 2

1

Introduction

Our understanding of the role and regulation of nucleic acids has come a long way since Friedrich Miescher first isolated what he termed *nuclein*, subsequently known as nucleic acid, from whole cell nuclei in 1869 (Dahm, 2005; Miescher-Rüsch, 1871). Allen (1941) was the first to describe the chemical differences between "plant" nucleic acids, Ribonucleic acid (RNA), and "animal" nucleic acids, Deoxyribonucleic acid (DNA), and suggest that both are present in the cells of animals, plants and fungi. Caspersson and Schultz (1939) identified an association between proteins and nucleic acids in the cytoplasm of proliferating cells. MacLeod and McCarty (1944) demonstrated, by means of bacterial transformation, that DNA was the likely molecule of heredity. Caspersson (1947) went onto to speculate that RNA and protein expression were interlinked because RNA levels increased in cells that were synthesising proteins. However, it was not until Watson and Crick (1953) determined the structure of DNA that it was possible for Crick (1958) to formulate his "Central Dogma of Molecular Biology". He proposed that, in general, biological information flows from DNA through RNA and into proteins. Jacob and Monod's seminal work on the *Escherichia coli lac* operon led them to theorise that a messenger RNA (mRNA) intermediate was necessary for the dynamic metabolic shift that occurs in response to the changing availability of lactose (Jacob and Monod, 1961a). The existence of mRNA was soon confirmed when Sydney Brenner identified an unstable RNA intermediate that transiently interacted with the protein-synthesising ribosomes (Brenner et al., 1961). This was followed by the discovery by Crick and Brenner of RNA molecules, transfer RNAs (tRNAs), that translate the mRNA triplet genetic code (codons) into amino acids (Crick et al., 1961). In turn, this led to the discovery that translation proceeds from the mRNA $5'$-hydroxyl group to $3'$-phosphate group (Lamfrom et al., 1966; Salas et al., 1965).

In contrast with James Watson's version of the Central Dogma (Watson, 1965), Francis Crick did not rule out the possibility of information flow from RNA to DNA. This less dogmatic view of the Central Dogma was supported by the discovery of retroviruses and viral reverse

transcriptase (RTase) that allows complementary DNA (cDNA) to be synthesised from an RNA template (Baltimore, 1964; Mizutani and Temin, 1970). The first full structure and nucleotide sequence of an RNA was determined from the *Saccharomyces cerevisiae* alanine tRNA by Holley et al. (1965). The discovery of RNA secondary structure led some to propose the existence of ribozymes, RNA molecules with catalytic activity (Crick, 1968; Orgel, 1968; Woese, 1967). This was later borne out by the discovery of self-splicing ribosomal RNAs (rRNAs) (Kruger et al., 1982). Carl Woese went so far as to propose that life started out as self-replicating ribozyme-like molecules.

Numerous breakthroughs in the field of RNA Biology have followed in the decades since the nucleic acid revolution of the 1950s and 1960s. Chief amongst them have been the dissection of the mechanisms underpinning transcriptional regulation and the revelation that many classes of RNA molecules, the non-coding RNAs (ncRNAs), are not directly involved in the synthesis of proteins. Rather, as will be discussed in due course, many of these are engaged in the post-transcriptional regulation of other RNA molecules. The existence of RNA-mediated mechanisms of post-transcriptional expression regulation was first hinted at when Napoli et al. (1990) inadvertently blocked a Petunia biosynthesis pathway *in trans* while trying to increase its activity with a transgene. Fire et al. (1998) later showed that the introduction of double-stranded RNA (dsRNA) into eukaryotic cells can trigger a homology-dependent RNA interference (RNAi) response that dramatically reduces the mRNA levels of endogenous target genes. Subsequent studies have revealed additional conserved mechanisms of RNA post-transcriptional regulation, most notably the microRNA (miRNA) mediated silencing pathway (Lagos-Quintana et al., 2001; Lau et al., 2001). Furthermore RNA molecules are not homogeneous, linear chains of adenines (As), cytosines (Cs), guanines (Gs) and uracils (Us), many are structured (Bartel, 2018; Blythe et al., 2016; Mayr, 2017) and indeed post-transcriptionally modified (Frye et al., 2016; Grozhik and Jaffrey, 2018; Pan, 2018; Vitsios and Enright, 2015).

Much of what has been achieved in the field of RNA Biology in recent decades has been accomplished due to the progress made in nucleic acid sequencing technologies and concomitant advances in computational analysis methods. Although RNA sequencing was developed first (Fiers et al., 1976), it was the application of Fred Sanger's eponymous dideoxy chain-termination method for sequencing DNA (Sanger et al., 1977) to libraries of short cDNA fragments that allowed the first eukaryotic *transcriptomes* to be interrogated (Adams et al., 1991; Sim et al., 1979; Velculescu et al., 1997). However it not until the advent of more high-throughput technologies like DNA microarrays and next-generation sequencing (NGS) (McGettigan, 2013; Nelson, 2001) that the field of *Transcriptomics* truly emerged. The widespread adoption of the NGS sequencing-by-synthesis method necessitated

the development of global sequence alignment tools capable of the fast mapping of short sequenced reads to a reference (Langmead and Salzberg, 2012; Langmead et al., 2009; Li and Durbin, 2010). Classically, investigations into transcriptional dynamics have focussed on changes in mean expression levels detected using bulk RNA sequencing or DNA microarrays. But with the development of reliable protocols for performing single-cell RNA sequencing (Picelli et al., 2014; Zheng et al., 2017) and computational methods for detecting sources of technical noise therein (Brennecke et al., 2013; Vallejos et al., 2015, 2016) there is a growing appreciation that alterations to cell-to-cell transcriptional heterogeneity have functional consequences (Martinez-Jimenez et al., 2017; Paul et al., 2015). Finally, third-generation technologies such as the Oxford Nanopore Technologies (ONT) Nanopore sequencing (Garalde et al., 2018; Jain et al., 2015) and Pacific Biosciences (PacBio) Single Molecule, Real-Time sequencing (SMRT-seq) methods (Eid et al., 2009) have the potential to further revolutionise the fields of genomics and transcriptomics by allowing long, single molecules to be sequenced (Niedringhaus et al., 2011). Longer nucleotide sequences have fewer possible matches in the genome (or transcriptome) and this reduces the computational burden of sequence alignment (Li, 2018). Longer reads better resolve regions of structural variation and stretches of repetitive sequence. The ONT direct RNA sequencing protocol eliminates the biases introduced by reverse transcription and amplification with polymerase chain reactions (Garalde et al., 2018). Furthermore, it is already possible to detect certain DNA and RNA modifications by interrogating the raw signal intensities underlying ONT base calls (Simpson et al., 2017; Workman et al., 2018).

1.1 Eukaryotic Transcriptional Regulation

Eukaryotic transcription is the episodic, three stage process by which RNA polymerases (RNAPs) synthesise RNA from a DNA template. There are three DNA-dependent RNAPs in higher eukaryotes (Roeder and Rutter, 1969). According to its modern definition, a gene is a genomic sequence that encodes a functional molecule, either RNA or protein (Gerstein et al., 2007). As all genes express RNA but not necessarily a protein, RNAPs are the real *Gene Machines*. RNAP I synthesises most rRNA precursors (Paule and White, 2000). RNA polymerase II (RNAP II) transcribes, amongst others, mRNA and miRNA encoding genes (Lee et al., 2004; McCracken et al., 1997). RNAP III catalyses the synthesis of tRNAs and other small RNA precursors (Cramer et al., 2008; Dieci et al., 2007; Paule and White, 2000). Transcription initiates at the transcriptional start site (TSS) within DNA cis-regulatory elements (CREs) known as promoters (Haberle and Stark, 2018). The promoter serves ad a platform for the recruitment and assembly of the pre-initiation complex (PIC). The core

PIC of RNAP IIs-transcribed genes is comprised of the RNAP II holoenzyme and six general transcription factors (GTFs), TFIIs A, B, D, E, F and H (Orphanides et al., 1996; Thomas and Chiang, 2006). While less than 1.5% of the human is comprised of protein-coding sequences (Dunham et al., 2012; Lander et al., 2001), the Encylcopedia of DNA Elements (ENCODE) Project has shown us that mammalian genomes specifically, and eukaryotic genomes in general, are rich in *cis* and *trans* regulatory elements. Furthermore, RNAP II-mediated transcription is influenced by distal CREs including so-called enhancer sequences (Banerji et al., 1981; Bulger and Groudine, 2011; Sagai et al., 2005). Most RNAP I and III genes encode 'housekeeping' functions and are constitutively expressed as a result (Dieci et al., 2007; Voit and Grummt, 2011). In contrast, many RNAP II-transcribed genes are dynamically regulated (Kornberg, 1999) with many of them being facultatively and/or inducibly expressed due to the existence of tissue-specific promoters and enhancers (Lenhard et al., 2012; Lettice et al., 2003; Pennacchio et al., 2007; Sagai et al., 2005; Weake and Workman, 2010).

1.1.1 The Histone Code

Eukaryotic genomes are more than one-dimensional DNA polymers, instead nuclear double-stranded DNA (dsDNA) is tightly packaged into a higher-order DNA-protein structures known as chromatin (Kornberg, 1977). The structural unit of which is the nucleosome. The nucleosome core particle consists of 146 bp superhelical DNA wrapped around a histone H2A, H2B, H3, and H4 hetero-octamer (Luger et al., 1997). Further compaction is mediated by interactions involving the basic histone amino-terminal tails and the DNA itself or the acidic regions of neighbouring histone octamers (Dorigo et al., 2003; Luger and Richmond, 1998; Zheng et al., 2005). Strikingly, the 6.6 gigabase (Gb) human diploid genome would be 2 metres long if fully unwound (de Wit and de Laat, 2012) yet it manages to be contained in a 5-20 micrometre diameter cell nucleus (Lammerding, 2011; Mirny, 2011). While compaction of the genome into chromatin and higher order structures elegantly solves the issue of storing the nuclear genome efficiently it also poses its own set of challenges to the cell. Heitz (1928) observed differentially staining regions of chromatin in the interphase nuclei of *Pellia epiphylla* that were consistently more or less condensed. He termed these regions hetero- and euchromatin; heterochromatin is associated with transcriptional repression (Brown, 1966) and euchromatin with active transcription (Chesteron et al., 1974). The more compact a region of chromatin is, the less accessible the DNA will be for the trans-acting factors mediating transcription and other cellular activities (Radman-Livaja and Rando, 2010). Unsurprisingly, mechanisms have evolved to modulate chromatin compaction and accessibility.
Allfrey et al. (1964) were the first to report that histone acetylation modulates transcription. Luger and Richmond (1998) reported that acetylation of the tails of histones H3 and H4

increases accessibility for chromatin remodelling complexes such as SWI/SNF which promote the formation of transcriptionally-permissive euchromatin. Studies have since shown that there are at least 80 covalent modifications to tails of histones H3 and H4 (Bannister and Kouzarides, 2011). This *histone code* defines one aspect of the eukaryotic *epigenome* and supports a dynamic genomic regulatory landscape (Barski et al., 2007; Benevolenskaya, 2007; Creyghton et al., 2010; Koch et al., 2007; Rosenfeld et al., 2009; Steger et al., 2008). If the ENCODE Project to identify functional sequences in the human genome has taught us nothing else, it demonstrates that transcription is pervasive and tightly linked to histone epigenetic states and chromatin accessibility (Birney et al., 2007; Dunham et al., 2012).

1.1.2 DNA Methylation

In addition to histone tail modifications, DNA itself is also epigenetically modified. Vertebrate CpG islands (CGIs) are short stretches of GC-rich, CpG-rich sequence that co-localise with 50% of annotated TSSs and the majority of promoters (Deaton and Bird, 2011). The cytosine-5 methylation of palindromic CpG dinucleotides is associated with gene silencing (Jones et al., 1998; Nan et al., 1998). 5-Methylcytosine (5mC) recruits histone deacetylases (HDACs) that compact chromatin and reduce chromatin accessibility. The symmetry of CpG dinucleotides allows cytosine-5 methylation to be stably inherited by daughter cells but methylation is not permanent (Li and Zhang, 2014). The global erasure of paternally- and maternally-inherited DNA methylation occurs at the zygotic, morula and blastula stages of mammalian embryogeneis (Kafri et al., 1992; Monk et al., 1987). The genome is remethylated after implantation. Epiblast-derived primordial germ cells (PGCs) undergo a further round of near total demethylation, in mice this occurs at embryonic day (E) 10–12 (Kafri et al., 1992; Popp et al., 2010). The spermatogonia of male neonates undergo remethylation with the *de novo* DNA methyltransferases (DNMTs) DNMT3A/B and the catalytically inactive DNMT3L as part of spermatogenesis (Bourc'his and Bestor, 2004; Ernst et al., 2017). Similarly, DNMT3A-DNMT3L together with DNMT1, which normally maintains DNA methylation in the soma, are responsible for *de novo* DNA methylation during oogenesis (Li et al., 2018). Indeed 5mC is so important that some have termed it the 5th nucleotide base (Lister and Ecker, 2009).

Promoter CGIs are predominantly unmethylated but methylation states may change with ageing and tumourigenesis (Esteller, 2002; Jung and Pfeifer, 2015). Furthermore, based on the age-associated changes in DNA methylation it has been suggested that the cytosine-5 methylation states at certain CpG sites can be used to create an *Epigenetic Clock* for measuring biological ageing (Horvath, 2013). Biological ageing being the age-associated progressive decline in function that is usually attributed to the accumulation of molecular

changes with time (Gems and Partridge, 2013; López-Otín et al., 2013). Outside of promoter regions, CpGs in repeats elements are also heavily methylated. Transcription of endogenous retroviruss (ERVs) is silenced by DNA methylation (Walsh et al., 1998). It is thought that DNA methylation initially evolved as a mechanism to silence these 'selfish' elements and maintain genome integrity and transcriptional regulation. The de-repression and mobilisation of ERVs is linked to insertional mutagenesis, the transcriptional activation of neighbouring genes and formation of chimaeric host-ERV transcripts (Maksakova et al., 2006; Peaston et al., 2004; Walsh et al., 1998). The germline remethylation of repeat loci is regulated by nuclear members of the P-element Induced WImpy testis (PIWI) family (Ernst et al., 2017). Studies in mice and *Drosophila* have shown that ageing leads to the increased expression and mobilisation of transposable elements (TEs) due to senescence-related epigenetic and chromatin changes (De Cecco et al., 2013; Wood and Helfand, 2013; Wood et al., 2016).

1.1.3 Chromatin Conformation

The compaction and folding of eukaryotic genomes as fractal globules (Lieberman-Aiden et al., 2009) is an obstacle for trans-acting factor accessibility but it also brings together regions in 3D space that would otherwise be distant in a linear sequence or even on distinct chromosomes (Lieberman-Aiden et al., 2009; Rao et al., 2014). There is strong evidence that the 3D organisation of the genome is a significant factor in the regulation of transcription (Bulger and Groudine, 2011; Rao et al., 2014). Microscopy-based assays have long shown that chromosomes occupy distinct, non-random positions within the interphase nucleus (Cremer and Cremer, 2006; Cremer et al., 1982) (Fig. 1.1a). Gene-rich chromosomes occupy the interior of nuclei while gene-poor ones are found at the peripheries (Cremer et al., 2001; Croft et al., 1999). Chromosomes only show limited intermingling at the borders of their territories (Branco and Pombo, 2006). Furthermore, euchromatin and heterochromatin are segregated within these chromosomal territories (Kosak et al., 2002; Meister et al., 2010; Noordermeer et al., 2008; Ragoczy et al., 2006). Techniques like 3D fluorescence *in-situ* hybridisation (FISH) give an overview of the nuclear chromatin organisation but lack the scalability to detect novel chromatin-chromatin interactions genome-wide at kilobase (kb) resolution (Solovei et al., 2002).

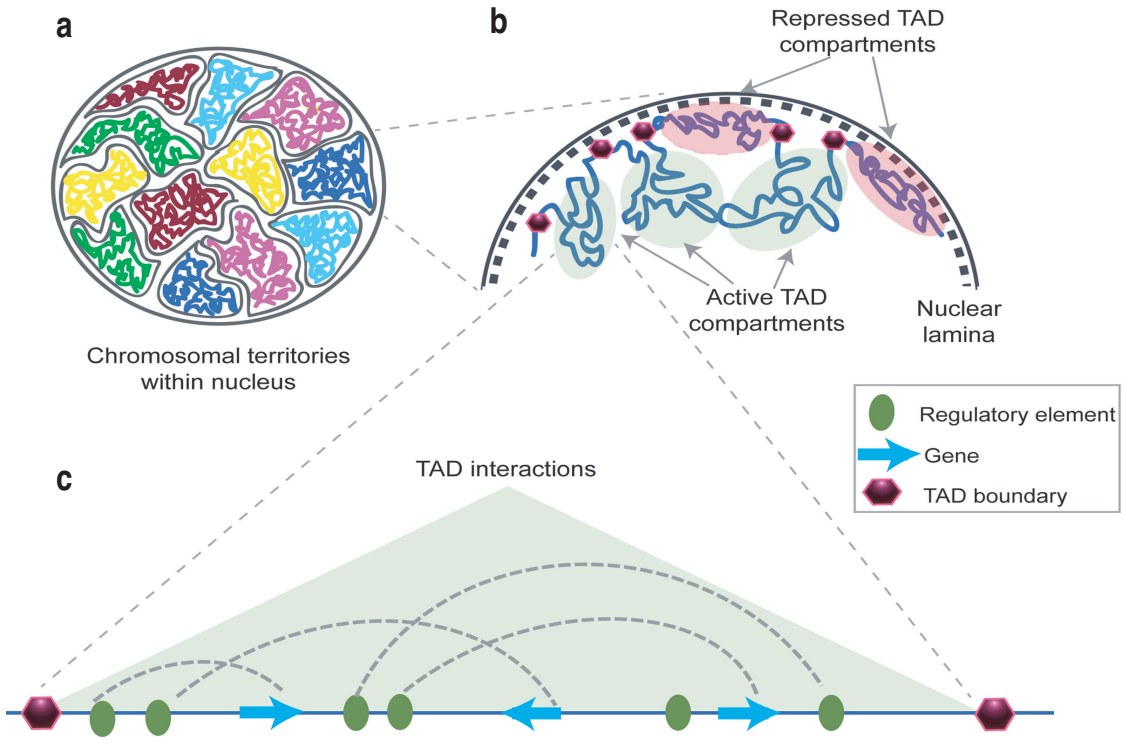


Fig. 1.1: Structural organization of the genome.

(a) Eukaryotic chromosomes occupy distinct territories in the interphase nucleus (Cremer et al., 2001; Cremer and Cremer, 2006). **(b)** Expression regulation on each chromosome is partitioned between the A- and B-type compartments. These compartments characterised by Lieberman-Aiden et al. (2009) are associated with euchromatin and heterochromatin. Type-A compartments are found towards the centre of the nucleus. Type-B compartments are associated with low transcriptional activity and their topologically associating domains (TADs) tend to congregate at or near the nuclear lamina. **(c)** Schematic of an active TAD. Dashed lines indicate looping interactions within the domain between distal CREs (green ovals) and genes (blue arrows). Figure taken from Matharu and Ahituv (2015).

The Chromatin Conformation Capture (3C) technique, developed by Jaob Dekker *et al.* (2002), and its derivatives have provided us with unprecedented access to the spatial organisation of eukaryotic genomes and its impact on transcription (de Wit and de Laat, 2012; Lajoie et al., 2015). These techniques are based on the principle that regions of chromatin that are in close proximity in 3D-space can be captured by performing formaldehyde cross-linking, a restriction digest and fragment re-ligation. 3C itself is limited to validating suspected interactions between pairs of restriction fragments ('one vs. one') but when coupled with NGS, 3C-based techniques can detect many novel chromatin interactions at kb resolution in a unbiased manner (Dixon et al., 2012; Dostie et al., 2006; Lieberman-Aiden et al.,

2009; Rao et al., 2014; Zhao et al., 2006). Hi-C, an 'all vs. all' approach, in particular has revealed the principle that chromatin-chromatin interactions are pervasive. Using this technique, Lieberman-Aiden et al. (2009) confirmed the existence of chromosome territories and demonstrated that euchromatin and heterochromatin are segregated in two genome-wide compartments, A and B. Further studies with higher resolution Hi-C have revealed that these genome compartments are organised into ~1 megabase (Mb) long regions of highly self-interacting chromatin known as TADs (Dixon et al., 2012; Nora et al., 2012) (Fig. 1.1c). Looping interactions within these domains frequently link enhancers and promoter regions (Rao et al., 2014) while insulator CREs and CCCTC-binding factor (CTCF) sites coincide with their boundaries. These domains are stable across different cell-types and conserved between species (Dixon et al., 2012). While only initially detected in animals, TADs are also found in plant and fungal genomes (Eser et al., 2017; Liu et al., 2017a).

Chromatin looping interactions with enhancer(s) facilitate and stabilise the assembly of the PIC at promoters (Allen and Taatjes, 2015; Plaschka et al., 2015; Soutourina, 2018), see **Fig. 1.2**. The binding of pioneer transcription factors (TFs) to enhancers recruits chromatin remodellers and co-activators that allow further TF binding (Chan and La Thangue, 2001; Yudkovsky et al., 1999).

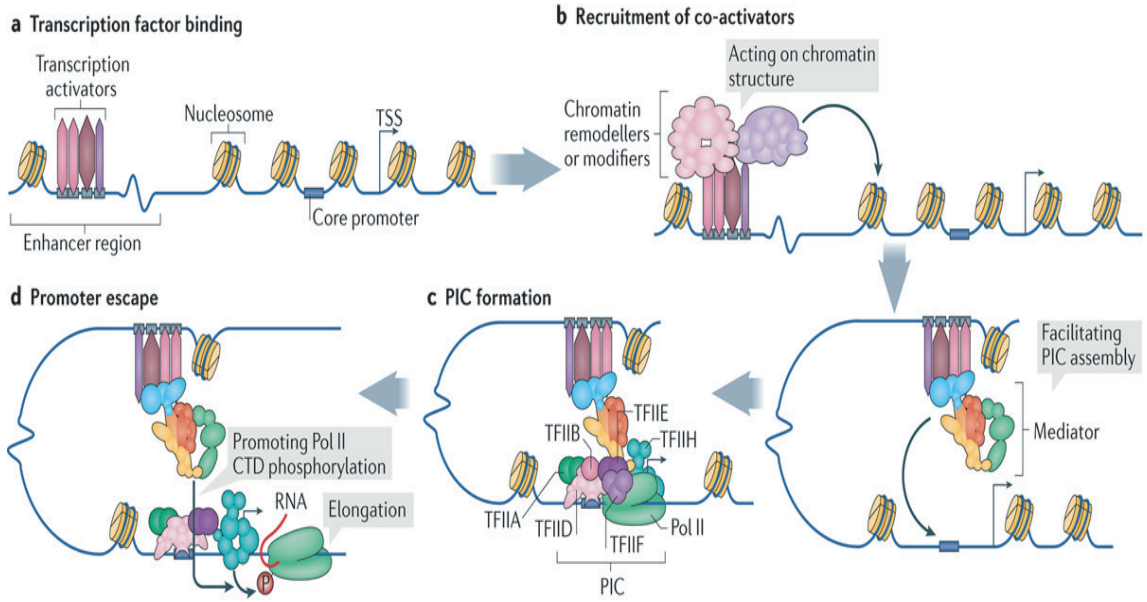


Fig. 1.2: Regulation of RNA polymerase II transcriptional initiation.

(a) Transcription factor binding to distal enhancer elements. Enhancers are typically a few hundred base-pair (bp) in length. They are able to influence their targets independently of their genomic location and orientation (Banerji et al., 1981). Enhancers rarely interact with their nearest gene (Sanyal et al., 2012).

Fig. 1.2: Regulation of RNA polymerase II transcriptional initiation (continued).
(b) Pioneer transcription factors facilitate the binding of other TF including the Mediator complex. Mediator interacts with cohesin to facilitate promoter-enhancer looping (Kagey et al., 2010). **(c)** The Mediator complex serves as a scaffold for the nucleation of pre-initiation complex subunits. **(d)** Mediator interacts with RNA Polymerase II CTD and triggers promoter escape. Mediator facilitates phosphorylation of the RNAP II carboxyl-terminal domain (CTD) (Plaschka et al., 2015). Figure taken from Soutourina (2018).

Enhancer-bound TFs recruit the Mediator complex (Poss et al., 2013; Soutourina, 2018) (Fig. 1.2b). This interacts with cohesin to facilitate promoter-enhancer looping (Kagey et al., 2010). Mediator serves as a scaffold for PIC assembly (Plaschka et al., 2015; Poss et al., 2013) (Fig. 1.2c). Mediator interacts with the RNAP II CTD and facilitates 'promoter escape' and transcriptional elongation by promoting its phosphorylation (Plaschka et al., 2015) (Fig. 1.2d).

1.2 Eukaryotic Post-Transcriptional Regulation

The existence of RNA-mediated mechanisms of post-transcriptional expression regulation was first hinted at when Napoli et al. (1990) inadvertently blocked the Petunia anthocyanin biosynthesis pathway *in trans* while trying to increase its activity with a chalcone synthase transgene. This resulted in a 50-fold decrease in endogenous chalcone synthase mRNA levels. Fire et al. (1998) later showed that dsRNA can trigger the homology-dependent RNAi response that dramatically reduces mRNAs levels of endogenous target genes. The fact that only exonic sequences triggered a RNAi response confirms that this does not target pre-mRNAs and is a post-transcriptional pathway. Subsequent studies have revealed additional conserved mechanisms of RNA post-transcriptional regulation, most notably the miRNA mediated silencing pathway (Lagos-Quintana et al., 2001; Lau et al., 2001).

1.2.1 Regulation of cytoplasmic mRNA degradation

Transcript steady-state abundances are lower than the quantities predicted purely from their rates of transcription and also reflect the contribution of their rates of turnover (Maekawa et al., 2015). The cytoplasmic stability of eukaryotic mRNAs, as measured by their half-lives, is influenced by many general and transcript-specific decay factors. The most important determinants are 3′-end polyadenylation (Chang et al., 2014), the formation of the N^7-methylguanosine (m^7G) 5′-cap (Mukherjee et al., 2012) and the presence of secondary structure (Geisberg et al., 2014) and CREs in the 3′-untranslated regions (UTRs) (Bartel, 2018).

The 3′-end processing of nuclear pre-mRNAs is indispensable for the generation of mature mRNAs (Dominski and Marzluff, 1999; Zhao et al., 1999). The RNAP II CTD is responsible for coordinating this and the other aspects of mRNA co-transcriptional processing (Custódio and Carmo-Fonseca, 2016) (Fig. 1.3a). polyadenylation signal (PAS) sequences are a defining feature of eukaryotic RNAP II transcribed genes (Proudfoot, 2011). With the exception of histone mRNAs, the vast majority of mammalian pre-mRNAs undergo 3′ polyadenylation, the non-templated addition of adenosine monophosphates, with polyadenylate polymerase (PAP) (Davila Lopez and Samuelsson, 2007). This is coupled to the endonucleotlytic cleavage by the Cleavage and polyadenylation specificity factor (CPSF) of a UG-rich sequence 10-30 bp downstream of the hexanucleotide PAS AAUAAA motif (Clerici et al., 2018; Haddad et al., 2012). Some messengers possess multiple PAS sequences and the final sequence of the 3′-end will depend on the chosen site of cleavage (Ransom et al., 2008) (Fig. 1.3b). This alternative polyadenylation (APA) will impact upon the formation of RNA secondary structure and the *cis*-regulatory sequences present in the 3′-UTR. Furthermore, cryptic PAS

motifs can be found within introns and the activation of these sites can result in truncated transcripts and proteins (Early et al., 1980; Rogers et al., 1980; Singh et al., 2018) (Fig. 1.3c). The spliceosomal U1 small nuclear ribonucleoprotein (RNP) normally suppresses intronic and premature PAS sites (Gunderson et al., 1998; Langemeier et al., 2012). The typically 150-250 nt long mammalian polyadenylate (poly(A)) tail is synthesised in one processive step (Brawerman, 1981; Darnell et al., 1971; Kühn et al., 2009).

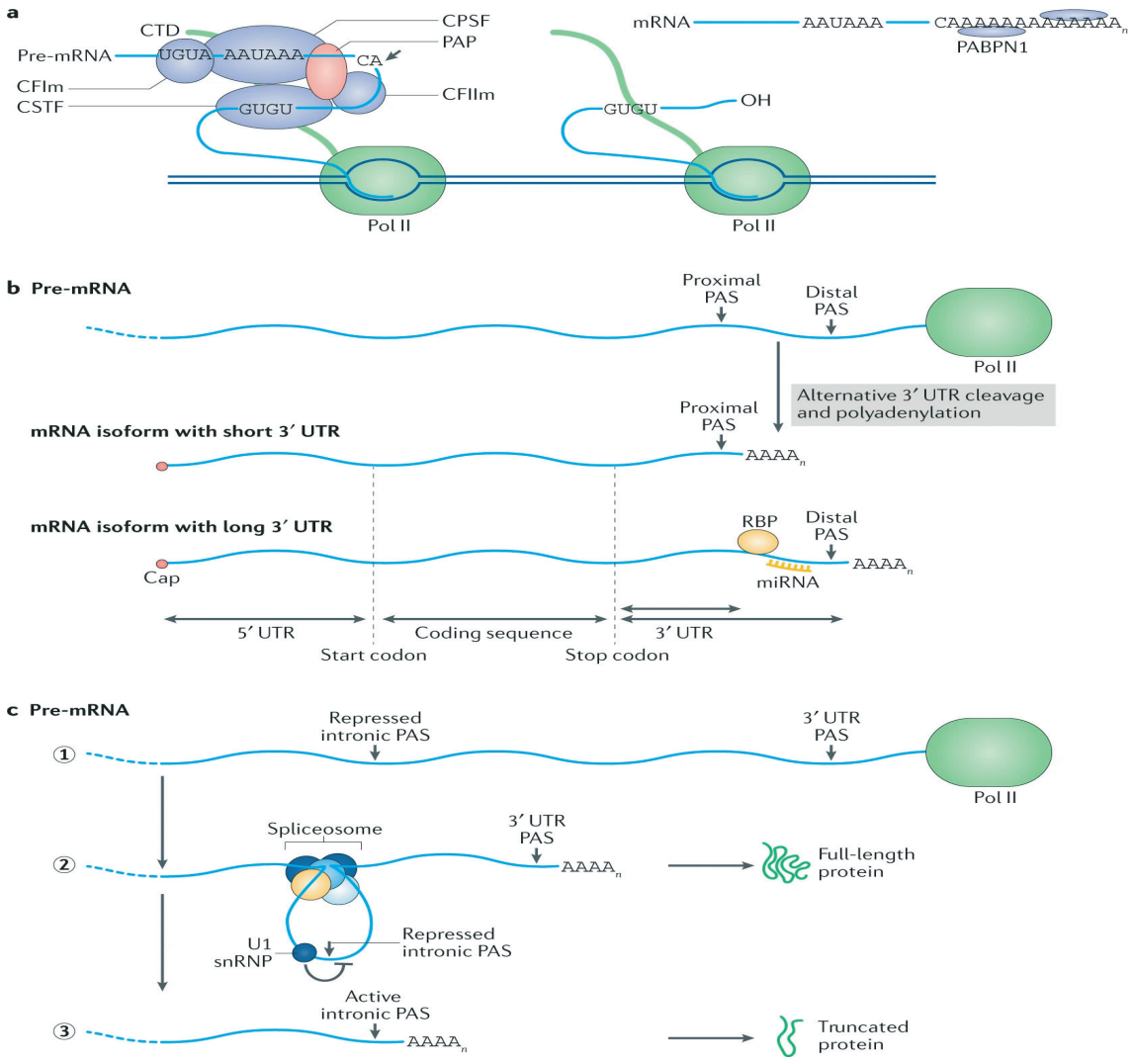


Fig. 1.3: Transcript 3′-end processing.

(a) Mammalian mRNA 3′-ends are formed from the endonucleolytic cleavage of the nascent transcript by CPSF that is coupled to PAP polyadenylation. CPSF specifically recognises the canonical PAS AAUAAA motif and recruits PAP (Murthy and Manley, 1995). Polyadenylate-binding nuclear protein 1 (PABPN1) associates with the nascent poly(A)-tail when it is 10-12 nucleotides long and stimulates further PAP activity (Kühn et al., 2009).

Fig. 1.3: Transcript 3′-end processing (continued).
(b) Usage of alternative proximal or distal PAS sites can result in different transcript isoforms. This affects which *cis*-regulatory sequences will be present in the 3′-UTR. **(c)** Cryptic PAS sites are also found within introns. PAS sites located near the transcript 5′-end are usually silenced to prevent premature 3′-end processing (Guo et al., 2011). Figure taken from Desterro et al. (2019).

poly(A)-tails facilitate mRNA nuclear export (Fuke and Ohno, 2008) but the main purpose of poly(A)-tails is to stabilise the rate of translation by enhancing its initiation (Meijer et al., 2007; Richter, 2000; Wakiyama et al., 2000) and protecting mRNAs from precocious degradation (Lim et al., 2014). The presence of a poly(A)-tail stabilises transcripts by inhibiting both of the two main mRNA decay pathways, 3′→5′ exonucleolytic degradation and 5′ m^7G-cap hydrolysis 5′→3′ degradation (Eckmann et al., 2011; Muhlrad et al., 1994). As a result most mRNA degradation pathways involve a deadenylation step (Beilharz et al., 2009; Muhlrad et al., 1994; Mukherjee et al., 2002; Yamashita et al., 2005). 3′ deadenylation is the rate-limiting step in cytoplasmic mRNA degradation (Chen and Shyu, 2011). An early indication of the protection afforded by poly(A)-tails was observed when RNase digestions failed to fully degrade polysome-associated mRNAs (Edmonds et al., 1971; Lim and Canellakis, 1970). In contrast, transcript polyadenylation promotes degradation in bacteria (Steege, 2000).
Although nascent mammalian poly(A)-tails are 150-250 nt long, the NGS TAIL-seq method for measuring tail lengths has revealed that there are populations of cytoplasmic transcripts with far shorter tails (50–100 nt) (Chang et al., 2014). Many housekeeping genes appear to have short tails (Subtelny et al., 2014). Furthermore, it has been suggested that the poly(A)-tails of certain transcripts change dynamically with the circadian rhythm (Kojima et al., 2012). Traditionally, mRNA deadenylation in the processing bodies (P-bodies) is viewed as a biphasic process (Bresson and Tollervey, 2018; Chen and Shyu, 2011; Wolf and Passmore, 2014). The first phase involves rapid deadenylation with the poly(A)-exonuclease PAN2/3 complex until protective the cytoplasmic Polyadenylate-binding cytoplasmic protein 1 (PABPC1) is no longer able to bind tails and the remaining residues removed by CCR4-NOT complex. In recent years this has been shown be an overly simplistic view of deadenylation regulation. Using TAIL-seq, Lim et al. (2014) and Morgan et al. (2017) have shown that post-transcriptional 3′ oligo-uridylation represents another layer in the regulation of tail lengths. It rings the death knell for transcripts with short tails (<25 nt) (Fig. 1.4b). This terminal uridylation is regulated in part by miRNAs. Yi et al. (2018) and Webster et al. (2018) recently revealed that the while PAN2/3 trims excessively long tails, this has relatively little impact on mRNA stability. Furthermore, CCR4-NOT complex is the main non-specific deadenylase.

Its CAF1 subunit deadenylates PABPC1-free tails while CCR4 removes PABPC1-bound tails (Webster et al., 2018; Yi et al., 2018) (Fig. 1.4a). Additionally, PABPC1 can facilitate mRNA deadenylation but also prevents precocious terminal uridylation-mediated decay (Yi et al., 2018). An mRNA is vulnerable to decay pathways after the initial rapid deadenylation of long poly(A)-tails by PAN2/3, followed by the slow iterative deadenylation by CAF1/CCR4 and a final 3′ oligo-uridylation by terminal uridyl transferases (TUTases) 4 and 7 (Fig. 1.4b).

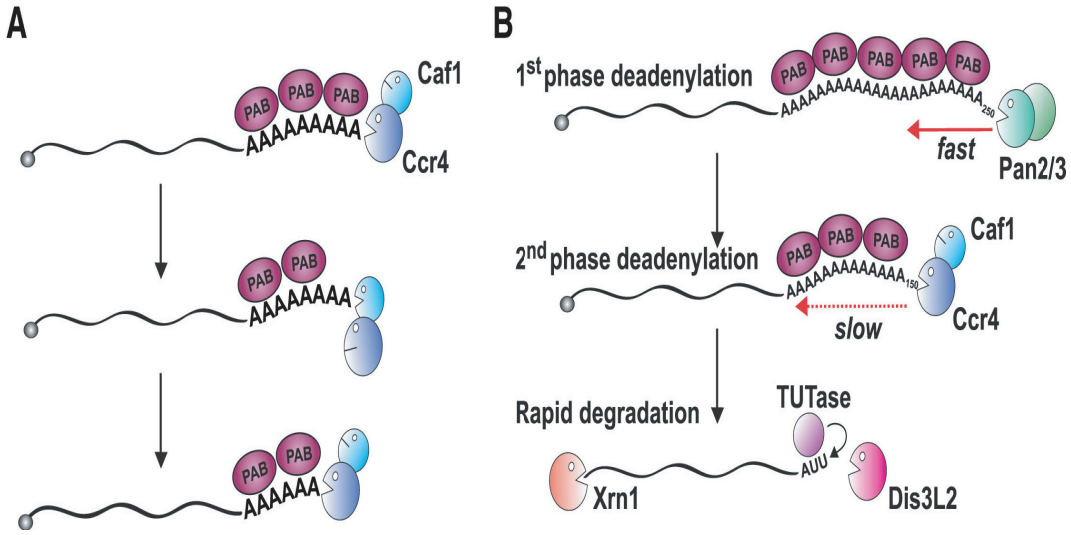


Fig. 1.4: Models for mammalian cytoplasmic mRNA deadenylation.
(a) CCR4-NOT complex-mediated deadenylation. Cytoplasmic mRNA poly(A)-tails are bound by multiple copies of PABPC1. These can activate CCR4-mediated deadenylation. Deadenylation removes PABPC1 binding sites revealing adensoine monomers. Free 3′ poly(A) is the ideal substrate for CAF1. CAF1 deadenylation removes free adensosines but it is inhibited until CCR4 removes the next PABPC1 monomer. **(b)** Full Kim/Passmore model for cytoplasmic deadenylation. Cytoplasmic poly(A) is rapidly shortened to ~150 nt by PAN2/3. This is followed by iterative cycles of CCR4/CAF1 deadenylation as in **(a)**. In the absence of PABPC1-binding, short tails (<25 nt) can be targeted for terminal uridylation-mediated decay via 3′ exonuclease decay with DIS3L2 or decapping and degradation by the 5′ exonuclease XRN1. Figure taken from Bresson and Tollervey (2018).

1.2.2 | mRNA N^6-methyladenosine methylation

RNA molecules are not homogeneous, linear chains of adenines (As), cytosines (Cs), guanines (Gs) and uracils (Us), many are post-transcriptionally modified and edited (Frye et al., 2016; Gott and Emeson, 2000; Grozhik and Jaffrey, 2018; Pan, 2018; Vitsios and Enright, 2015). As we have already seen in the case of 3′ terminal oligo-uridylation, RPTMs are important regulators of expression and RNA-associated activities (Lim et al., 2014; Morgan

et al., 2017; Saletore et al., 2012). Some have termed these RNA epigenetic or *epitranscriptomic* modifications given the parallels with DNA methylation (Liu and Pan, 2015; Meyer et al., 2012; Saletore et al., 2012), but I will avoid these terms due to some of the connotations attached to epigenetics. RPTMs have a long history with the first non-canonical ribonucleosides being identified in the 1960s (Cohn, 1960; Dunn, 1960). Indeed the first full RNA sequence determined, that of the *S. cerevisiae* alanine tRNA, contained 10 non-canonical bases (Holley et al., 1965).

Over 150 different RPTMs have been identified so far in cellular RNAs (Roundtree et al., 2017; Shi et al., 2019), **Fig. 1.5** shows examples of commonly detected RPTMs. N^1-methyladenosine (m^1A) and m^7G have positive electrostatic charges under physiological conditions (Agris, 1996). m^7G-modified miRNAs precursors have a greater propensity to form G-quadruplex structures (Pandolfini et al., 2019) but presence of m^7G disrupts their formation and facilitates their miRNA processing. Adenosine deamination to inosine affects base pairing and has the potential to recode mRNA codons (Gott and Emeson, 2000). The isomerisation of uridine to pseudouridine stabilises RNA secondary structure affecting RNA–RNA and RNA–protein interactions (Carlile et al., 2014). These few examples already illustrate that RNA chemical modifications can affect the properties of their RNA molecules.

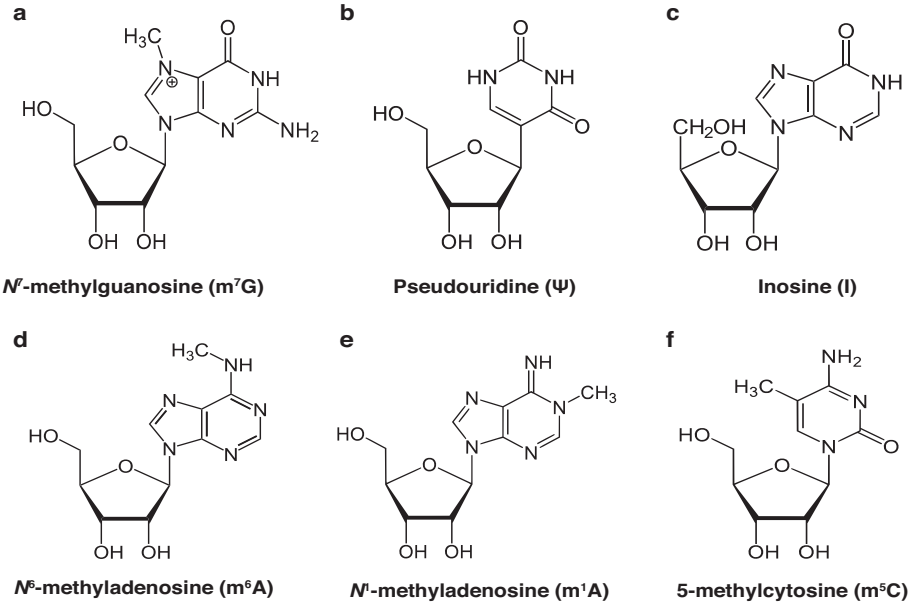


Fig. 1.5: RNA post-transcriptional modifications.

Chemical structures for a selection of commonly detected modified nucleosides. **(a)** N^7-methylguanosine, **(b)** Pseudouridine, **(c)** Inosine, **(d)** N^6-methyladenosine, **(e)** N^1-methyladenosine, **(f)** 5-methylcytosine.

The standard approach for detecting many covalent RPTMs is to perform antibody-based RNA-immunoprecipitation sequencing (RIP-seq) and standard cDNA sequencing experiments in parallel followed by enrichment analysis with a peak-calling algorithm like *MACS* (Dominissini et al., 2012, 2016; Meyer et al., 2012). Although m^1A has been reported in transcript 5′-UTRs, there is a strong suspicion that the antibody used for m^1A RIP-seq cross-reacts with m^7G in the 5′-cap (Grozhik et al., 2019; Safra et al., 2017).
m^6A is the most common internal modification to mammalian transcripts (Bokar, 2005). It was first detected in mRNAs using chromatography in the 1970s (Desrosiers et al., 1974; Perry et al., 1975). m^6A comprises ~0.5% of all cellular adenines and accounts for over half of all methylated ribonucleotides (Wei et al., 1975). Most mRNA m^6A is deposited co-transcriptionally by a METTL3-METTL14 heterodimer core (Bokar et al., 1997; Ke et al., 2017; Liu et al., 2013) in a complex containing multiple regulatory and adaptor subunits. Known subunits include WTAP, VIRMA, ZC3H13, HAKAI and RBM15 (Liu et al., 2013, 2018; Patil et al., 2016; Wen et al., 2018) (Fig. 1.6a). METTL3-METTL14 deposit m^6A in the context of a D/RR(m^6A)CH sequence motif (D = A, G or U, R = A or G; H = A, C, or U) (Dominissini et al., 2012; Meyer et al., 2012). Furthermore, an additional RNA m^6A methylase METTL16 has been shown to selectively methylate certain pre-mRNAs (Mendel et al., 2018; Warda et al., 2017).

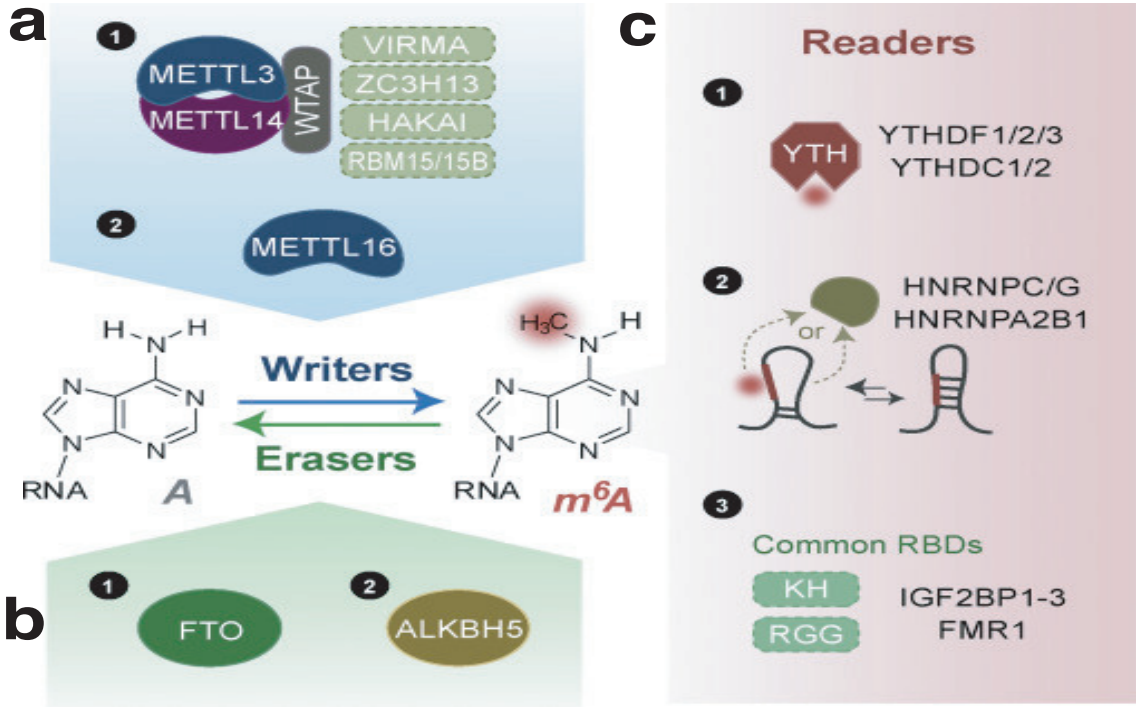


Fig. 1.6: Summary of m^6A Readers, Writers and Erasers.

Fig. 1.6: Summary of m^{6}A Readers, Writers and Erasers (continued).
(a) m^{6}A writers. The vast majority of m^{6}A methylation on mRNA is written by a METTL3-METTL14 core in association with WTAP, VIRMA, etc. in a sequence specific manner. Another writer METTL16 has been shown to deposit m^{6}A in *Mat2a* pre-mRNA hairpin structures in the context of a UAC(m^{6}A)GAGAA motif (Mendel et al., 2018). m^{6}A methylation is also regulated epigenetically, METTL14 directly binds to histone H3 lysine 36 trimethylation and is associated with transcriptional elongation (Huang et al., 2019). Additionally, certain miRNAs enhance methylation by guiding METTL3 binding (Chen et al., 2015). **(b)** m^{6}A erasers. FTO and ALKBH5 are the only characterised m^{6}A RNA demethylases in vertebrates. Evidence suggests that ALKBH5 is the main m^{6}A demethylase under physiological conditions (Mauer et al., 2017). **(c)** m^{6}A readers. YT521-B homology (YTH) domain containing proteins (YTHDF1-3, YTHDC1-2) directly recognise m^{6}A via their YTH-domains. The m^{6}A-mediated disruption of RNA secondary structure facilitates mRNA recognition by certain heterogeneous nuclear ribonucleoproteins (HNRNPs) (Alarcón et al., 2015; Liu et al., 2015, 2017b). Evidence suggests that some well characterised mRNA binding proteins, such as IGF2BP1-3 and FMR1, are in fact m^{6}A-readers (Edupuganti et al., 2017; Huang et al., 2018). Figure taken from Shi et al. (2019).

Although m^{6}A methylation seemed to be essential for mammalian cells (Bokar, 2005), it was largely ignored until it was determined that it is a reversible modification that can be erased by the RNA demethylase, FTO (Jia et al., 2011) (Fig. 1.6b). Subsequent studies identified a second RNA m^{6}A demethylase ALKBH5 (Zheng et al., 2013). ALKBH5 is the main m^{6}A demethylase *in vivo* while FTO is a facultative m^{6}A demethylase *in vitro* (Mauer et al., 2017). The existence of m^{6}A-regulating enzymes suggested that m^{6}A has a biological function.
The development of reliable high-throughput NGS-techniques for detecting m^{6}A by Dominissini et al. (2012) and Meyer et al. (2012) has allowed the m^{6}A methylome to be interrogated in unprecedented detail. They demonstrated that m^{6}A is found along the length of transcripts but it is particularly abundant towards their 3′-ends. Furthermore, although m^{6}A is usually found in the context of an D/RR(m^{6}A)CH motif, the majority of motif sites will be unmethylated at any given time (Dominissini et al., 2012; Meyer et al., 2012). Similar to the situation for m^{1}A, it is thought that the enrichment of m^{6}A near TSSs, initially reported by Dominissini et al. (2012); Meyer et al. (2012), instead reflects antibody cross-reactivity with the related RPTM N^6,2′-O-dimethyladenosine (m^{6}Am) (Linder et al., 2015). Using the single-nucleotide resolution UV-crosslinking and immunoprecipitation (CLIP) seq technique, Ke et al. (2015) showed that m^{6}A is enriched in the final exon particularly within the 3′-UTR. While m^{6}A is reversible, there is debate over the extent to which it is a dynamic RPTM (Ke et al., 2017; Meyer and Jaffrey, 2014; Rosa-Mercado et al., 2017; Roundtree et al., 2017). Most m^{6}A is added co-transcriptionally prior to splicing (Ke et al., 2017) and many sites are constitutive across tissues (Schwartz et al., 2014) but there is evidence that m^{6}A methylation is altered as a part of the heat shock response (Zhou et al., 2015).

m^{6}A methylation has been shown to affect numerous aspects of transcriptional and post-transcriptional regulation ranging from splicing, APA, RNA localisation, translation to stability and degradation (Edupuganti et al., 2017; Geula et al., 2015; Kasowitz et al., 2018; Louloupi et al., 2018; Meyer et al., 2015; Wang et al., 2013) (Fig. 1.7). The m^{6}A-mediated disruption of RNA secondary structure facilitates mRNA recognition by certain HNRNPs (Alarcón et al., 2015; Liu et al., 2015, 2017b). Evidence suggests that some well characterised mRNA binding proteins, such as IGF2BP1-3 and FMR1, are in fact m^{6}A-readers (Edupuganti et al., 2017; Huang et al., 2018). However, the best characterised m^{6}A readers in vertebrates are the YTH-domain containing proteins (Xu et al., 2014; Zhu et al., 2014).

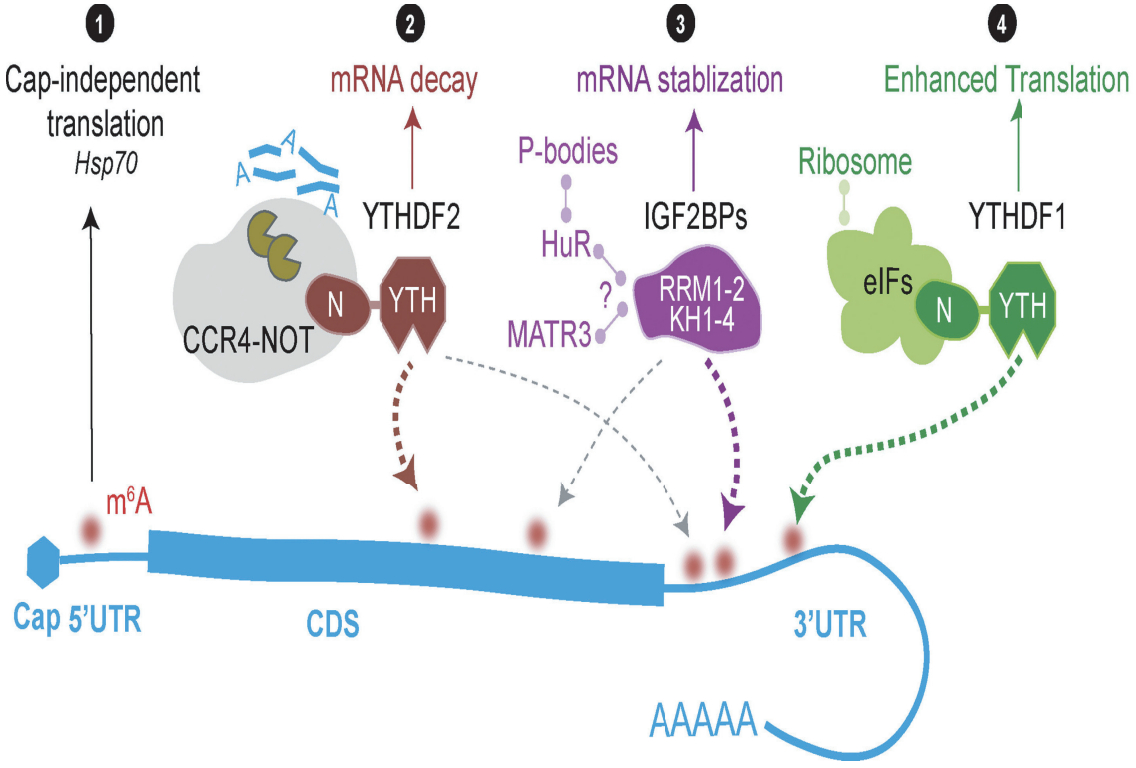


Fig. 1.7: Context-dependent functions of m^{6}A methylation in mRNA.
The location of m^{6}A within a transcript can determine its fate. m^{6}A in 5′-UTRs has been linked to conditional cap-independent translation. Whereas YTHDF2 destabilises transcripts by recruiting the CCR4-NOT complex, IGF2BP1-3 in association with HuR and MATR3 co-factors stabilise mRNA. IGF2BP1-3 preferentially bind to 3′-UTR m^{6}As (Huang et al., 2018) and it is possible that differential binding affinities based on the location of m^{6}A determines the impact of methylation on transcript stability. YTHDF1 also binds to 3′-UTR m^{6}A and promotes translation by interacting with initiation factors. Figure modified from Shi et al. (2019).

YTH m^{6}A RNA binding protein 2 (YTHDF2) is the major effector of m^{6}A-regulated transcript deadenylation (Geula et al., 2015; Wang et al., 2013, 2014; Zaccara et al., 2019). YTHDF2-bound transcripts are directed to cytoplasmic P-bodies where YTHDF2 recruits the CCR4-NOT deadenylase complex via an interaction with the SH-domain of NOT1 (Du et al., 2016). Deadenylated transcripts undergo $3'\rightarrow5'$ exonuclease-mediated decay. The localisation of transcripts in the P-bodies limits their potential translation by sequestering them away from the ribosomes. There may be some element of cooperative binding in the recruitment of CCR4-NOT; in general the greater the number of m^{6}As per transcript the shorter its half-life (Ke et al., 2017). In general, mRNAs with a high rate of turnover and multiple m^{6}A modifications are associated with regulatory functions while those more stable transcripts devoid of m^{6}A are associated with general 'housekeeping' functions. This m^{6}A-mediated decay (m^{6}AMD) regulates the transcriptional shifts necessary for the vertebrate MZT (Ivanova et al., 2017; Zhao et al., 2017) and endothelial-to-haematopoetic transition (EHT) (Lv et al., 2018). It facilitates the transition from a pluripotent to differentiating state in embryonic stem cells (mESCs) (Geula et al., 2015). As some transcriptional programmes in vertebrate embryonic development have parallels in tumourigenesis (Aiello and Stanger, 2016; Youssef et al., 2012), it is unsurprising that YTHDF2 expression is clinically relevant for certain cancer types (Chen et al., 2017, 2018; Paris et al., 2019; Yang et al., 2017). The role of YTHDF2 in regulating transcriptional dynamics during the murine MZT is discussed in **Chapter 4**.

1.2.3 Direct detection of modifications using Native Nanopore RNA Sequencing

Although RNA sequencing was developed first (Fiers et al., 1976), it was the application of Fred Sanger's eponymous dideoxy chain-termination method for sequencing DNA (Sanger et al., 1977) to libraries of short cDNA fragments that allowed the first eukaryotic *transcriptomes* to be interrogated (Adams et al., 1991; Sim et al., 1979; Velculescu et al., 1997). However it not until the advent of more high-throughput technologies like DNA microarrays and NGS (McGettigan, 2013; Nelson, 2001) that the field of *Transcriptomics* truly emerged. The development of third-generation sequencing technologies such as the ONT nanopore sequencing (Garalde et al., 2018; Jain et al., 2015) and PacBio SMRT-seq methods (Eid et al., 2009) have the potential to innovate the field of transcriptomics (Niedringhaus et al., 2011). ONT nanopore-based sequencing platforms detect single DNA molecules without the need for enzymatic synthesis and amplification reactions. Their sequencing platform consists of individual protein nanopores embedded in arrays of thousands of synthetic polymer membranes in a single flowcell (Jain et al., 2015). Perturbations in the nanopore current are generated

when a single motor protein-bound DNA molecule is captured by a pore and pulled through at a consistent rate. These signal perturbation can be converted into base space using machine learning techniques like recurrent neural networks (Wick et al., 2019). The signal for a given base depends on the 5 nt k-mer context in which it is found (Jain et al., 2015). Recently, it has become possible to directly sequence poly(A)$^+$ RNA molecules by performing splint ligation with a poly(T) adaptor and an RNA-specific motor protein (Garalde et al., 2018) (Fig. 1.8a). As the adaptor-ligated RNA enters a pore, the adaptor oligomer generates a characteristic shift in the current which is followed by a flat signal indicative of the 3′ poly(A)-tail and finally the RNA itself (Fig. 1.8b). The signal is basecalled from the 3′-end with a slight decline in basecall quality score towards the 5′-end (Fig. 1.8c). The basecall accuracy for RNA with Guppy, the main ONT basecaller, is currently >90% (Garalde et al., 2018; Wick et al., 2019). As poly(A)-tails have such a clearly defined signal in Nanopore RNA 'squiggles', it was possible for Jared Simpson to generate a tail length estimator from *in vitro* transcribed (IVT) *S. cerevisiae* transcripts with poly(A)-tails of known length (Workman et al., 2018). Nanopore direct RNA-seq generates strand-specific, long transcript sequences without any of the biases introduced by polymerase chain reaction (PCR) amplification or reverse transcription (RT) that normally affect cDNA based NGS experiments (Kozarewa et al., 2009).

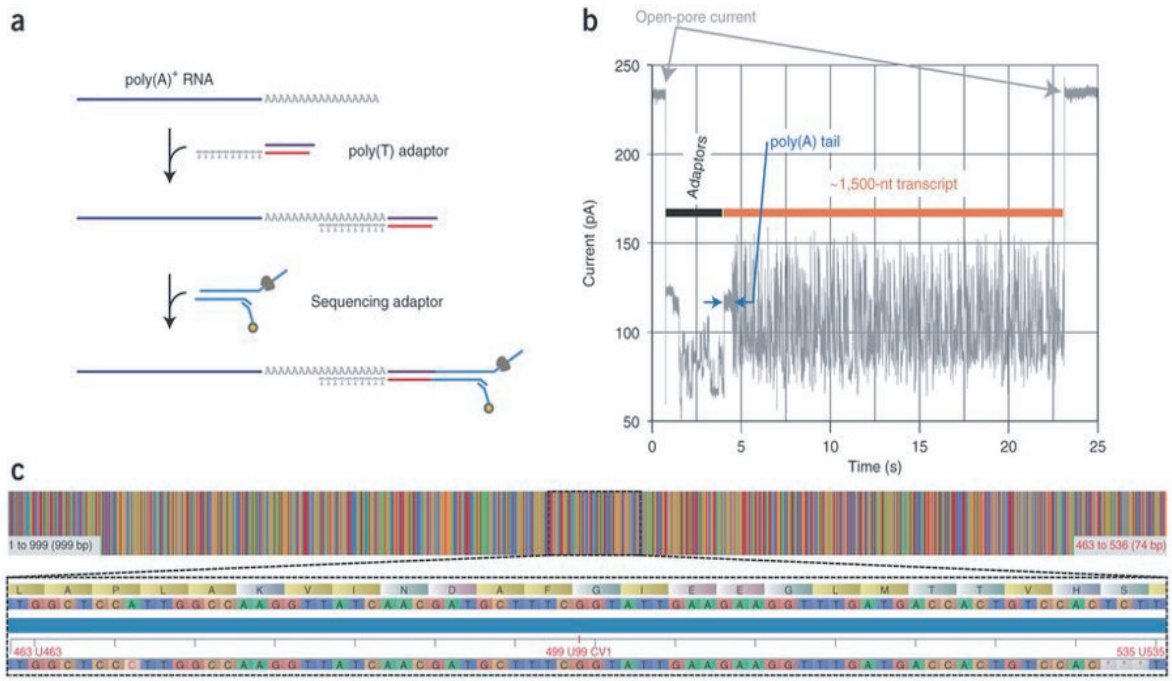


Fig. 1.8: Oxford Nanopore Technologies native or direct RNA-sequencing.

Fig. 1.8: Oxford Nanopore Technologies native or direct RNA-sequencing (continued).
(a) Library preparation method for Nanopore direct RNA-seq. **(b)** Representative raw data 'squiggle' from the translocation of a single transcript through a pore. **(c)** Alignment of a basecalled *S. cerevisiae* transcript against the reference transcriptome. Figure taken from Garalde et al. (2018).

Over 500 studies have profiled the m^6A-methylome (Zaccara et al., 2019). However, all current NGS methods for detecting m^6A in mRNAs are indirect. The two most widely used techniques for detecting m^6A, RIP-seq (Dominissini et al., 2012; Meyer et al., 2012) and CLIP-seq (Ke et al., 2015; Linder et al., 2015), involve the use of m^6A-recognising antibodies and as we have already seen, one of these antibodies also reacts with m^6Am. Neither of these families of techniques are entirely satisfactory. RIP-seq requires the sequencing of input material to call m^6A peaks and is prone to false positives (Zaccara et al., 2019). It is not possible to detect which strand the peak is on or determine how many m^6A sites are in each peak (Grozhik et al., 2017; Linder et al., 2015). While CLIP-seq is able to map m^6A at single-nucleotide resolution, you lose any information about the absolute number of modified sites. It is a labourious process that requires large quantities of RNA. Both of these techniques are affected by the biases introduced by PCR amplification and RT. In contrast, Nanopore direct RNA-seq holds the potential to directly identify m^6A from RNA molecules without the loss of information from the synthesis and amplification of cDNA. Garalde et al. (2018) showed that it possible to distinguish m^6A from adenine in synthetic RNA oligos based on subtle shifts in the pore current between modified and unmodified k-mers. Furthermore, (Liu et al., 2019) trained Support Vector Machines (SVMs) on all possible k-mers for the RRACH version of the METTL3/14 motif using IVT m^6A-modified and unmodified synthetic sequences. Their *EpiNano* algorithm is able to predict m^6A in *S. cerevisiae* transcripts with 87% accuracy. While these findings are very encouraging, it was perhaps unwise to restrict their SVM training set to the RRACH motif given that m^6A can be deposited in other contexts and by other RNA methylases (e.g. METTL16). Additionally, as the shifts in current between m^6A and A and between m^6A and m^6Am are very subtle it would have been logical to have included m^6Am-modified sequences in their training set. The ability to simultaneously interrogate poly(A)-tail lengths and the m^6A methylome is particularly useful for dissecting YTHDF2-regulated transcriptional dynamics. Nanopore native RNA-seq holds great promise for the direct detection of m^6A and other RPTMs but for the moment is held back by the need to generate vast training sets based on different RPTMs in various 5 nt k-mer contexts.

1.3 Transcriptional Noise

Stochasticity is an inherent feature of biological processes involving molecular interactions (Ecker et al., 2017). When components in a system are present in large quantities, random fluctuations have little impact on the overall system (Swain et al., 2002). However in a small system like a biological cell, molecules are generally found in relatively low abundances. Even amongst isogenic populations of cells a small, transient fluctuation in the quantity of an mRNA molecule has the potential to introduce non-genetic variation (Dong et al., 2011; Elowitz et al., 2002). *Transcriptional noise*, the unstructured cell-to-cell variation in transcript abundances, reflects the contributions of intrinsic and extrinsic factors (Swain et al., 2002; Valadares Barroso et al., 2018). Sources of intrinsic noise include gene-specific factors that influence transcriptional initiation like epigenetic modifications and the chromatin accessibility and nucleosome occupancy around TSSs (Buenrostro et al., 2015b; Radman-Livaja and Rando, 2010; Smallwood et al., 2014). Extrinsic noise encompasses cell-specific features like the cell cycle and the availability of regulators like RNAP and GTFs (Raser and O'Shea, 2005; Sherman et al., 2015; Valadares Barroso et al., 2018; Zopf et al., 2013). The transcript abundances of genes affected by extrinsic noise will covary and correlate. As some sources of extrinsic noise are also gene products they have the potential to further propagate noise (Sherman et al., 2015). Intrinsic noise is the more significant source of cell-to-cell transcriptional heterogeneity in mammals (Levesque and Raj, 2013; Raj et al., 2006). The reverse is the case for unicellular fungi (Raser and O'Shea, 2005; Sherman et al., 2015).

1.3.1 Detection of Transcriptional Noise

Classically, investigations into transcriptional dynamics have focussed on changes in mean expression levels. While it is possible to detect transcriptional heterogeneity from DNA microarray and bulk RNA-seq data (Hansen et al., 2011; Ho et al., 2008), it is with the development of reliable plate- and droplet-based protocols for performing scRNA-seq (Picelli et al., 2014; Zheng et al., 2017) that this facet of expression regulation can be reliably interrogated. With the advent of computational methods for detecting sources of technical noise in scRNA-seq (Brennecke et al., 2013; Vallejos et al., 2015, 2016) there is a growing appreciation that is cell-to-cell transcriptional heterogeneity is quantifiable and consequential. scRNA-seq generates individual transcriptional profiles that can be used to assess differences in mean gene expression and transcriptional variability in population of cells. This has allowed cell-to-cell transcriptional heterogeneity to be dissected at unprecedented resolution genome-wide and at the single-gene level (Goolam et al., 2016; Grün et al., 2014). Tools for analysing differential gene expression in bulk RNA-seq data such as *DESeq2* (Love

et al., 2014) and *edgeR* (Robinson et al., 2009) use a Negative Binomial Regression to model and normalise counts. The high gene 'dropout rate' and sparsity of counts per cell mean that the application of this model is not entirely appropriate for scRNA-seq data. Bayesian analysis of single-cell sequencing data (BASiCS) (Vallejos et al., 2015, 2016) uses a Bayesian probabilistic model based on the Poisson distribution to decompose expression variance into technical and true biological variability to generate 'denoised' counts. From this a gene-specific overdispersion estimate, the statistical tendency for observed cell-to-cell transcriptional variability to exceed what is predicted by Poisson sampling, is calculated. Other methods for assessing transcriptional heterogeneity have employed the coefficient of variation, squared coefficient of variation or the variance scaled by mean (Brennecke et al., 2013; Chen et al., 2016; Satija et al., 2015). Furthermore, BASiCS can correct for the statistical dependence of the variance on the mean typically observed for scRNA-seq data by fitting a regression trend between the gene overdispersion and mean expression (Eling et al., 2018). This allows for simultaneous differential testing of changes in mean expression and transcriptional variability.

1.3.2 Noise Control

There is mounting evidence that transcriptional variability is a necessary and regulated feature of cell populations (Antolović et al., 2017; Dueck et al., 2016). It is likely that plasticity and transcriptional noise are coupled and co-evolve (Lehner, 2010). The low-level, leaky expression from the *Eschercia coli lac* operon promoter (Jacob and Monod, 1961b) is a classic example of this "bet hedging". This phenomenon allows a cell to rapidly alter its metabolism in response to the presence of β-galactosides. In a pool of differentiating eukaryotic cells, heterogeneity allows progenitors to explore the cell fate space before committing to a particular decision or lineage (Dueck et al., 2016; Trapnell et al., 2014). Transcriptional variability increases in haematopoietic stem and progenitor cells (HSPCs) directly prior to state transitions (Mojtahedi et al., 2016). However, excessive heterogeneity in transcript abundance may compromise homoeostasis. To overcome deleterious effects of heterogeneity, transcriptional noise must either be regulated or integrated into the cell system (Dueck et al., 2016).

Steady-state transcript abundances in a cell are lower than the quantities predicted purely from their specific rates of transcription and also reflect the contribution of their rates of turnover (Maekawa et al., 2015). Both of these rates are potential sources of noise but can also be used to modulate transcriptional variability (Baudrimont et al., 2019; Dueck et al., 2016). It has been observed that most eukaryotic genes are transcribed periodically and experience large, transcriptional bursts (Hnisz et al., 2017; Larsson et al., 2019). As we have seen in

Section 1.1.1, the eukaryotic DNA genome is compacted into chromatin. Transcription from promoters is dependent on the assembly of a PIC in regions of accessible chromatin (Allen and Taatjes, 2015; Plaschka et al., 2015; Soutourina, 2018). The nucleation of RNAP II and GTF at promoters is enhanced by looping interactions with CREs that are regulated by the Mediator complex (Kagey et al., 2010; Poss et al., 2013). Evidence suggests that histone H3 lysine 27 acetylation in regions near TSSs regulates the frequency of transcriptional bursts (Nicolas et al., 2018) while chromatin looping interactions regulate the amplitude of these bursts (Hnisz et al., 2017). Histone H3 lysine 4 methylation is associated with both burst size and frequency (Wu et al., 2017). Short refractory periods between bursts reduce transcriptional noise (Harper et al., 2011). The presence of promoter CGIs is associated with reduced transcriptional noise (Faure et al., 2017) while the presence of TATA-box CREs increases it (Ravarani et al., 2016). Furthermore, weak and/or distal enhancer activity (Fukaya et al., 2016) and high nucleosome occupancy are associated with noisy transcription (Cairns, 2009).

In general, unstable mRNAs are more susceptible to noise than stable transcripts (Baudrimont et al., 2019). Cytoplasmic mRNA stability is anti-correlated with mRNA expression noise in mESCs (Faure et al., 2017). The cytoplasmic stability of eukaryotic mRNAs, as measured by their half-lives, is influenced by many general and transcript-specific decay factors. The most important determinants are $3'$-end polyadenylation (Chang et al., 2014), the formation of m^7G $5'$-caps (Mukherjee et al., 2012) and the presence of secondary structure (Geisberg et al., 2014) and *cis*-regulatory sequences in the $3'$ UTRs (Bartel, 2018). The presence of a poly(A)-tail stabilises transcripts by inhibiting both of the two main mRNA decay pathways, $3'\rightarrow5'$ exonucleolytic degradation and $5'$ m^7G-cap hydrolysis $5'\rightarrow3'$ degradation (Eckmann et al., 2011; Muhlrad et al., 1994). Most mRNA degradation pathways involve a deadenylation step (Beilharz et al., 2009; Muhlrad et al., 1994; Mukherjee et al., 2002; Yamashita et al., 2005). Notably one of these pathways, the miRNA-mediated decay (miRMD), directly impacts upon cell-to-cell heterogeneity (Gambardella et al., 2017; Schmiedel et al., 2015). miRNAs are short ($\sim$22nt) single-stranded ncRNAs molecules that when loaded into Argonaute (AGO) as part of the RNA-induced silencing complex (RISC), can specifically bind to mRNA $3'$-UTR *cis*-regulatory sequences complementary to their short $5'$ seed regions (Bartel, 2018). In animals, depending on the extent of complementarity, the transcript will either undergo deadenylation (Chen et al., 2009) or, more rarely, endonucleolytic cleavage (Xu et al., 2016; Yekta et al., 2004). miRNAs buffer expression variability during the noisy transcriptional shifts that facilitate differentiation as part of metazoan development (Posadas and Carthew, 2014; Siciliano et al., 2013). Schmiedel et al. (2015) have shown that miRNAs reduce expression noise for lowly expressed genes and that this effect is enhanced when multiple

miRNAs target the same transcript. This canalises expression and confers developmental robustness.

1.3.3 Transcriptional Noise in Biological Ageing

Organismal ageing or *senescence* is the time-dependent, progressive decline in biological function and is usually attributed to the accumulation of molecular damage over time (Gems and Partridge, 2013; López-Otín et al., 2013). Both genetic and epigenetic factors have been implicated in the senescence-associated dysregulation (Fig. 1.9). Ageing is associated with global DNA hypo- and more localised instances of hyper-methylation (Ciccarone et al., 2018; López-Otín et al., 2013). Furthermore, the abundance and distributions of various activating and repressive histone tail modifications also change with ageing (Fraga and Esteller, 2007; Han and Brunet, 2012). Additionally, the activity and efficiency of many DNA repair pathways declines with age as a result of the accumulation of molecular damage (Gorbunova et al., 2007). The impairment of these pathways leads to the further accrual of genetic alterations.

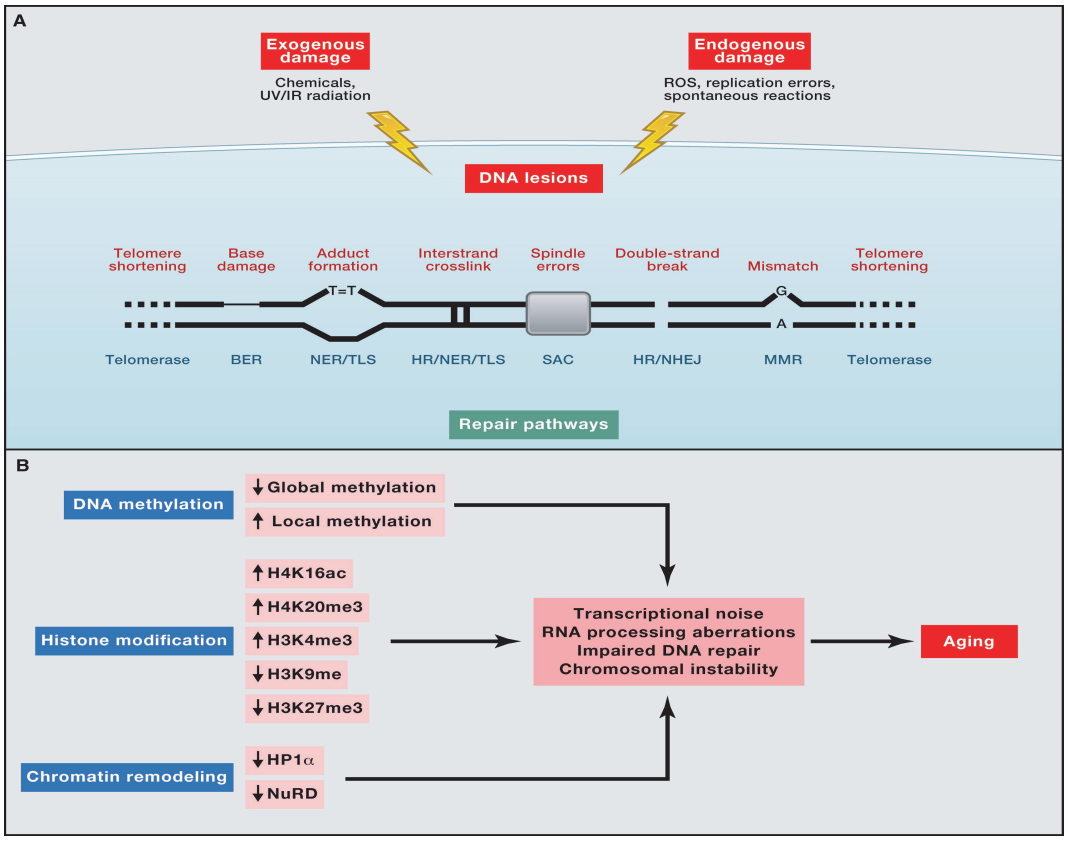


Fig. 1.9: Age-associated genetic and epigenetic alterations.

Fig. 1.9: Age-associated genetic and epigenetic alterations (continued).
(a) Nuclear DNA damage and repair. Endogenous and exogenous damage can lead the accrual of a various genetic lesions with increasing age. These are normally repaired through a variety of mechanisms. DNA damage and insufficient/inefficient repair promotes biological ageing (Gorbunova et al., 2007). Mitochondrial genomes also accumulate damage with increasing age. BER, base excision repair; HR, homologous recombination; NER, nucleotide excision repair; NHEJ, nonhomologous end-joining; MMR, mismatch repair; ROS, reactive oxygen species; TLS, translesion synthesis; SAC, spindle assembly checkpoint. **(b)** Epigenetic alterations. There is no definitive 'aged epigenome' but ageing is associated with a general global DNA hypo- and more localised instances of hyper-methylation (Ciccarone et al., 2018). The abundance and distributions of various activating and repressive histone tail modifications also change with age (Fraga and Esteller, 2007; Han and Brunet, 2012). These epigenetic changes directly affect for the recruitment of chromatin remodellers and impact upon chromatin acessibility.

Only a relatively small proportion of an organism's transcriptome is altered with increasing age (Stegeman and Weake, 2017). Most of these changes in mean expression are tissue- and/or cell-type specific. However, it has recently been shown that ageing increases cell-to-cell transcriptional heterogeneity in somatic tissues (Angelidis et al., 2019; Enge et al., 2017; Martinez-Jimenez et al., 2017). Martinez-Jimenez et al. (2017) profiled the affect of ageing on the transcriptomes of naïve and activated $CD4^+$ T cells from two *Mus musculus ssp*. The expression of genes related to the adaptive immune response increased more heterogeneously upon stimulation in aged cells from both sets of mice. This suggests that the less robust adaptive immune-response associated with ageing (Tsukamoto et al., 2009) can be partly explained by the less coordinated activation of the immune-related genes upon immune stimulation. Angelidis et al. (2019) observed a similar increase in cell-to-cell transcriptional heterogeneity with age across the cell-types present in the mouse lung, including for $CD4^+$ T cells. While mean gene expression is relatively stable with age, it is likely that altered transcriptional variability is hallmark of ageing in somatic tissues. Thus far, the impact of ageing on transcriptional heterogeneity has only been assessed in the soma. Thee impact of ageing and testicular injury and subsequent regeneration on transcriptional variability in mouse undifferentiated spermatogonia is explored in **Chapter 3**.

1.4 Outline

The aim of my PH.D. was to profile the impact of altered eukaryotic transcriptional regulation on gene expression as measured by bulk and single-cell RNA-sequencing.
Chapter 2 explores the involvement of high-risk HPV (HR-HPV) integration in alterations to chromatin conformation and gene expression in early cervical carcinogenesis using the W12 *in vitro* model. HPV16 integrants interact with host chromatin without disrupting 3D chromatin structure. Integrants affect the transcription of neighbouring and more distal genes on the same host chromosome. A manuscript is undergoing final preparations.
Chapter 3 presents a study into the affects of ageing and testicular damage and regeneration on gene mean expression and transcriptional noise in mouse undifferentiated spermatogonia. We see subtle changes in mean expression but a large decline in transcriptional variability with ageing and testicular regeneration. This is possibly linked to the phenomenon of "Selfish spermatogonial selection". A manuscript is undergoing final preparations.
Chapter 4 presents the results of an initial plate-based scRNA-seq experiment investigating the regulation of gene expression and transcriptional noise by YTHDF2, a reader protein for RNA N^6-methyladenosine. Gene expression and transcriptional noise increase in the absence of maternally supplied YTHDF2. This suggests that YTHDF2 has a role in buffering transcriptional noise in zygotes. A second follow-up experiment with droplet-based scRNA-seq was performed for mouse pre-LSCs.

1.5 Other contributions

Contributions to papers that are not discussed in this thesis are as follows:

Matthew P Davis, Claudia Carrieri, Harpreet Saini, Stijn van Dongen, Tommaso Leonardi, Giovanni Bussotti, **Jack M Monahan**, Tania Auchynnikava, Angelo Bitetti, Juri Rappsilber, Robin C Allshire, Alena Shkumatava, Dónal O'Carroll, Anton J Enright. Transposon-driven transcription is a conserved feature of vertebrate spermatogenesis and transcript evolution. *EMBO reports*, 18 (7), 1231-1247, 2017 ■

Ivayla Ivanova, Christian Much, Monica Di Giacomo, Chiara Azzi, Marcos Morgan, Pedro N Moreira, **Jack Monahan**, Claudia Carrieri, Anton J Enright, Dónal O'Carroll The RNA m6A reader YTHDF2 is essential for the post-transcriptional regulation of the maternal transcriptome and oocyte competence. *Molecular cell*, 67 (6), 1059-1067. e4, 2017 ■

Marcos Morgan, Yuka Kabayama, Christian Much, Ivayla Ivanova, Monica Di Giacomo, Tatsiana Auchynnikava, **Jack M Monahan**, Dimitrios M Vitsios, Lina Vasiliauskaitė, Stefano Comazzetto, Juri Rappsilber, Robin C Allshire, Bo Torben Porse, Anton J Enright, Dónal O'Carroll. A programmed wave of uridylation-primed mRNA degradation is essential for meiotic progression and mammalian spermatogenesis. *Cell research*, 29 (3), 221, 2019 ■

The github repository containing the source code of this thesis can be found here:
https://github.com/monahanj/phd_thesis

2

Altered transcriptional regulation in early cervical carcinogenesis

Declaration This work was a joint effort of the Coleman and Enright labs. Ian Groves and Nick Coleman conceived the study. Anton J. Enright and Nick Coleman supervised the study. Ian Groves, Emma Knight and Marco Michalski performed the experiments. I performed all computational analyses unless specified otherwise. The manuscript is undergoing final preparations.

2.1 Introduction

Cancer is the most common cause of premature mortality in developed and developing countries (Bray et al., 2018). Cancer encompasses a heterogeneous class of neoplastic diseases sharing a number of biological hallmarks defined in Hanahan and Weinberg (2000) and expanded upon in Hanahan and Weinberg (2011). A unifying feature of oncogenesis is the frequent occurrence of pre-malignant genetic changes that facilitate transformation as part of Knudson's Two-hit Hypothesis, see **Box 2.1**. This genomic instability provides genetic diversity that fuels emergent cancer hallmarks like the evasion of growth suppressors and sustained proliferation (Hanahan and Weinberg, 2011). Unsurprisingly these general and tumour-specific changes to the DNA have consequences for transcriptional regulation (Uhlen et al., 2017).

Box 2.1: Knudson's Two-hit hypothesis

Retinoblastoma, an ocular paediatric cancer, can develop bilaterally in both retinas or unilaterally in a single retina. Knudson (1971) observed distinctive dominant inheritance pattern for the bilateral form of the disease. Children with bilateral retinoblastoma were more likely to have had relatives with retinoblastoma than those with later onset and more sporadic, unilateral tumours. This suggested to him that the retinoblastoma inheritance pattern is more easily explained if pre-existing, deactivating germ-line mutation or 'hit' contributes to the emergence of retinoblastoma. These children would be susceptible to developing cancer as it would only require a second somatic 'hit', in any cell, at the other allele encoding what was termed the retinoblastoma protein (pRb). While in contrast, those without a family history would require two independent somatic hits to occur in the same cell, at both alleles of the *RB1* locus. From this he inferred that pRb must normally act to prevent the emergence of these tumours, thus identifying the first of what were later termed *tumour-suppressor* proteins. Knudson's Two-hit hypothesis holds true for the majority of tumour suppressor genes. Exceptions to the rule include genes encoding proteins that function in complexes and require a minimum dosage or those for which a deleterious mutation in either allele compromises the function of the entire complex. A notable example of the latter being *dominant-negative* mutations in the *TP53* locus that prevent the P53 homo-tetramer from binding to it target loci (Willis et al., 2004).

Global cancer prevalence continues to increase with improving living standards and the associated changes in lifespan and population growth (Bray et al., 2018). Differences in the availability and quality of healthcare mean that certain neoplasms are more common in less developed regions than higher incomes countries. As a population has better access to high quality healthcare the cancer burden in that population shifts (Ferlay et al., 2015). This reflects the greater provision of early detection programmes and prophylactic measures

against the more preventable and treatable cancer types. Cervical cancer is the fourth most common cancer-mortality of women in developed countries but it is second only to breast cancer in the least developed countries (Bray et al., 2018). While the incidence of cervical neoplasms declines in affluent countries, it continues to increase in the less developed regions of the world (Bray et al., 2018; Drolet et al., 2019; Ferlay et al., 2015). This disparity reflects the greater provision of prophylactic vaccines against the vector of cervical neoplasms, the human papillomavirus, and screening programmes for pre-cancerous cervical intraepithelial neoplasia (CIN) in countries with a high Human Development Index (HDI) (Drolet et al., 2019). A third of females aged 10–20 years in developed countries have received a HPV vaccination as compared to fewer than 3% in less developed countries (Drolet et al., 2019). The prevention of future HPV infections has potential to eliminate cervical cancer as well as other HPV-related neoplasms including oropharyngeal, anogenital and lung cancer subtypes (de Sanjosé et al., 2018).

Papillomaviruses infect the epithelia of amniotes and these infections can develop into benign warts or papillomas (McBride et al., 2012). This follows the abrasion or removal of the stratified epithelium layers above the basal cells (Roden and Stern, 2018). The aetiology of cervical malignancies lies with persistent HPV infections in the cervical basal keratinocytes (Roden and Stern, 2018). HPV, like all papillomaviruses, is a dsDNA virus that requires the cellular replication machinery available in these proliferating cells for its own replication (zur Hausen, 2002). Viral replication is dependent on the proliferation and differentiation of infected cells (Bedell et al., 1991; Kajitani et al., 2012). The 8 kb HPV genome typically consists of six *early* and two *late* open reading frames (ORFs) (Zheng, Zhi Ming, Baker, 2008; zur Hausen, 2002). The expression of the HPV polycistronic transcripts is modulated by the viral regulatory region known as the long control region (LCR) or upstream regulatory region (URR) (Groves and Coleman, 2015). The spatial and temporal expression of the viral genes is necessary for the initiation and completion of the HPV infectious life-cycle (Roden and Stern, 2018). The extra-chromosomal HPV episome is usually present at a copy number of 50-100 per cell nucleus in low-grade CIN lesions (Bedell et al., 1991; Stanley et al., 1989). These copies segregate between daughter cells as the CIN undergo cytokinesis. The viral genome typically reaches a copy number of 10^3 per cell nucleus after infected cells have undergone terminal differentiation in the upper epithelium (Roden and Stern, 2018).

Although three viral proteins E5, E6 and E7 have been shown to have pro-oncongenic activities, the vast majority of HPV infections will not result in cervical neoplasms (zur Hausen, 2002). HPV infection is required but not sufficient for cervical carcinogenesis (Walboomers et al., 1999). HPV infection is the first 'hit' in cervical carcinogenesis, according to Knudson's Two-hit Hypothesis (see **Box 2.1**). HPV DNA is found in 90-95%

of cervical malignancies with HR-HPV types 16 (HPV16) and 18 (HPV18) detected in over 70% of cervical neoplasms (Bosch et al., 1995; de Sanjosé et al., 2018; Schiffman et al., 2011). HPV16 infection is associated with adenocarcinomata as well as the the most common form of cervical malignancy squamous cell carcinoma (SCC) (de Sanjosé et al., 2018). The HR-HPVs are distinguished from the low-risk HPVs by the capacity of their E6 and E7 proteins to synergistically immortalise human cells (Hawley-Nelson et al., 1989). The immortalised HeLa cell line, the workhorse of countless molecular biology labs across the globe, is derived from Henrietta Lacks' HPV18-infected cervical adenocarcinoma (Landry et al., 2013; Popescu et al., 1987). Under normal circumstances, the combined adaptive and innate immune response, of otherwise healthy individuals, is sufficient to clear HR-HPV infections (Denny et al., 2012; zur Hausen, 2002). Cervical cancer has a long latency period and persistent infections increase the likelihood of benign papilloma or CINs becoming high-grade dysplasia or even tumours (zur Hausen, 2002). Carcinogenesis is not beneficial to the virus as excessive cell proliferation antagonises differentiation and prevents the release of virions from the terminally-differentiated, infected keratinocytes (Hong and Laimins, 2013). The early protein E2 is a transcriptional repressor of the HPV early promoter (p97) and regulates its own expression and that of the *E6*/*E7* ORFs (Hong and Laimins, 2013). This prevents their ectopic expression in the basal epithelium (Dürst et al., 1992).

While it is possible for cervical SCCs to develop in the presence of HPV episomes (Gray et al., 2010), 85-90% of cases involve the integration of the dsDNA viral genome into chromosome(s) (Burk et al., 2017; Landry et al., 2013; Pett and Coleman, 2007). Although HR-HPVs should be able to integrate at random anywhere a DNA double strand break (DSB) has occurred, a number of genomic "integration hotspots" have been identified in SCCs (Groves and Coleman, 2018; Hu et al., 2015). Integrated HPV genomes, *HPV integrants*, are commonly associated with specific genes like c-Myc and are frequently at or near regions of genome, known as common fragile site (CFS), that are vulnerable to breakages (Bodelon et al., 2016; Burk et al., 2017; Dall et al., 2008; Thorland et al., 2000). HPV integrants are commonly associated with regions of open chromatin and active histone marks (Bodelon et al., 2016; Christiansen et al., 2015). HR-HPV-host integration breakpoints frequently contain short, flanking regions of homologous sequence (Akagi et al., 2014; Hu et al., 2015). This suggests that the microhomology-mediated end joining (MMEJ) DNA repair pathways play a role in HR-HPV integration (Hu et al., 2015). This is all the more likely given that E6 and E7 oncoproteins impair the canonical homologous recombination (HR) pathways for repairing DNA DSBs (Wallace et al., 2017). Akagi et al. (2014) detected recurrent tandem arrays of host and viral sequence at sites of viral integration. Based on this, they suggested that concatemers of HR-HPV integrant and host sequence can be amplified via a

homology-mediated 'looping' reinsertion model.
HR-HPV infection and the integration of HR-HPV genomes leads to a swathe of transcriptional changes to host and viral genes. Persistent HR-HPVs infections are typically accompanied by the transcriptional de-repression and amplification of *E6*/*E7* ORFs in the basal keratinocytes (Dürst et al., 1992; Paris et al., 2015; Pentland et al., 2018). This follows the epigenetic silencing and/or disruption of the E2 ORF that are frequently observed after HR-HPV integration (Groves and Coleman, 2018; Pentland et al., 2018; zur Hausen, 2002). Host-viral fusion transcripts and/or insertional mutagenesis are possible if HR-HPV integrants are located within genes (Bodelon et al., 2016; Burk et al., 2017; Hu et al., 2015). It has been long suspected that HPV integrants alter the transcriptional regulation of surrounding host genes *in cis* (Durst et al., 1987). In addition to two promoter regions and an enhancer CRE (Groves and Coleman, 2015), the HR-HPV 16 and 18 genomes contain a conserved, strong CTCF site within the *E2* ORF (Paris et al., 2015; Pentland et al., 2018). While it is probable that HR-HPV promoter in correct orientation can activate adjacent host genes *in cis*, it is also possible that an intact HR-HPV CTCF-binding site may alter the transcription of these and more distant genes by mediating novel host-host and host-HPV chromatin interactions in a site-specific manner. This is supported by the existence of a long-range *in cis* interaction on chromosome (chr) 8 in the HeLa cell line between a HPV18 integrant and the *c-Myc* locus (Adey et al., 2013). Depending on the orientation of an integrant CTCF site it may even play a role in TAD formation as discussed in **Section 1.1.1**.
Longitudinal, mechanistic studies of cervical neoplastic development and progression *in vivo* are impossible for ethical reasons. Fortunately, *in vitro* models that can give insight the molecular changes occurring as part of cervical carcinogensis exist. The W12 model is derived from a polyclonal low-grade HPV16-infected CIN explant (Stanley et al., 1989). There are textasciitilde100-200 episomal copies of the viral genome per nucleus in this karyotypically normal line (Stanley et al., 1989). The cells grow as a monolayer culture similar to the conditions found in the basal layer of the cervical epithelium where HR-HPV oncogenic transcriptional deregulation occurs. W12 cells passaged over 9-12 months recapitulate the viral and host genetic changes, including HPV16 integration and the loss of viral episomes, that occur *in vivo* (Gray et al., 2010; Stanley et al., 1989).

2.1.1 Overview

The lab of Nick Coleman, Department of Pathology, University Cambridge previously characterised twenty-four W12-derived integrant clones under non-competitive conditions (Dall et al., 2008). This ensured that all integrants were detected, not just those that confer a selective advantage. The resulting clones, all of which are episome-free, differ only in

their sites of HPV16 integration. In collaboration with Emma Knight, Ian Groves and Cinzia Scarpini from the Coleman Lab and Marco Michalski from Peter Fraser's Lab at the Babraham Institute, I investigated the impact of HPV16 integration on 3D chromatin structure and the regulation of gene expression in five W12 clones characterised by Dall et al. (2008), see **Table 2.1**. To this end Emma Knight and Marco Michalski prepared 'in-nucleus ligation' Hi-C libraries (Lieberman-Aiden et al., 2009; Nagano et al., 2015; Rao et al., 2014), see **Chapter 1** (Section **1.1.3**), to capture viral-host and host-host chromatin interactions. The small size of the viral genome and the low integrant copy numbers meant that normal Hi-C approaches lacked the resolution to confidently detect viral-host interactions. To address this limitation, Emma and Marco adapted the capture Hi-C SCRiBL method, modified from the Promoter Capture Hi-C protocol (Schoenfelder et al., 2015), to specifically enrich for these contacts, see **Fig. 2.1d**. To maximise the resolution, MboI, a 4nt-cutting restriction enzyme was used in all Hi-C library preparations. The integration sites and DNA DSBs characterised by Dall et al. (2008) were refined, and in some cases corrected, using a Capture-sequencing (capture-seq) based enrichment method. RNA-seq libraries were prepared by Cinzia Scarpini to assess the host transcriptional changes that occur as a consequence of viral integration and alterations to host 3D chromatin organisation. I analysed all generated data unless stated otherwise.

Table 2.1: HPV16 Integration Sites in W12 Clones.

W12 Clone	Ploidy	Integration Site*	Locus*	Integrant copy number†
A5	2N	8p11.21	intergenic	1
D2	2N	18q21.2	intergenic	4
F	2N	4q13.3	*RASSF6*	1
		8q24.21		
G2	2N	21q22.1	intergenic	3
H	2N	4q21.23	*MAPK10*	1

* As reported by Dall et al. (2008).

† As determined by Scarpini et al. (2014).

2.2 Results

2.2.1 Generation of Hi-C libraries

CHi-C libraries in biological duplicate, were successfully generated for the five W12 clones in **Table 2.1** and standard in-nucleus Hi-C produced for three of these. Likewise, standard Hi-C libraries were prepared from normal cervical tissue (NCx) in duplicate. No CHi-C libraries were generated from these samples which are HPV16^{neg}. Library preparation steps for standard in-nucleus and SCRiBL CHi-C protocols are summarised in **Fig. 2.1a-c**.

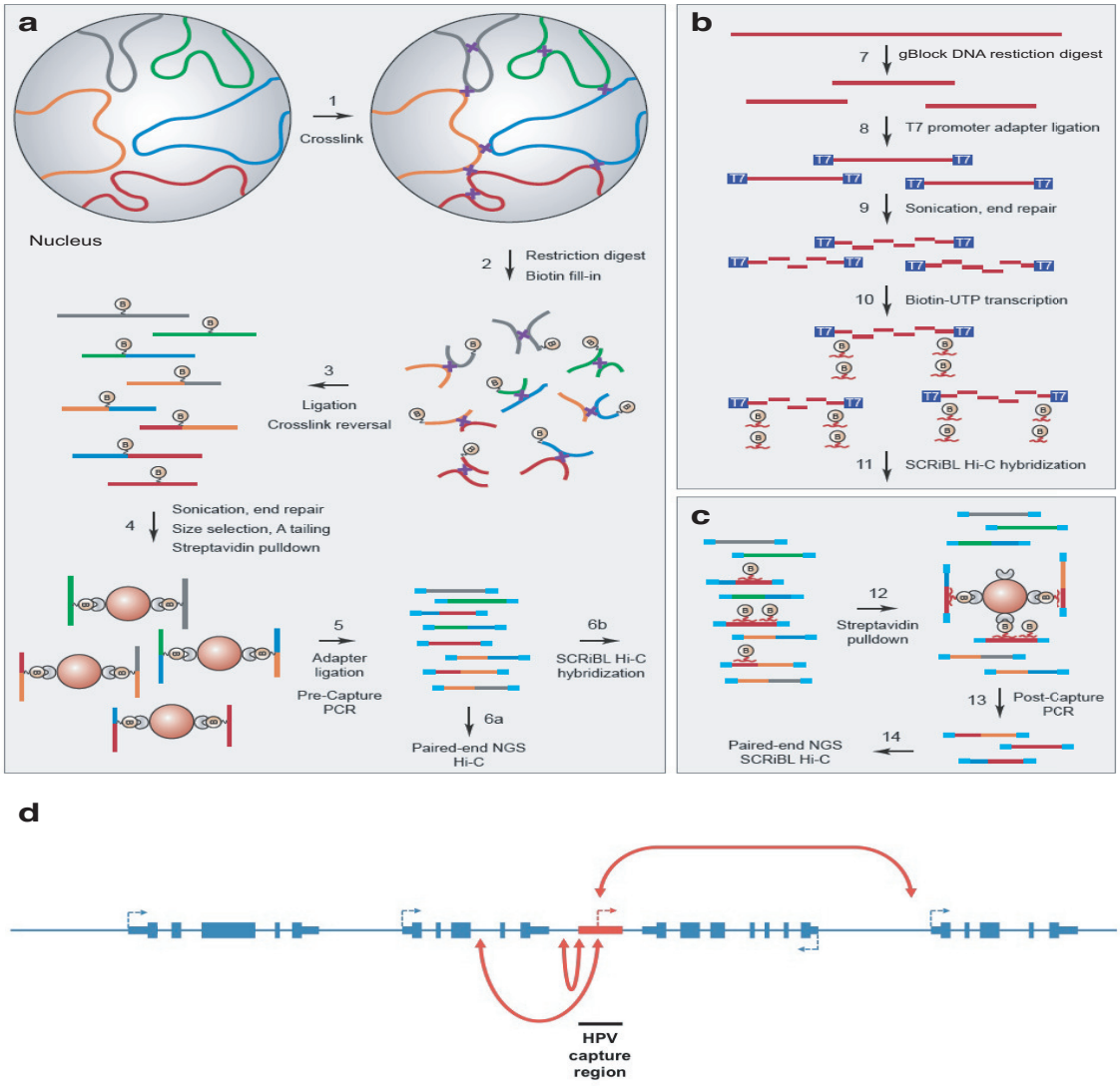


Fig. 2.1: Overview of in nucleus Hi-C and SCRiBL CHi-C library preparation.

Briefly, chromatin-chromatin contacts are captured by the methanol-free, formaldehyde cross-linking of nuclei in 90% confluent 15cm^2 plates. Each replicate consists of two plates and ~15 million cells. Cells are lysed and the cross-linked chromatin digested with the MboI restriction enzyme. Fragment sticky-ends are filled-in with biotinylated deoxyribonucleoside triphosphates (dNTPs). Cross-links are reversed for fragment ends in close proximity by blunt-end ligation and the molecules are sonicated to ~400 bp. The fragment-ends are repaired and size selected. 3′ Illumina adapters are ligated following biotin-streptavidin pull-down. A preliminary PCR determined how many amplification cycles were necessary for the final Hi-C library (Fig. 2.1a). At this point, the Hi-C library was either sent for 50 bp paired-end NGS (standard Hi-C) or for capture and subsequent sequencing with the SCRiBL protocol (CHi-C). HPV16 chromatin contacts were captured by hybridising the Hi-C library with ~120bp biotinylated-RNA baits complementary to the 5′-ends of the HPV16 MboI restriction fragments (Fig. 2.1b & c). This modified SCRiBL protocol enriches for chromatin contacts with HPV integrants allowing for the detection of short- and long-range interactions (Fig. 2.1d).

An initial quality control (QC) PCR digest assay performed by Emma Knight prior to sequencing (data not shown) demonstrated that a known short-range *cis* interaction involving the TSS of human *RPL13A* was captured in all Hi-C libraries but not genomic DNA (gDNA) controls. The ligation efficiency was estimated by quantifying the intensity of the cut and uncut PCR products using *ImageJ*. Ligation efficiencies were high and ranged from 71.1% for A5 replicate I to 94.2% for G2 replicate I. This suggested that the Hi-C libraries were successfully prepared. Hi-C and CHi-C libraries were sequenced on an Illumina® HiSeq 2500 (Sequencing Facility, Babraham Institute) and base-called with CASAVA (v1.8.2, Illumina). As part of the Hi-C User Pipeline (HiCUP) (Wingett et al., 2015), paired reads in FASTQ format were mapped individually with Bowtie 2 (Langmead and Salzberg, 2012) against a modified human reference genome (GRCh37.p13/hg19) containing a HPV16 pseudo-chromosome. HiCUP removes invalid and artefactual C/Hi-C di-tags by overlaying them on an *in silico* restriction digest of the reference genome. HPV16 genome and transcript annotations were obtained from the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena/data/view/K02718). An initial SCRiBL pilot study using different concentrations of the HPV16 bait RNA (1ng, 5ng, 25ng and 125ng) with the G2 Hi-C library, established that was 25 ng the optimal concentration for capture (Data not shown) The ratio of valid to invalid di-tags determined by HiCUP varied across the Hi-C (Fig. 2.2a) and SCRiBL libraries (Fig. 2.2b). The main source of error was from 'dangling ends', due to the presence of biotinylated but non-ligated fragments (Wingett et al., 2015). More promisingly, the ratio of *in cis* to *in trans* contacts within the valid pairs was high for all Hi-C libraries (Fig. 2.2c). Less than 10% of di-tags

were from inter-chromosomal *in trans* contacts and this is a strong indicator of library quality (Nagano et al., 2015). This metric is largely uninformative for the SCRiBL libraries as captured HPV16 di-tags would be classed as *in trans* contacts because the viral genome was considered a pseudo-chromosome in the reference we used. The SCRiBL protocol significantly increased the number and proportion of HPV16 contacts sequenced as compared to the standard in nucleus Hi-C, see **Fig. 2.2d**.

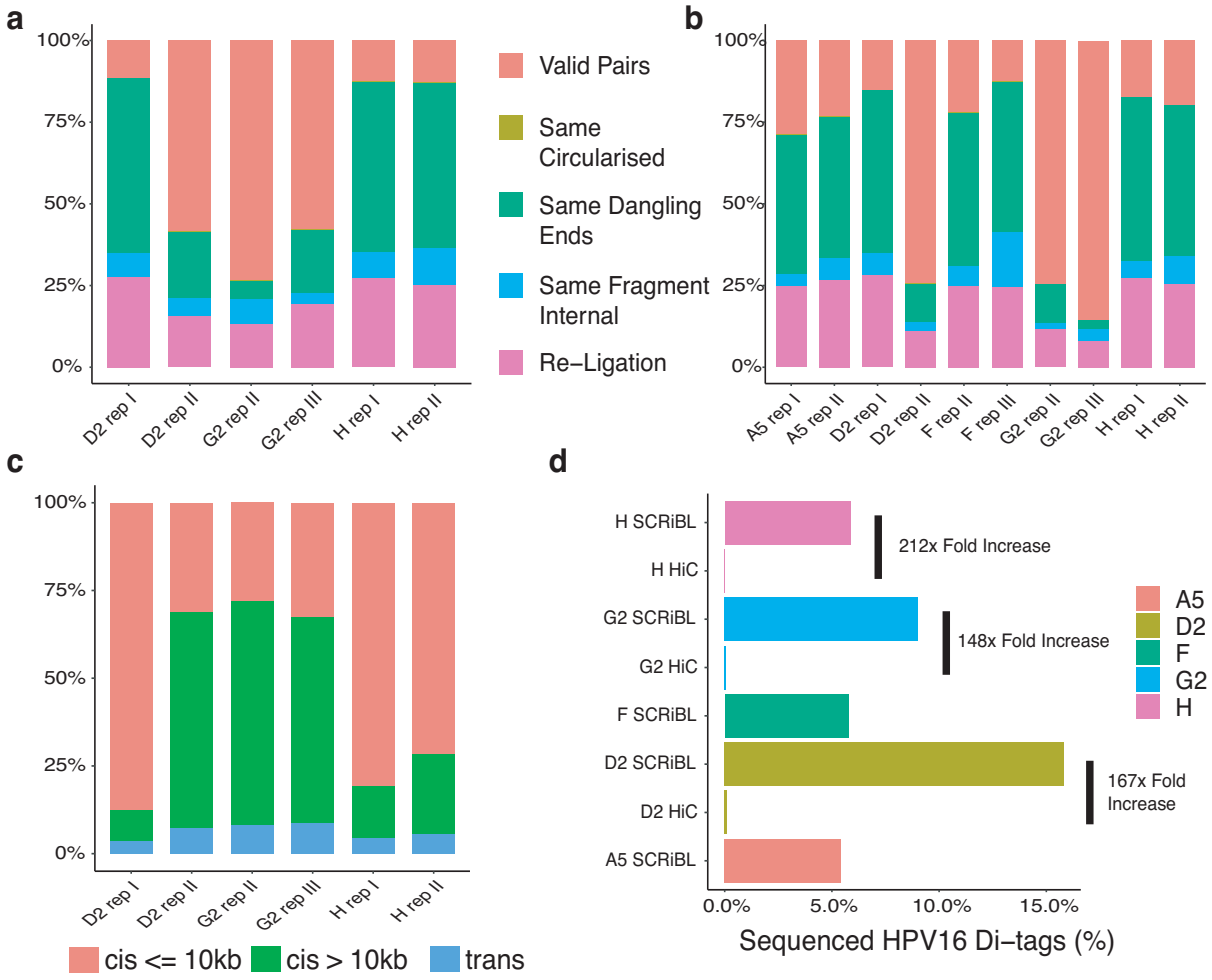


Fig. 2.2: HiCUP statistics for sequenced SCRiBL and Hi-C libraries.
(a) Ratio of valid to invalid di-tags for Hi-C libraries. **(b)** Ratio of valid to invalid di-tags for SCRiBL libraries. Valid di-tags are in red and invalid di-tags consist of circularised, dangling-end, internal and re-ligated fragments. Only di-tags originating from ends coming of different restriction fragments contain information about the 3D organisation. Invalid di-tags are filtered out by HiCUP. **(c)** Bar chart displaying the percentage of near *in cis* ($\leq$10 kb), distant *in cis* ($>$10 kb) and *in trans* reads for the valid read-pairs in each of the generated Hi-C libraries. **(d)** Fold-enrichment for HPV-di-tags in SCRiBL libraries.

Table 2.2: Numbers of sequenced Hi-C di-tags.

	D2 rep I	D2 rep II	G2 rep II	G2 rep III	H rep I	H rep II
Total	57,550,037	59,673,157	95,005,122	89,354,281	75,579,186	94,958,017
Paired	33,631,990	38,187,271	58,018,705	52,348,568	42,441,972	51,456,905
Valid	3,851,501	22,321,321	42,530,343	30,219,773	5,323,477	6,531,624
Invalid	29,780,489	15,865,950	15,488,362	22,128,795	37,118,495	44,925,281
***cis* $\leq$ 10 kb**	3,282,991	6,792,827	11,263,407	9,644,209	4,185,130	4,564,829
***cis* $>$ 10 kb**	332,414	13,519,640	25,624,455	17,455,609	779,562	1,459,227
trans	137,896	1,628,405	3,289,085	2,592,134	229,930	358,098

	Ncx rep I	Ncx rep II
Total	51,682,808	71,563,483
Paired	28,210,670	41,525,436
Valid	2,672,137	8,190,075
Invalid	25,538,533	33,335,361
***cis* $\leq$ 10 kb**	2,027,026	4,357,464
***cis* $>$ 10 kb**	426,737	3,195,209
trans	151,340	493,392

Table 2.3: Numbers of sequenced SCRiBL di-tags.

	A5 rep I	A5 rep II	D2 rep I	D2 rep II	F rep II	F rep III
Total	52,501,708	69,614,938	63,716,356	69,039,987	65,370,911	82,741,321
Paired	24,612,018	25,842,139	29,573,469	31,868,806	30,302,984	38,598,527
Valid	7,061,185	5,971,338	4,435,821	23,656,746	6,631,712	4,836,688
Invalid	17,550,833	19,870,801	25,137,648	8,212,060	23,671,272	33,761,839

	G2 rep II	G2 rep III	H rep I	H rep II
Total	61,723,893	98,143,540	54,790,015	71,105,583
Paired	29,518,707	52,334,332	24,443,572	26,010,504
Valid	21,961,514	44,689,076	4,210,049	5,114,898
Invalid	7,557,193	7,645,256	20,233,523	20,895,606

The proportion of HPV-containing di-tags ranged from a minimum of 3% (H replicate II) to a maximum of 27% (D2 replicate I) in the individual SCRiBL replicates as compared

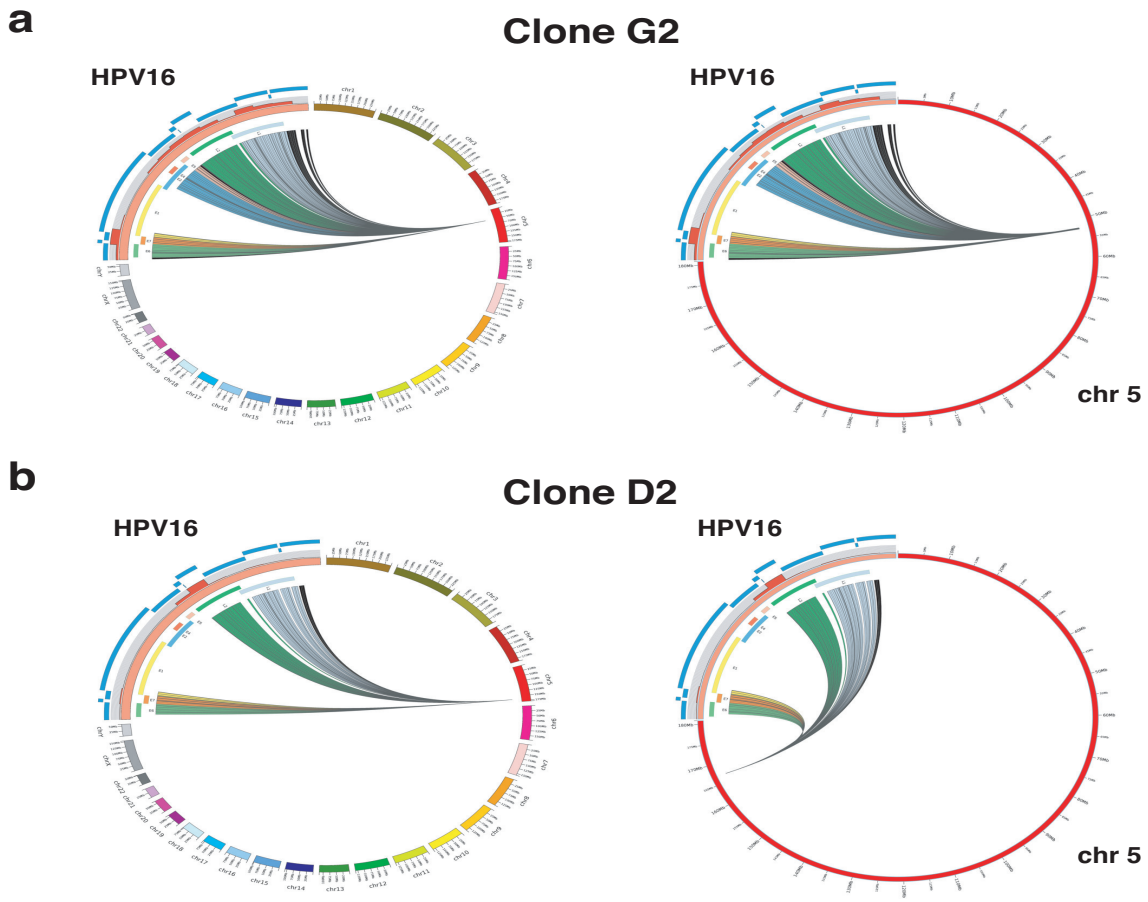
to a maximum of 0.01% (D2 replicate I) in the Hi-C libraries. This represents a 150-200x fold enrichment for the W12 clones with Hi-C and SCRiBL data: clones D2, G2 and H. The number captured and sequenced reads is likely influenced by the HPV16 integrant copy number. The D2 clone has the highest integrant copy number and had the greatest fold enrichment from the capture.
These initial results demonstrate that we successfully adapted the 'in nucleus ligation' Hi-C protocol for use with the 4-cutter restriction enzyme, MboI, in the genomes of human cervical keratinocyte-derived cell lines. Furthermore, we developed a method of enriching for HPV16-containing chromatin contacts based on hybridisation with short RNA baits complementary to the 5′-ends of the HPV16 MboI fragments.

2.2.2 HPV16 integrants interact with host chromatin

As discussed in **Chapter 1** (1.1.3), the frequency of *in cis* chromatin interactions for any two given loci declines with increasing genomic distance (Lajoie et al., 2015). I used the R/Bioconductor package GOTHiC (Mifsud et al., 2017) to identify which regions of the HPV16 integrants are in close spatial proximity with regions of the host chromatin and identify the likely site of HPV16 integration. These were visualised with Circos plots (Krzywinski et al., 2009). The viral integrants should interact with neighbouring regions of host chromatin more significantly and frequently than with chromatin from chromosomes lacking HPV16 integrants. GOTHiC implements a binomial probabilistic model to correct for the biases introduced into Hi-C data during library preparation by factors such as the frequency of restriction enzyme cutting and the ligation efficiency. GOTHiC assumes that these biases are captured by the total number of reads mapping to the two interacting loci. These biases affect each end of the di-tag independently and the probability of observing a random interaction between loci is calculated from the product of their coverages. Using the filtered SCRiBL di-tags obtained from HiCUP, I was able to determine the significant HPV16-human *in cis* chromatin interactions at a 1 kb resolution and thus map the probable sites of HPV16 integration in all five W12 clones.
Each panel in **Fig. 2.3** shows a single representative biological replicate for each W12 clone. The HPV16 integrant interacts exclusively with two narrowly separated loci on a single chromosome in each clone. These likely being the 5′ and 3′ host breakpoints at the sites of integration. No significant HPV16 *in trans* interactions were detected by GOTHiC. All viral interactions originated from at or near the ends of the HPV16 MboI restriction fragments, as is expected from the SCRiBL bait design. This suggests the capture has been successful and has enriched for the HPV16-containing Hi-C di-tags. Few Hi-C di-tags correspond to actual chromatin interactions, most are chance, transient contacts where 2 regions of chromatin find

themselves in close proximity. Filtering with GOTHiC removed these.

For the clone G2, the greatest proportion of di-tags originated from the viral MboI fragment containing the *E7* oncogene. The HPV16 integrant interacts exclusively with a region on chromosome 5 at around 52 Mb (**Fig. 2.3a**). Likewise, the HPV16 integrant in clone D2 also interacts solely with chromosome 5 but with loci at around 167 Mb (**Fig. 2.3b**). Most host-pathogen interactions were formed with the region containing the viral *L1* ORF. This suggests that HPV16 has integrated on chromosome 5 in the case of clones D2 and G2. For clone H, the captured reads indicate that interactions are mainly mediated by the early genes *E6*, *E7*, *E2* and *L1*, with the majority coming from *E2* and are uniquely formed with 2 loci on chromosome 4 (Fig. 2.3c). Strikingly, we found that W12 clones F and A5 had the same integration site, with virus-host reads converging on the same region of chromosome 4 (Fig. 2.3d & e), with the majority of interactions being mediated by the viral *E2*. The absence of di-tags originating from the HPV16 *E1* ORF is not evidence for the absence of interactions involving this region of the genome but reflects the SCRiBL RNA bait design.



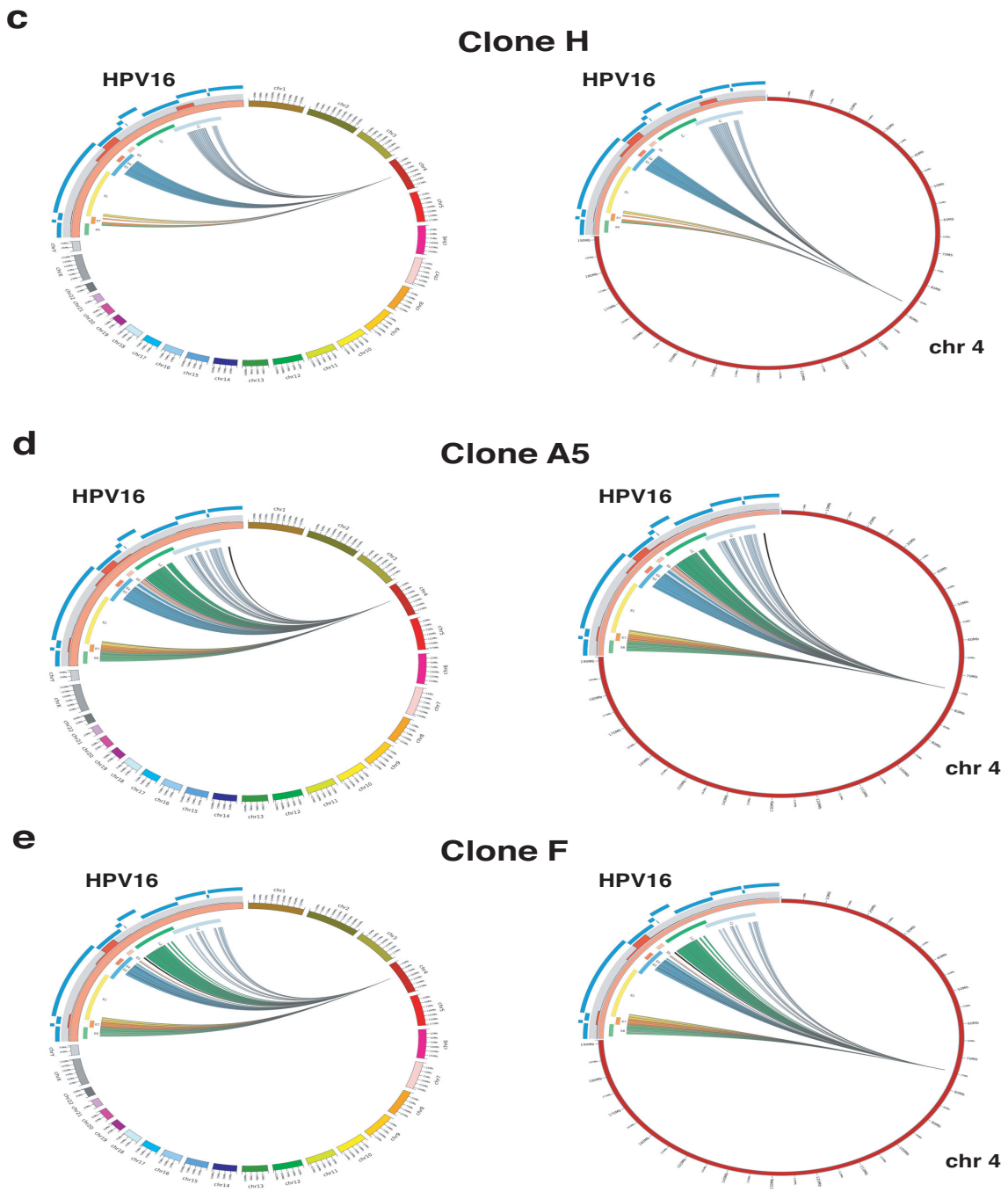


Fig. 2.3: Circos plots showing 3D interactions between the HPV16 and neighbouring regions of the host genome for W12 clones.

Clones **(a)** G2, **(b)** D2, **(c)** H, **(d)** F and **(e)** A5. The upper part of each panel shows significant interactions between HPV genome (in orange) and the human genome and the lower part the interactions with the likely chromosome of integration. Each Circos link is a single di-tag supporting a significant viral-human chromatin interaction determined by GOTHiC. Links are coloured according to where on the HPV genome the significant SCRiBL Hi-C chromatin contact emanates from; green = *E6*, orange = *E7*, yellow = *E1*, blue = *E2*, red = *E4*, pink = *E5*, dark green = *L1*, light blue = *L2* and black = intergenic regions.

Fig. 2.3: Circos plots showing 3D interactions between the HPV16 and neighbouring regions of the host genome (continued).
The proportion of reads emanating from different 500bp bins across the viral genome is indicated by the histogram outside the HPV16 chromosome. The HPV16 MboI restriction fragments are indicated in blue on the outside of the Circos plots.

2.2.3 Identification of virus-host integration breakpoints

Having determined that novel 3D virus-host chromatin interactions likely form after HPV16 integration, we sought to precisely identify the virus-host junctions so as to better characterise these putative interactions. It was also clear from the initial analysis of the SCRiBL Hi-C data that viral integration sites characterised by Dall et al. (2008) for the D2, A5 and F clones differed from what we observed. It was necessary to validate and re-characterise all integration sites using a more accurate means of identifying the breakpoint junctions. Previous studies have used commercially available, HPV-specific probes to enrich for breakpoint junctions in gDNA libraries (Liu et al., 2016). Instead of using these commercial probes, Emma Knight and Ian Groves adapted the capture-seq protocol for enriching target sequences in cDNA libraries (Mercer et al., 2012) to enrich for HPV16 sequences in undigested Hi-C libraries. Briefly, four fragments of roughly equal length were produced from an EcoRI and BamHI restriction digest of the HPV16 genome (Cinzia Scarpini, Coleman Lab). Full-length biotinylated-RNA molecules were IVT from the four regions in the presence of biotin-UTPs. The final 150 bp long baits were generated by chemical fragmentation with Tris (pH 8) and 4mM $MgCl_2$. HPV16 DNA fragments were enriched by hybridising them with the baits and libraries captured with a streptavidin bead pull-down and sent for 50 bp paired-end sequencing on an Illumina® HiSeq 2500 (Sequencing Facility, Babraham Institute). I determined the host-HPV16 breakpoints computationally from the sequenced reads.
Raw FASTQs were filtered down to HPV16-human pairs using *BLAST* (Altschul et al., 1990) to search against the HPV16 genome. *Usearch* (Edgar, 2010), with a sequence identity score of 0.65, was used to find clusters of sequences in the human and HPV16 reads. Consensus sequences for non-singleton clusters were obtained by aligning clustered reads against each other with *Clustal Omega* (Sievers et al., 2011). Cluster consensus sequences were mapped to the GRCh37-HPV16 reference with Bowtie 2 and the locations of host and viral breakpoints identified. Breakpoints were amplified by Emma Knight using two sets of primer pairs complementary to the viral genome and the host flanking sequences at the $5'$ and $3'$ breakpoints. The precise breakpoints were determined from the amplified, chimaeric DNA

with Sanger Sequencing (Sanger et al., 1977) at the Department of Biochemistry, University of Cambridge (data not shown).

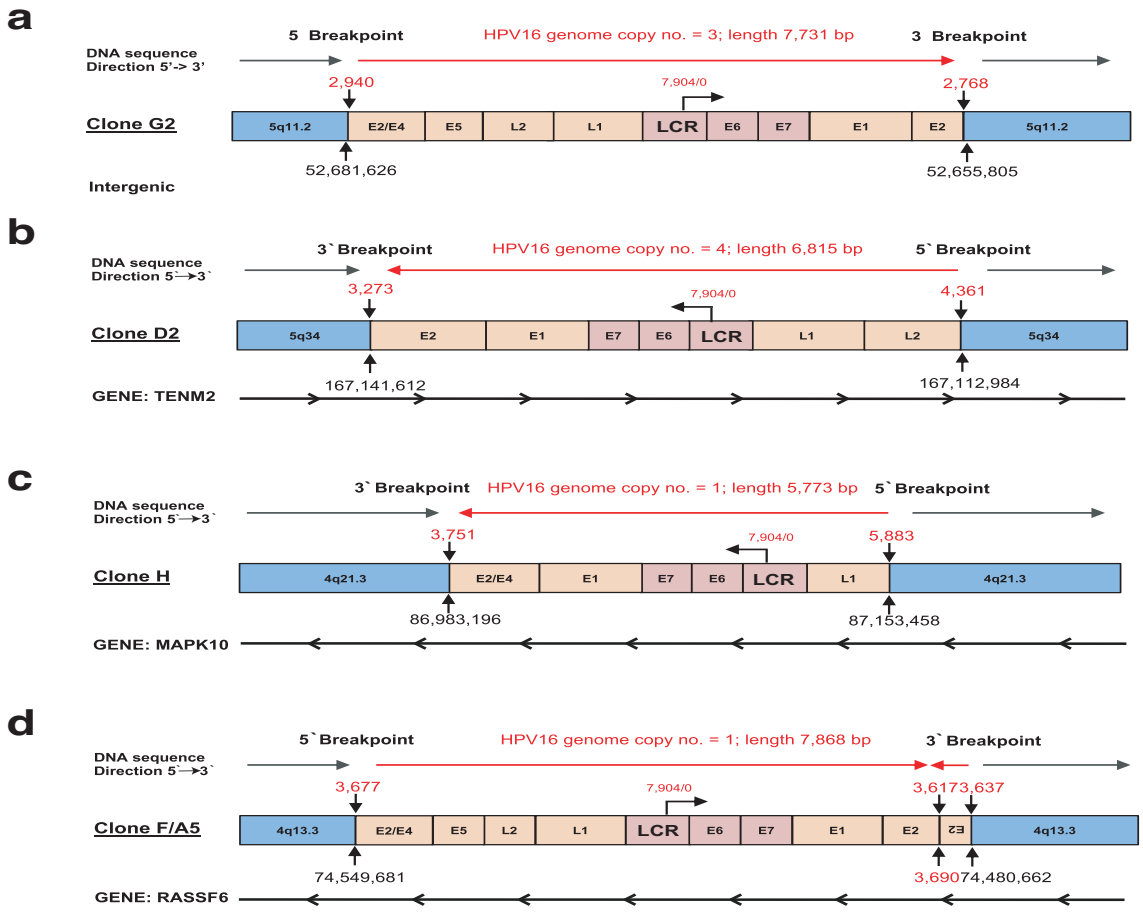


Fig. 2.4: Host-virus junctions at the different integration sites.

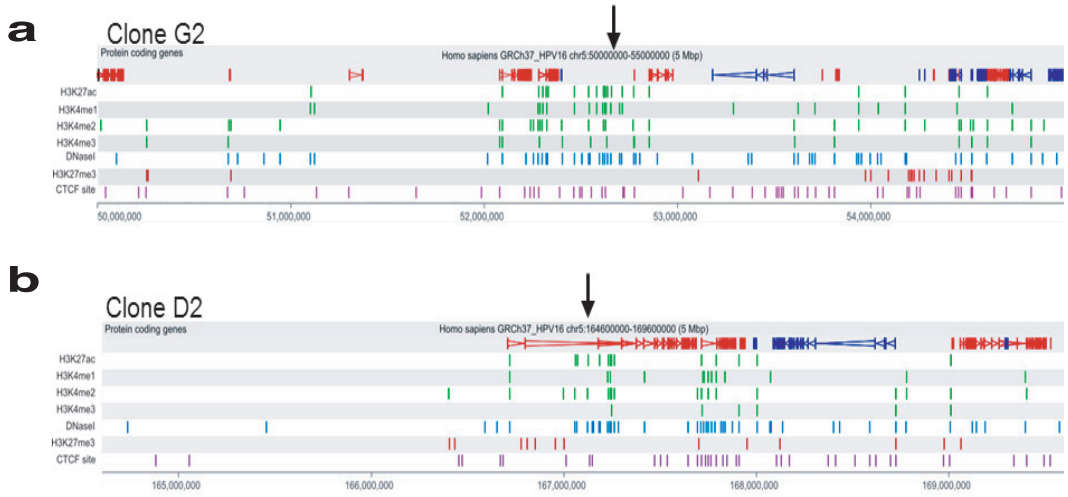
Host chromosomal DNA is shown in blue and orientation indicated by the grey arrow above (5′ to 3′). Integrated HPV16 DNA is shown in orange, with the viral oncogenes and LCR highlighted in red, and the direction of transcription from the viral early promoter indicated by an arrow from the LCR. The viral breakpoint in bp is above the junction. The host breakpoint is below the junction. The genome copy number and length of the integrated HPV16 genome are indicated in red. Clones **(a)** G2, **(b)** D2, **(c)** H and **(d)** A5 and F. Viral genome copy numbers from Scarpini et al. (2014).

With the exception of Clone H, all clone integration sites consisted of tandem arrays of human and viral sequence. The HPV16 integrant in the G2 clone is linearised by a breakpoint in the *E2* ORF [5′: 2,940bp and 3′: 2,768bp] resulting in 173 nt deletion. The majority of *E2* (913bp) is upstream of the virus early promoter as a result. The virus is integrated into an intergenic region on chromosome 5 [5′: 52,681,626 bp and 3′: 52,655,805 bp], see **Fig. 2.4a**. In clone D2, HPV16 has integrated into an intron of *TENM2* on chromosome 5

[5′: 167,112,984 bp and 3′: 167,141,612 bp]. The linearisation of HPV16 genome occurs via breakage in the *L2* [5′: 4,361 bp] and *E2* [3′: 3,272 bp] ORFs, resulting in a 1,089 bp truncation of the viral genome. Furthermore, the viral early promoter is in the opposite orientation to the *TENM2* promoter, see **Fig. 2.4b**. In case of clone H, HPV16 integrated within *MAPK10* intronic sequence on chromosome 4, see **Fig. 2.4c**. Transcription from the early promoter of the viral integrant is in the same orientation as the *MAPK10* promoter. In addition to a large truncation of the viral genome, viral integration resulted in a large deletion of host sequence. 170 kb separate the 5′- (86,983,196 bp) and the 3′-breakpoints (87,153,458 bp). The HPV16 integrant is disrupted within the *L1* [5′: 5,883 bp] and *E2* [3′: 3,751 bp] ORFs, reducing the length of the viral genome to 5,773bp. The capture-seq confirmed that clones A5 and F have the same site of integration on chromosome 4 within an intron the host gene *RASSF6* [5′: 74,549,681 bp and 3′: 74,480,662 bp], see **Fig. 2.4d**. The length of the integrated HPV16 genome was same for both clones. The Sanger Sequencing also detected a 54bp inversion at the HPV16 3′ breakpoint within the *E2* ORF. Short stretches of microhomologous sequence were detected at 5′-breakpoints for clones G2, D2 and H and at the 3′-breakpoints of clones G2, H and F/A5 (data not shown).

2.2.4 HPV16 integrates into open chromatin

With the identification of the exact breakpoints, we were able to profile the chromatin states for 5 Mb regions containing the integration sites based on the ENCODE annotation data for normal human epidermal keratinocytes (NHEK) cells (Encode Consortium, 2012) from Ensembl release 75 (Cunningham et al., 2019). These regions were visualised with the *SeqMonk* (v0.34.1) genome viewer (Andrews, 2007), see **Fig. 2.5**.



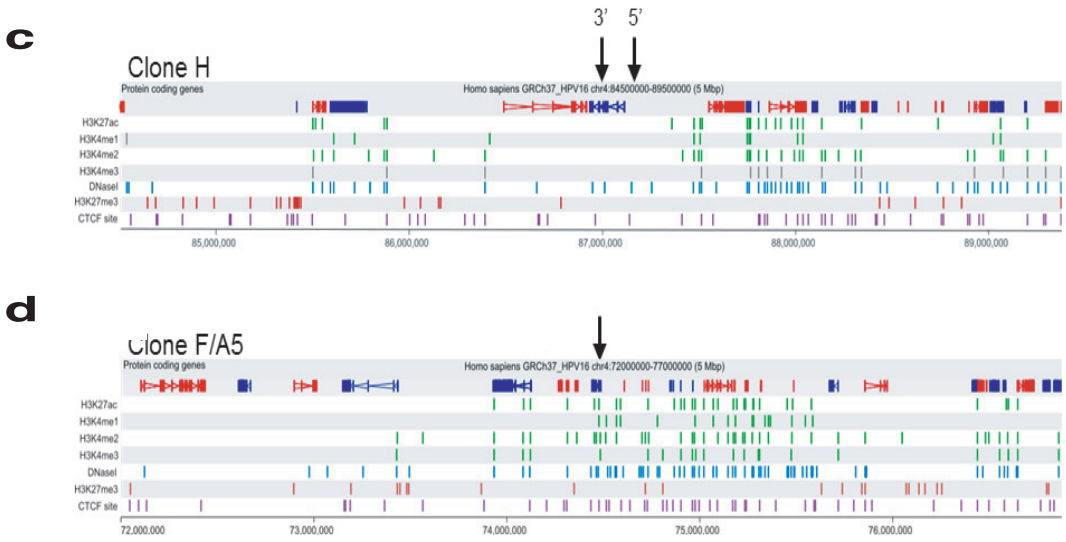


Fig. 2.5: Chromatin marks in regions of HPV16 integration.

Clones **(a)** G2, **(b)** D2, **(c)** H and **(d)** F/A5. Each panel shows the 5 Mb host genomic region surrounding the viral integration site (black arrow; 5′ and 3′ separated in clone H due to deletion). Protein coding genes are shown and colour coded based on orientation (red = + strand; blue = - strand). Active histone marks, such as H3K4me1/2/3 and H3K27Ac are coloured in green; DNAseI hypersensitivity sites are in blue; the repressive heterochromatin mark H3K27me3 is in red and CTCF binding sites are in purple.

HPV16 in clone G2 integrated into an open, active enhancer region as determined from ENCODE DNase I hypersensitivity, H3K4me1 and H3K27Ac peaks (Fig. 2.5a). Active histone marks can also be found at the integration sites of the clones D2, F and A5 (Fig. 2.5b & d). While there are regions of DNase I hypersensitivity within the site of clone H integration, active epigenetic marks are depleted (Fig. 2.5c). Repressive heterochromatin marks, such as H3K27me3 are absent from the integration sites in all clones.

2.2.5 3D interactions between viral integrants and host chromatin

A CHi-C experiment approximates a multiplexed Circularised Chromosome Conformation Capture (4C) experiment with multiple anchors. A virtual 4C (v4C) analysis can be performed by treating the entire captured region as a single anchor. SeqMonk was used to quantitate and profile the density of the HiCUP-filtered SCRiBL di-tags in linear space. Each peak in the v4C profiles represents a 3D chromatin interaction between the HPV16 integrants and host chromatin. Peak heights and intensities reflect the SeqMonk-normalised read depths supporting a particular interaction. The largest peaks were observed at the 5′ and 3′ host integration breakpoints identified from the capture-seq analysis (Figs. 2.6, 2.7 and 2.9). The

largest contact peaks in clone D2 coincided with the host integration breakpoints within an intron of *TENM2* on chromosome 5 (Fig. 2.6a). We were able to identify several candidate short- to medium-range chromatin interactions within the *TENM2* locus by expanding the search window (Fig. 2.6b). The majority of which were formed with regions downstream of the site of integration, 49 to 527 kb away on the linear sequence. All of these were with regions of open chromatin or CTCF-binding sites.

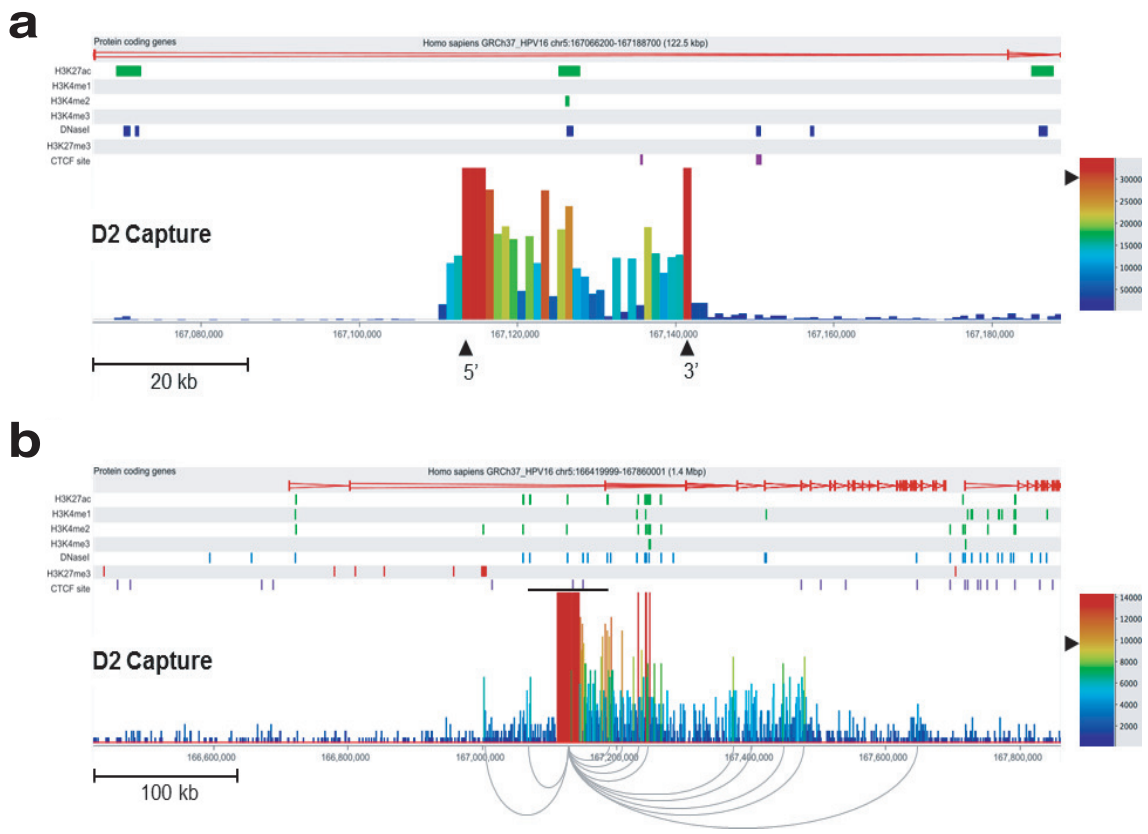


Fig. 2.6: Short and long-range interactions between HPV16 integrant and the host genome in clone D2.

(a) HPV16 SCRiBL di-tags in 122.5 kb region on chromosome 5 containing the viral integration site in clone D2. Host breakpoints coincide with the largest peaks and are denoted with arrows. **(b)** HPV16 SCRiBL di-tags in a 1.4 Mb region on chromosome 5 containing the viral integration site in clone D2. Protein coding genes are coloured based on their orientation (red = + strand; blue = - strand). Active histone marks are in green; DNAse I hypersensitivity sites are in blue; repressive histone marks are in red and CTCF binding sites in purple.

Similar to clone D2, we were able to identify several possible short- to medium-range interactions (34-238 kb) with the host chromatin in the G2 clone (Fig. 2.7b). Interestingly,

these contact peaks appeared to align with CTCF-binding sites. Furthermore, the majority of these peaks overlap with epigenetic marks denoting enhancers (H3K27ac and H3K4me1) and regions of DNaseI hypersensitivity. By expanding the search window to 5 Mb, a number of long-range 3D interactions (>500 kb) between the viral integrant and host chromatin were detected. The furthest and most prominent of these being with the first intron of host gene *ARL15* (Fig. 2.7c).

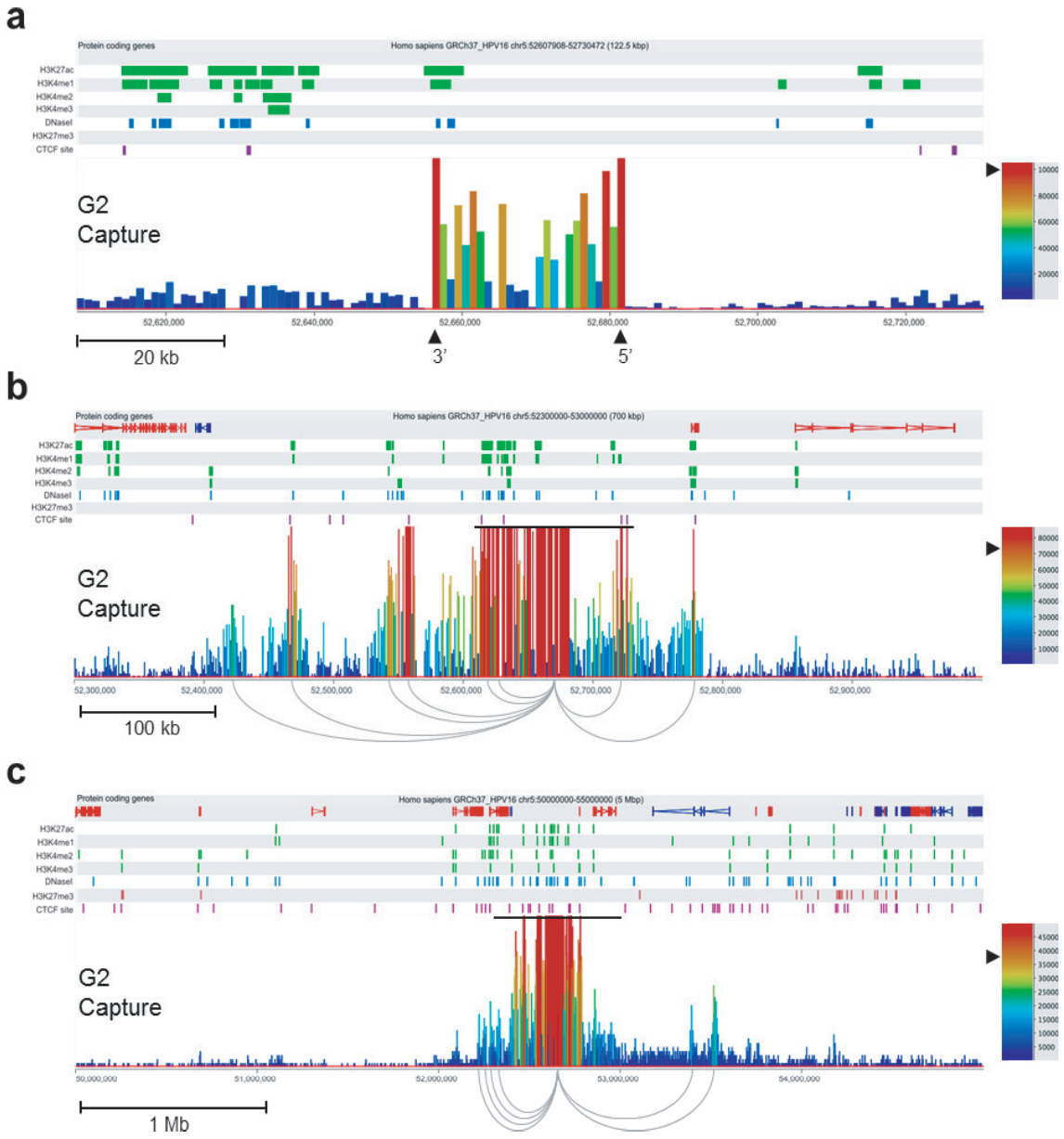
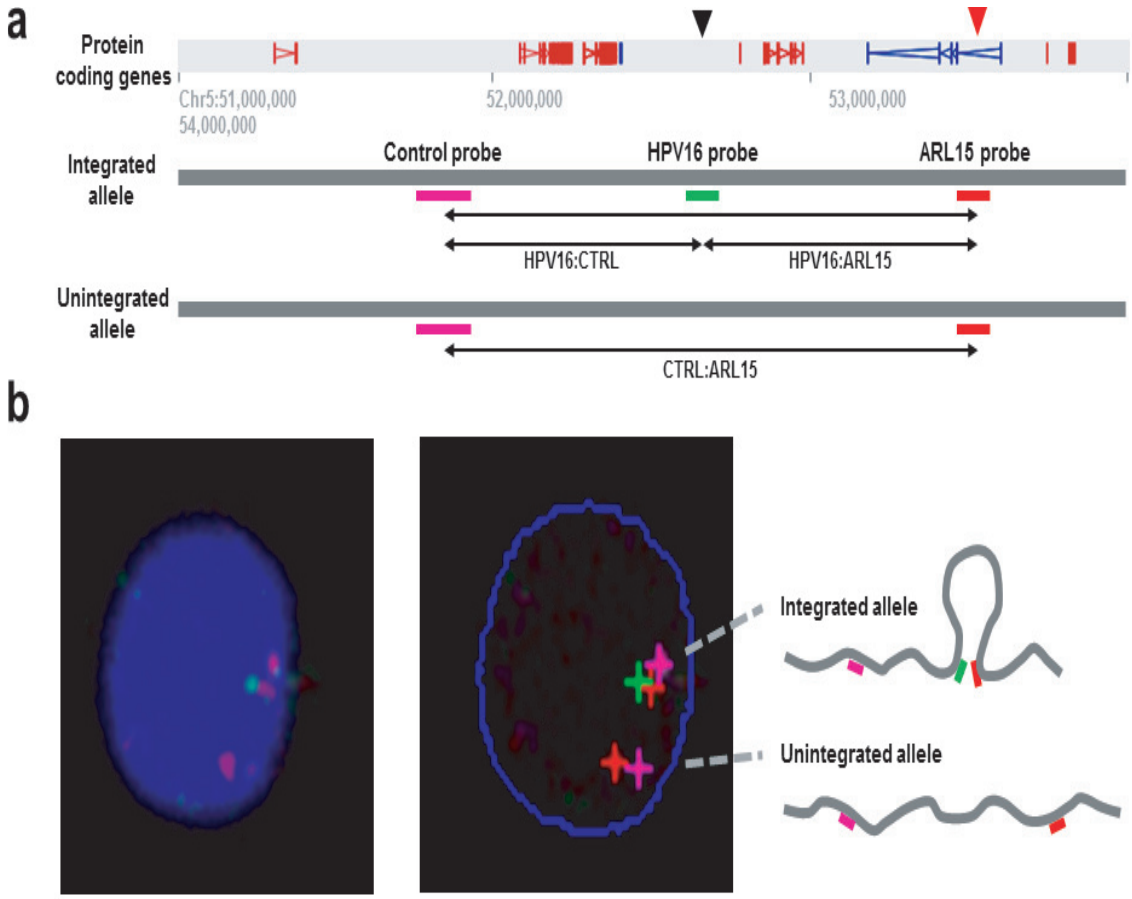


Fig. 2.7: Short to long-range interactions between the HPV16 integrant and host genome in clone G2.

Fig. 2.7: Short to long-range interactions between the HPV16 integrant and host genome in clone G2 (continued).
(a) HPV16 SCRiBL di-tags in 122.5kb region on chromosome 5 containing the viral integration site in clone G2. Host breakpoints coincide with the largest peaks and are denoted with arrows. **(b)** HPV16 SCRiBL di-tags in 700kb region on chromosome 5 containing the viral integration site in clone G2. **(c)** HPV16 SCRiBL di-tags in 5Mb region on chromosome 5 containing the viral integration site in clone G2. Loops are drawn for interactions, with more than 16,000 normalised reads. Protein coding genes are coloured based on their orientation (red = + strand; blue = - strand). Active histone marks are in green; DNAse I hypersensitivity sites are in blue; repressive histone marks are in red and CTCF binding sites are in purple.

Emma Knight used the 3D DNA FISH protocol (Bolland et al., 2013) to verify and validate the long-range interaction I detected between the viral integrant and the *ARL15* intron ~900 kb downstream in the G2 clone. Briefly, three fluorophore-labelled DNA probes were generated via 'nick translation' to hybridise to the HPV16 genome, the first intron of *ARL15* or to an upstream region equidistant from the HPV16 integrant (Fig. 2.8a). A representative image of the hybridised regions in the nucleus of a cell from the G2 clone is shown in **Fig. 2.8b**.



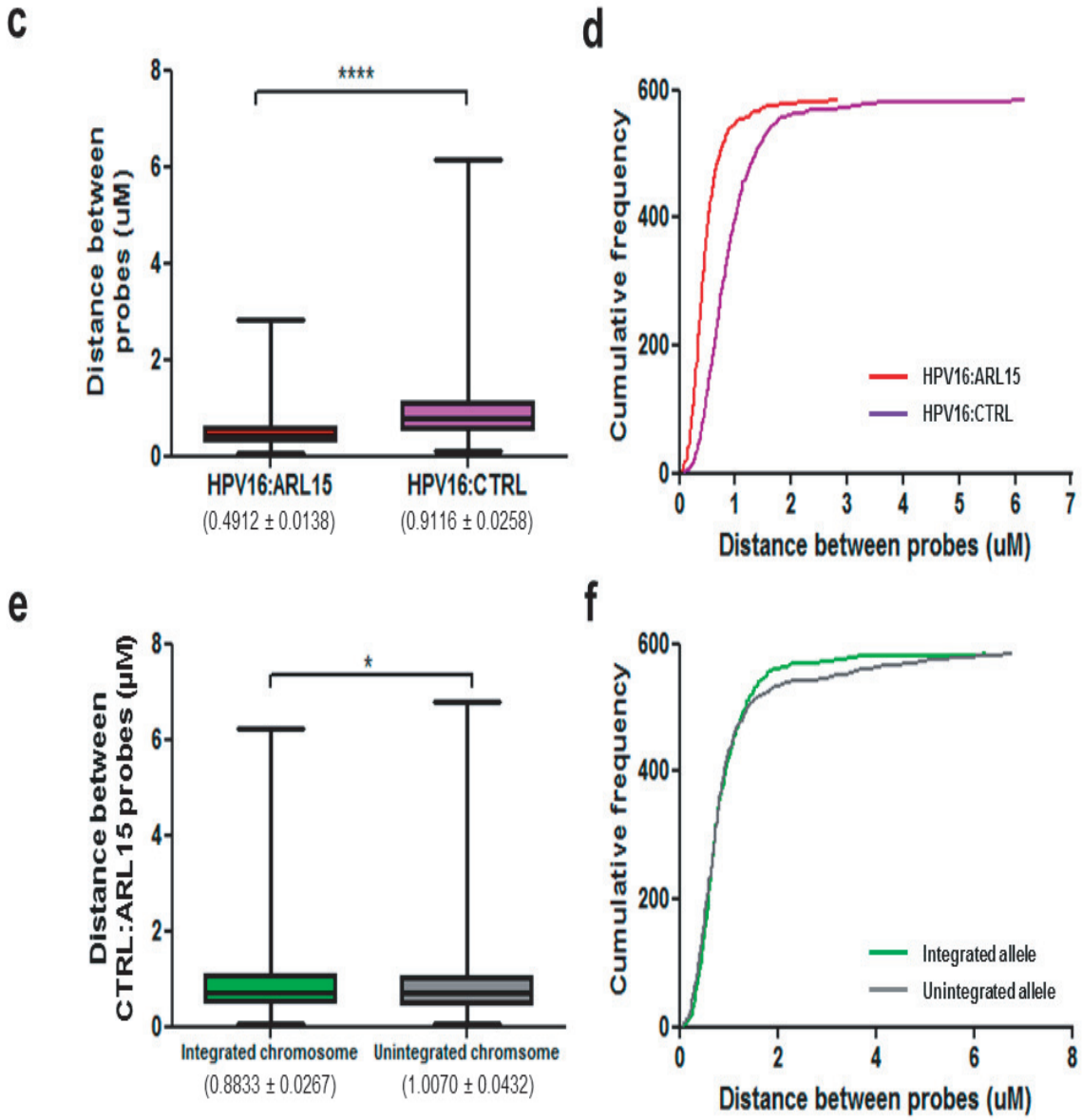


Fig. 2.8: Validation of long-range viral-host chromatin interaction in clone G2 by 3D FISH. **(a)** Location of the DNA probes used in the biallelic region of viral integration on chromosome 5 in clone G2. The control probe (purple) hybridises to a region of the host genome equidistant to *ARL15* intron 1 but upstream of the viral integrant. The *ARL15* intron 1 probe is in red and HPV16 probe in green. **(b)** Image of the three hybridised FISH probes in the nucleus of a representative cell (LHS: raw image with merged channels; centre: MetaCyte segmentation of probe locations; RHS: inferred chromatin looping interaction). Analysis of the 3D distance between both sets of probes: HPV16 to control region (purple) and HPV16 to *ARL15* intron 1 (red), in the copy of chromosome 5 that contained the viral genome is shown in **(c)** a box plot and **(d)** a plot of the cumulative frequency distributions. Analysis of the 3D distances between "control" and "interacting" probes in both the integrated (green) and the non-integrated (grey) chromosomes are shown in **(e)** a box plot and a **(f)** a plot of the cumulative frequency distributions. n cells = 585; mean ± standard error; using unpaired, two-tailed Student's T-test: * $p < 0.05$, *** $p < 0.0001$.

The 3D distances between hybridised regions were calculated in the MetaCyte 3D FISH imaging suite. The HPV16 genome was significantly closer to the *ARL15* intron 1 than the control region on the integrant-containing chromosome 5 (Fig. 2.8c & d). More surprisingly, the distance between the control region and *ARL15* intron differs for the two copies of chromosomes 5. This suggests that viral integration affects the host genome architecture such that the region containing the viral integrant is more closely interacting as a result. The FISH experiment also confirmed that HPV16 had integrated into a single copy of chromosome 5 and that the interaction with *ARL15* is an *in cis* contact with the chromosome 5 of integration. No episomal copies of HPV16 were detected.

The largest contact peaks for clones A5, F and H also coincided with their respective host integration breakpoints (Fig. 2.9a-c). However, no further chromatin interactions were observed between their viral integrants and host genome.

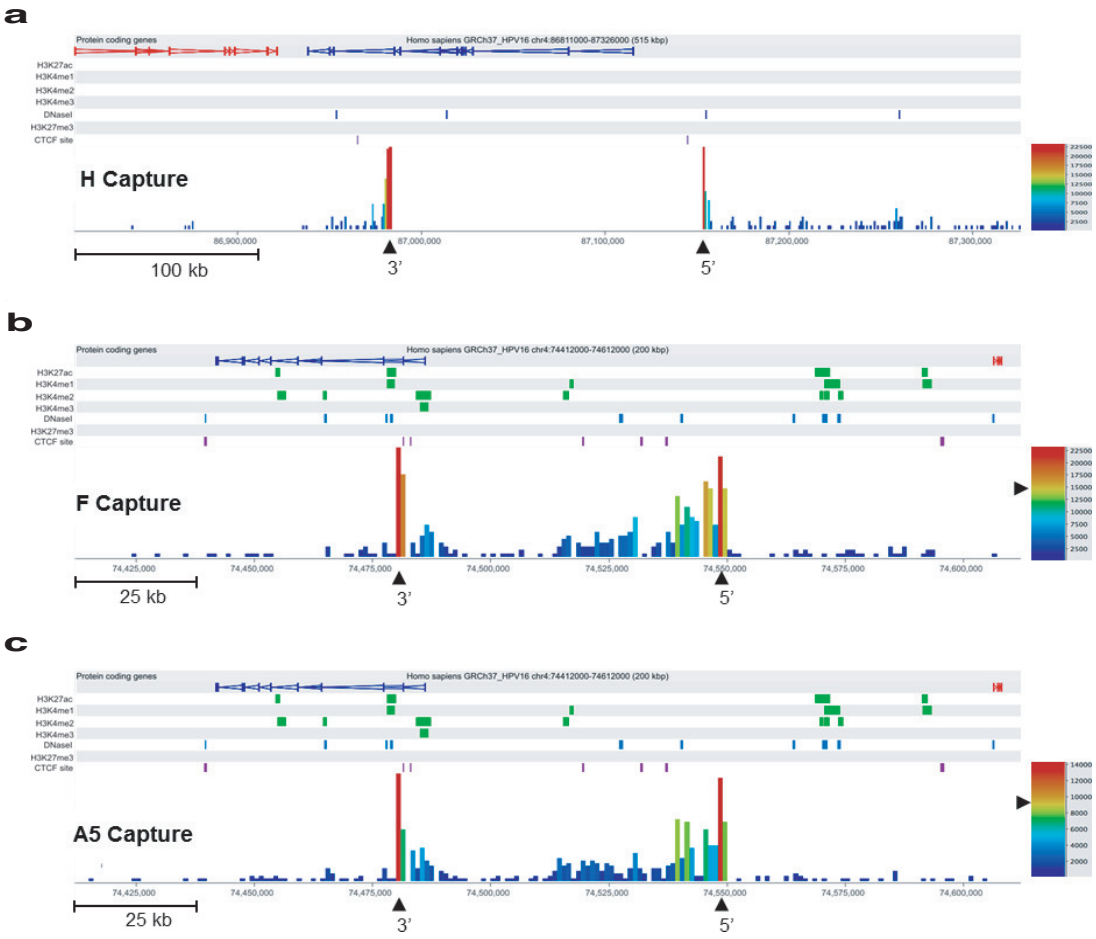


Fig. 2.9: H, F and A5 host-virus breakpoints.

Fig. 2.9: H, F and A5 host-virus breakpoints (continued).
HPV16 SCRiBL di-tag peaks at a 1 kb resolution across the regions containing the viral integration sites of **(a)** H, **(b)** F and **(c)** A5 clones. Host breakpoints coincide with the largest contact peaks and are denoted with arrows. Legends indicate the scaled/normalised read counts in each. Protein coding genes are coloured based on their orientation (red = + strand; blue = - strand). Active histone marks, such as H3K4me1/2/3 and H3K27Ac are in green; DNAseI hypersensitivity sites are in blue; the repressive heterochromatin mark H3K27me3 is in red and CTCF binding sites are in purple.

2.2.6 HPV16 integration alters expression of neighbouring genes

To evaluate the impact of HPV16 integration on the host nuclear architecture, HiCUP and the *Juicer* pipeline (Durand et al., 2016b) were used to generate normalised contact matrices from the raw FASTQs obtained from the in-nucleus Hi-C libraries. Briefly, Hi-C paired-end reads for each replicate are aligned against the hg19 human reference genome using HiCUP, the HiCUP-filtered alignments converted to BED format with *bedtools* (Quinlan and Hall, 2010) and modified to be compatible with Juicer Tools. The matrices were visualised as contact maps using *Juicebox* (Durand et al., 2016a). I restricted my analysis to clones G2 and D2 as these were the only ones for which long-distance interactions were detected in the SCRiBL data. I did not generate contact maps from the SCRiBL data due to the biases in library composition introduced by the capture.
Each column/row in the symmetrical heat map is a human chromosome; autosomes are sorted based on their linear sequence length in descending order with the sex chromosomes as the final two columns/rows. The strength of an interaction between any two loci is reflected in their intensity on the contact map, red = interacting & white = not interacting. Each sub-matrix contains the interactions between any two chromosomes. The sub-matrices that lie on the diagonal contain *in cis* interactions such that the upper leftmost sub-matrix will contain those for chromosome 1 while chromosome Y *in cis* interactions will be found the lower right matrix. Any matrices off this diagonal contain putative *in trans* interactions.
As expected from previous studies of nuclear architecture (Dixon et al., 2012; Rao et al., 2014), the majority of chromatin contacts in clones D2 (Fig. 2.10) and G2 (Fig. 2.11) are *in cis*. Unsurprisingly, the genome-wide profiles for these two W12 clones are very similar. The patterns of *in cis* interactions seem to replicated in both clones, suggesting that the higher order organisation of the nuclear architecture including TADs is largely stable. However, there is evidence for clone-specific *in trans* interactions at this resolution. In clone D2, there are *trans* interactions between chromosomes 3 and 5 (Fig. 2.10). Whereas in clone G2, interactions occur between chromosomes 3 & 7, chromosomes 3 & 10 and chromosomes 9 & 22 (Fig. 2.11). Alternatively, given that these are Mb long stretches of *in trans* interactions,

these may represent early evidence for genetic abnormalities (e.g. translocations) that are typically observed during malignancy. Given that these cells are derived from cervical keratinocytes, it is unsurprising that no interactions were observed for chromosome Y.

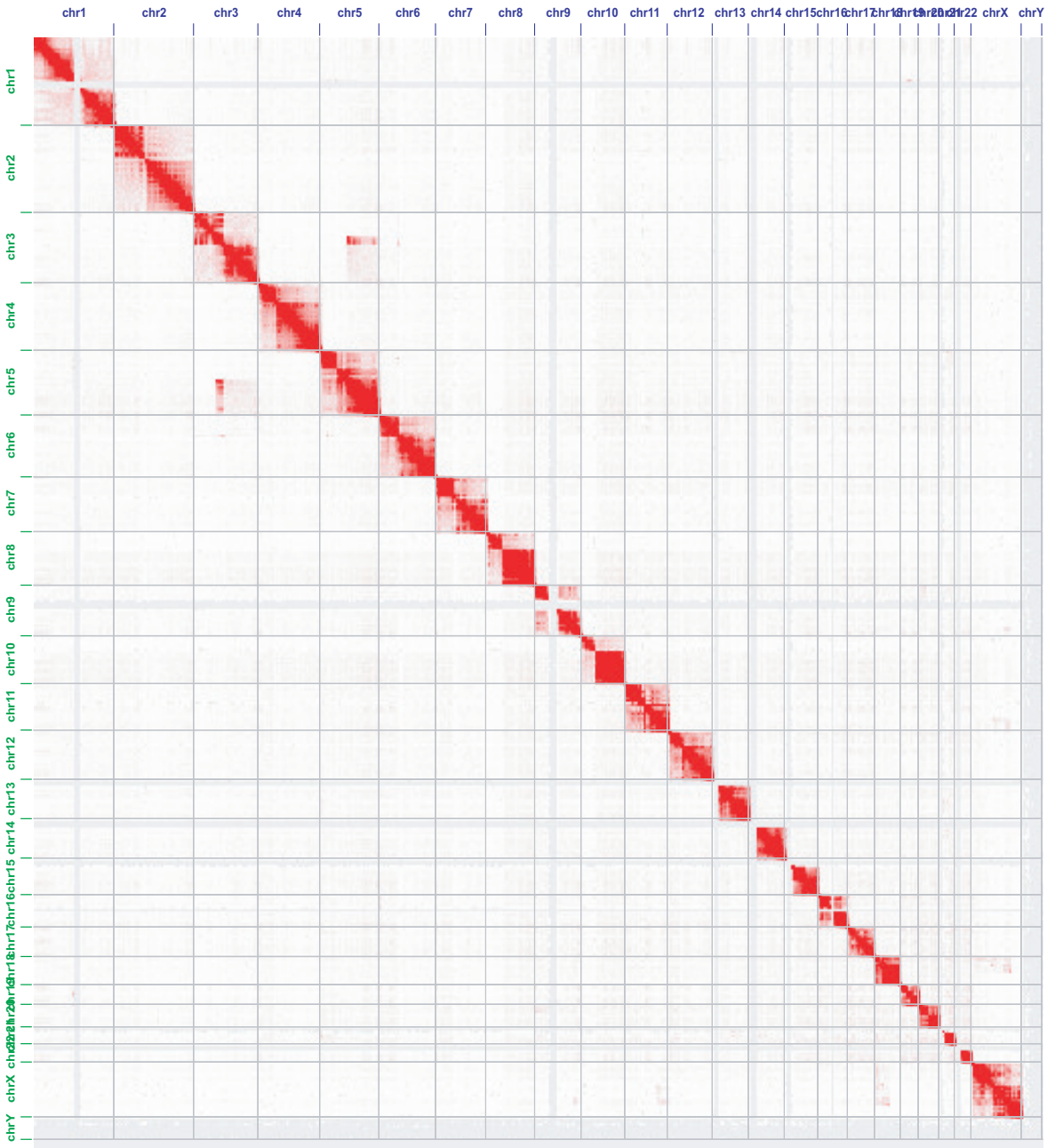


Fig. 2.10: Chromatin Contact Maps for Clone D2.

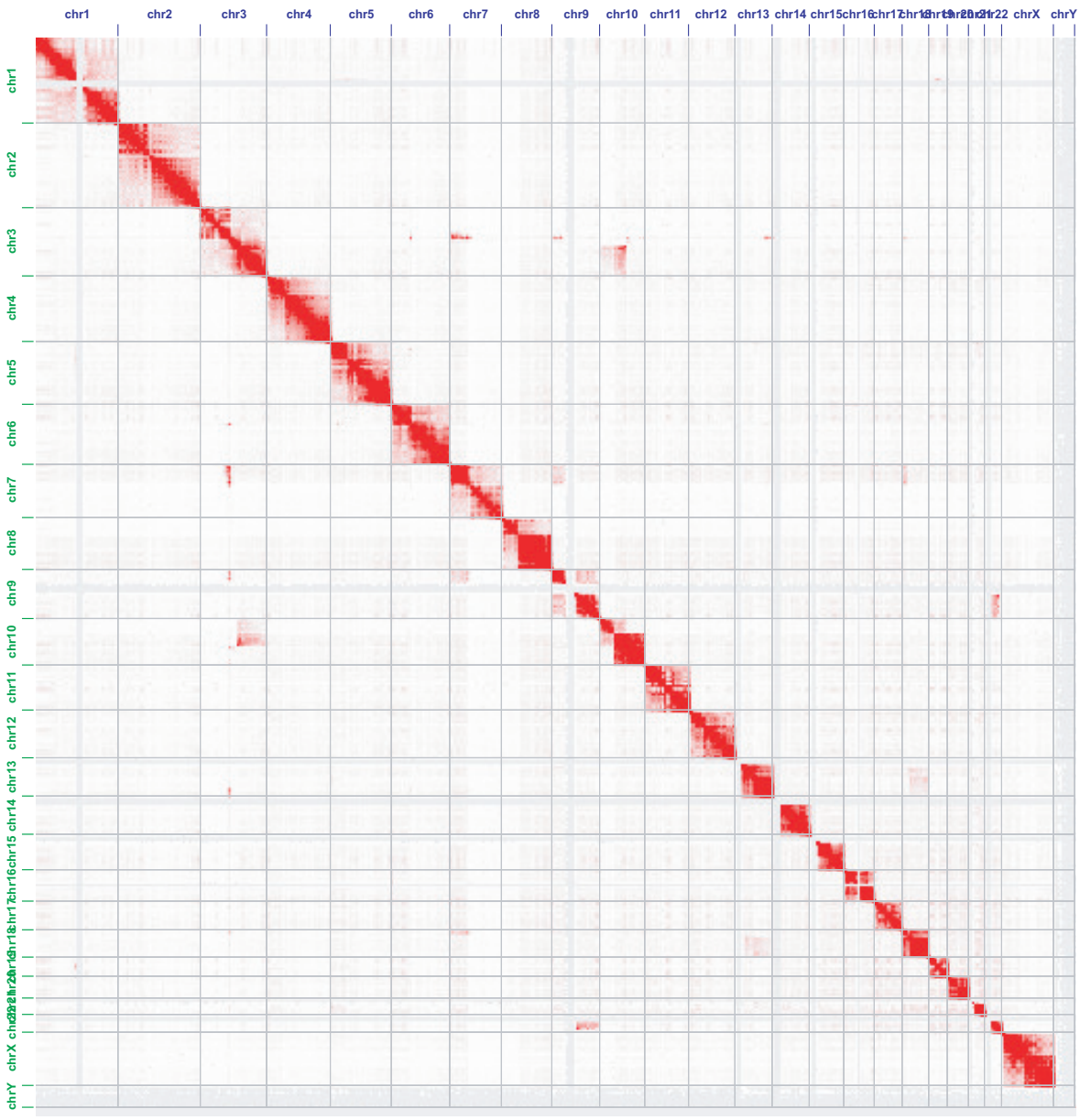


Fig. 2.11: Chromatin Contact Maps for Clone G2.

The HPV16 integration sites for clones G2 and D2 are at distinct, distant loci on chromosome 5 and as a such they can be used as controls when comparing host-host *in cis* interactions in the regions containing their respective HPV16 integrants. Chromatin contact maps at a 50 kb resolution were generated for two 5 Mb regions centred on the D2 and G2 viral intergants. The symmetrical heat maps were split along the diagonal containing *in cis* interactions such that clearly defined, highly self-interacting triangular regions (i.e. TADs) are identifiable. Multiple TADs, up to 1 Mb in length, are visible in the 5 Mb map containing the HPV16

integrant in clone G2 (Fig. 2.12a). However, inspection of the same region of genome in the clone D2 map reveals a very similar pattern of interaction frequencies (Fig. 2.12b). TADs structures are evident from the region containing the HPV16 integrant in clone D2 (Fig. 2.12e) but again the structures in the corresponding region in clone G2 are very similar (Fig. 2.12d). Moreover, the Hi-C-derived TAD insulation scores, devised by Crane et al. (2015), calculated for the two regions in both clones are very similar, see **Fig. 2.12c** & **f**. Insulation scores are highest at the probable TAD boundaries and lowest at their centre. This indicates that for these clones, HPV16 has integrated into two local minima seemingly without altering the structure of surrounding TADs.

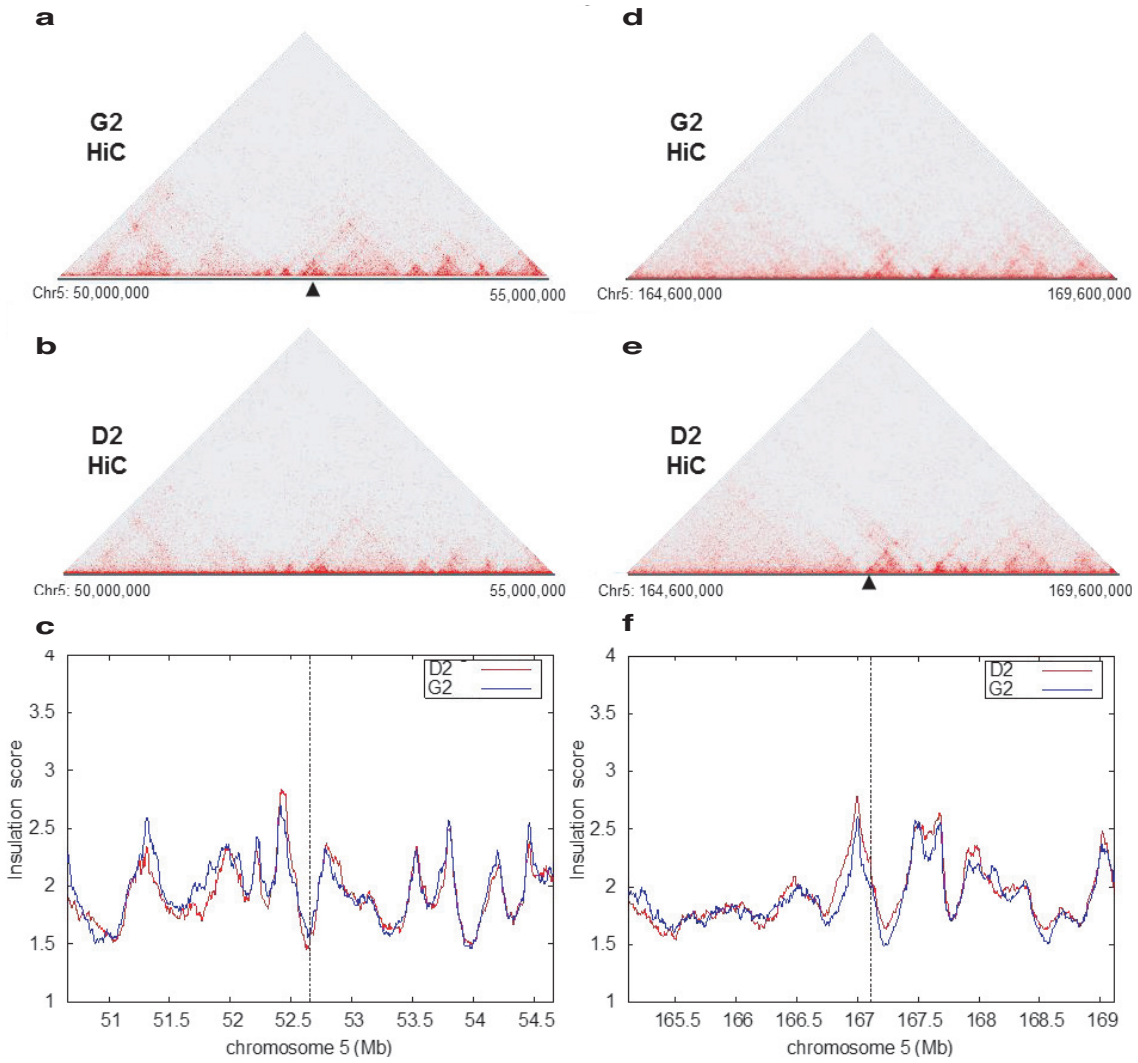


Fig. 2.12: Changes in host genome architecture and domain boundary strength upon HPV16 integration in clones G2 and D2.

Fig. 2.12: Changes in host genome architecture and domain boundary strength upon HPV16 integration in clones G2 and D2. (continued).
Upper triangular matrices generated from a 5 Mb contact map containing the G2 integration locus (chr5:50 Mb – 55 Mb) for clones **(a)** G2 and **(b)** D2. Viral integration sites are denoted with arrows. **(c)** Insulation scores across the same locus as above for clone G2 (blue) and D2 (red). Upper triangular matrices generated from a 5 Mb contact map containing the clone D2 integration site (chr5:164.6 Mb - 169.6 Mb) for clones **(d)** G2 and **(e)** D2, with the integration site being depicted by a black arrowhead. **(f)** Insulation scores calculation obtained for the D2 integration locus in clones G2 (blue) and D2 (red).

In order to evaluate the consequences of differences in chromatin contacts on gene expression, Cinzia Scarpini (Coleman Lab) prepared 50bp paired-end cDNA libraries, with two biological replicates each, for the five W12 clones. Briefly, total RNA was extracted from confluent cells and subjected to Ribo-Zero rRNA depletion and DNAse treatment. cDNA libraries were prepared from the ribo-depleted RNA with the TruSeq™ RNA and DNA Sample Prep Kit (Illumina) and 50bp paired-end cDNA libraries sequenced on an Illumina HiSeq 2000 (Genomics Core Facility, EMBL Heidelberg).
Sequence adapters were trimmed from the reads with *Kraken* (Davis et al., 2013). Trimmed FASTQs were mapped against a GRCh37.p13 reference transcriptome (Ensembl version 75) that included HPV16 transcript annotation using *STAR* (Dobin et al., 2013) with its default parameters. RNA-seq read alignment rates are contained in **Table 2.4**. Strand-specific gene counts were obtained from the alignments with *HTSeq* (Anders et al., 2015) and differential gene expression analysis performed using the R/Bioconductor package *DESeq2* (Love et al., 2014). The impact of viral integration on transcriptional regulation was evaluated by comparing gene expression in the clone of interest versus the mean expression of the other four clones.

Table 2.4: RNA sequencing mapping rates.

	A5 rep I	**A5 rep II**	**D2 rep I**	**D2 rep II**
Sequenced reads	200,526,758	219,342,483	208,473,526	240,417,401
Uniquely mapping	121,039,938	132,527,811	144,075,292	166,800,967
Multi-mapping	24,054,483	25,118,974	20,887,545	25,588,121
Unmapped	55,432,337	61,695,698	43,510,689	48,028,313
Chimaeric	11,617,179	13,376,336	14,342,801	15,664,746

	F rep I	F rep II	G2 rep I	G2 rep II
Sequenced reads	229,801,440	228,349,304	200,319,256	237,877,929
Uniquely mapping	165,180,395	159,676,032	82,425,748	95,643,695
Multi-mapping	24,043,777	25,551,198	50,456,275	54,744,857
Unmapped	40,577,268	43,122,074	67,437,233	87,489,377
Chimaeric	16,015,888	19,712,748	3,458,546	3,468,896

	H rep I	H rep II
Input reads	249,136,137	241,493,648
Uniquely mapping	91,464,098	87,293,759
Multi-mapping	55,198,474	55,627,704
Unmapped	102,473,565	98,572,185
Chimaeric	3,492,896	3,172,965

To visualise differences within the two 5 Mb regions, I generated differential chromatin contact maps at a 25kb resolution with Juicebox using both Hi-C datasets. The values at loci are the log2 enrichment value for the test versus control datasets; regions of the heat map in red have greater observed contact frequencies in the test dataset while those in blue are higher in the control. We combined these maps with the corresponding regions in the SCRiBL data, TAD data derived from human IMR90 and ESC lines (Dixon et al., 2012) and the results of differential gene expression. In clone G2, the long-range interaction between the viral integrant and intron 1 of *ARL15* is accompanied by a reduction in the frequency of host-host interactions within the TAD of integration as compared to the D2 clone (Fig. 2.13a & b). In support of the evidence from the insulation scores, it is clear that the G2 integrant is found within a previously annotated TAD. Furthermore all viral chromatin contacts identified from the G2 SCRiBL data are within the TADs of integration. From comparisons with the other 4 clones, it is clear that the expression of protein-coding genes within the 5 Mb region of the G2 genome is altered, up- and down-regulated, in response to viral integration (Fig. 2.13c). *ARL15* expression is slightly higher in the G2 clone than the other four clones (log2 fold-change = 0.30, adj. $p < 0.05$).

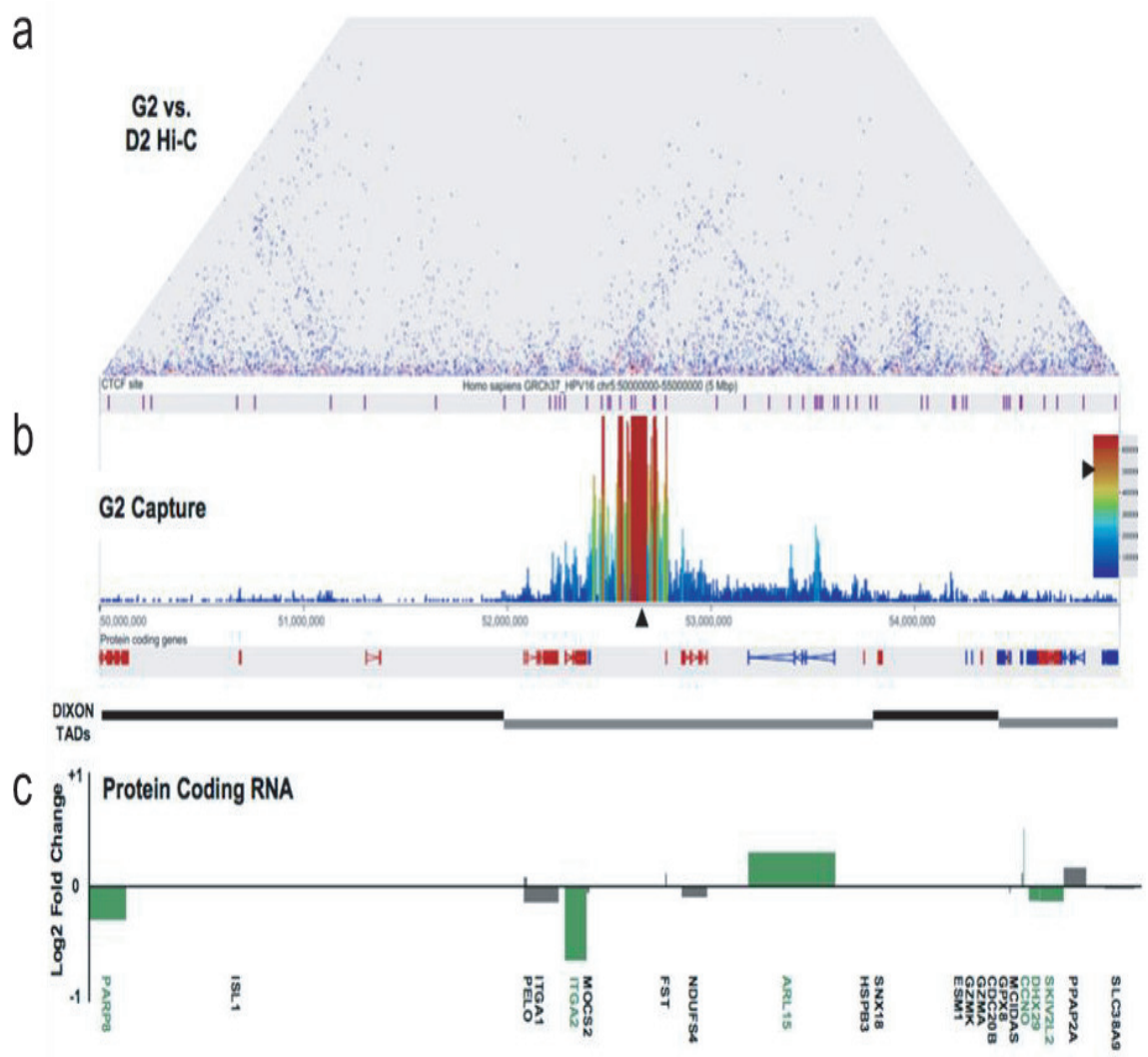


Fig. 2.13: Alterations to host chromatin architecture and gene expression in clone G2.
(a) Differential contact map of clones G2 and D2 covering a 5 Mb region centred on the G2 integrant. Red loci have greater observed host-host contact frequencies in the G2 Hi-C libraries while those in blue are higher in the D2 libraries. **(b)** SCRiBL data showing 3D interactions between the viral integrant and the host. CTCF-binding sites are in purple, protein-coding genes are coloured based on their orientation (red = + strand; blue = - strand). TADs determined by Dixon et al. (2012) are shown beneath. **(c)** Expression log2 fold-changes for G2 versus the average of the four other clones. Significant changes are depicted in green (adj. p value < 0.05, Wald test).

As was the case for the G2 integrant, the chromatin organisation of the 5 Mb region centred on the D2 integration site is very similar for clones D2 and G2 (Fig. 2.14a). Likewise, it is clear that the D2 integrant is found within a previously annotated TAD (Fig. 2.14b). Furthermore the majority of viral chromatin contacts identified from the D2 SCRiBL data are within a

single TAD (Fig. 2.14b). Viral integration has significantly upregulated the expression of *TENM2*, into which HPV16 has integrated, and 3 neighbouring genes downstream of the D2 viral intergant as compared to their expression in the unaltered regions in the other four clones (Fig. 2.14c).

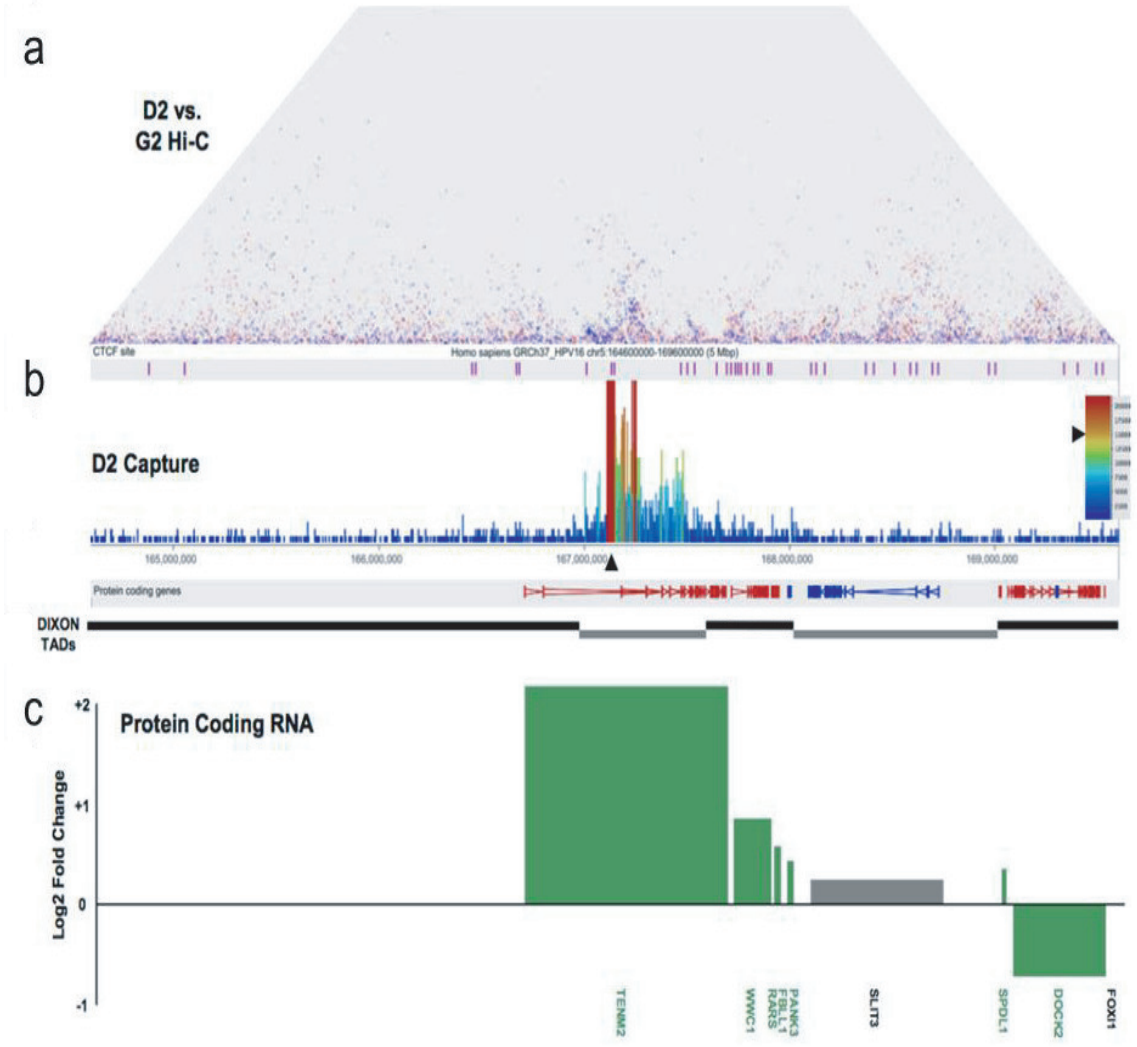


Fig. 2.14: Alterations to host chromatin architecture and gene expression in clone D2. **(a)** Differential contact map of clones D2 and G2 covering a 5 Mb region centred on the D2 integrant. Red loci have greater observed host-host contact frequencies in the D2 Hi-C libraries while those in blue are higher in the G2 libraries. **(b)** SCRiBL data showing 3D interactions between the viral integrant and the host. CTCF-binding sites are in purple, protein-coding genes are coloured based on their orientation (red = + strand; blue = - strand). TADs determined by Dixon et al. (2012) are shown beneath. **(c)** Expression log2 fold-changes for D2 versus the average of the four other clones. Significant changes are depicted in green (adj. p value < 0.05, Wald test).

2.2.7 HPV16 integration results in the formation of viral-host fusion transcripts

HPV16 was integrated within intronic regions for four of the five W12 clones. HPV16 integration consistently increased expression of the host genes containing integrants when compared to the other clones. *TENM2* expression was 4.79 log2-fold higher in D2, *MAPK10* was 4.47 log2-fold higher in H while *RASSF6* had 1.62- and 1.64 log2 fold enrichment in clones A5 and F, respectively. I used the *STAR-Fusion* pipeline (Haas et al., 2017) to determine if integration within any of these three genes resulted in the formation of chimaeric, fusion transcripts. STAR-Fusion filters split and discordant read alignments detected by STAR based on breakpoint proximity and support to infer novel splice events and fusions from existing transcriptome annotation.

STAR detected split viral-host alignments for all five clones but clone H was the only one in which STAR-Fusion detected valid fusion transcripts. Clone H is the sole clone in which the viral early promoter is in the same orientation as the promoter of its host gene. STAR-Fusion identified three viral-host fusion transcripts at the truncated *MAPK10* locus for both clone H replicates, see **Fig. 2.15**.

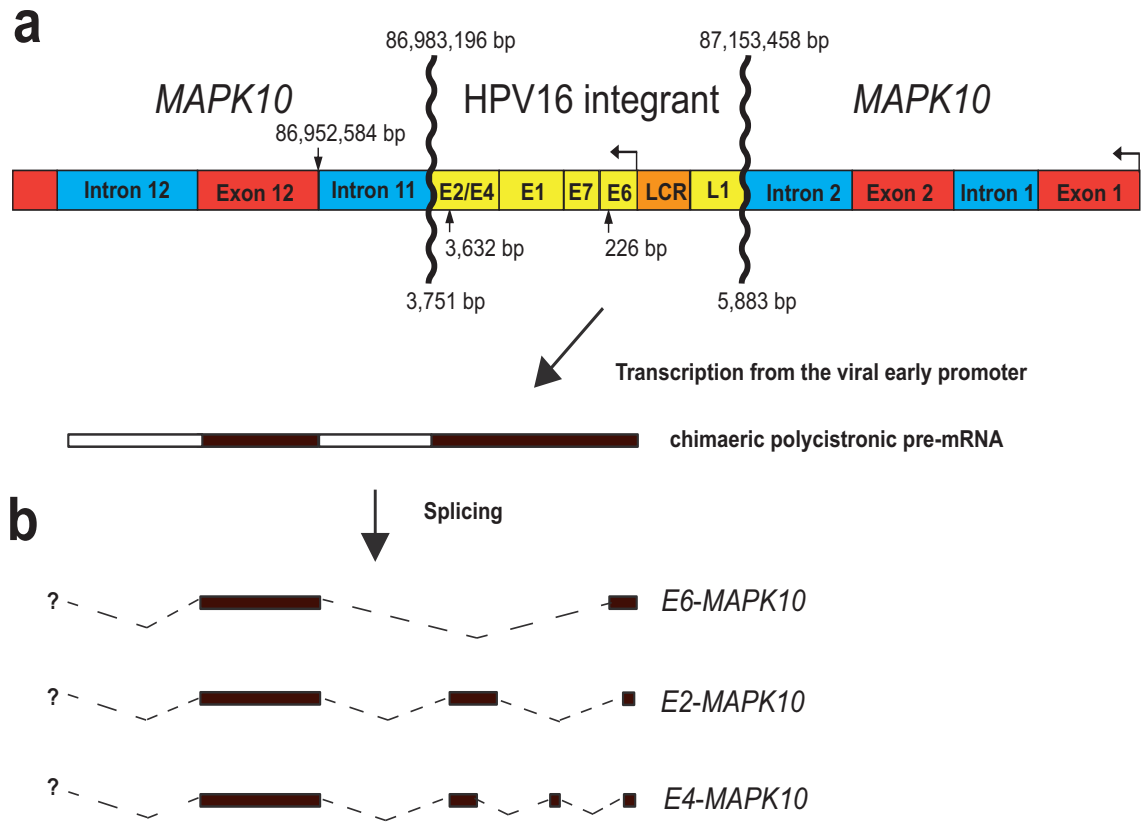


Fig. 2.15: Viral-host fusion transcripts in W12 Clone H.

Fig. 2.15: Viral-host fusion transcripts in W12 Clone H (continued).
(a) Schematic of the truncated *MAPK10* locus on chromosome 4 with the inferred polycistronic viral-host fusion transcript below. Exonic regions of *MAPK10* are in red, intronic regions in blue and viral ORFs in yellow. Direction of transcription from the viral and *MAPK10* promoters shown with arrows. Host-viral integration breakpoints denoted with wavy lines and the fusion transcript splice sites labelled with arrows. Positions above = human, positions below = virus. Splice site positions are based on the coordinates from the HPV16 and hg19 references. *MAPK10* exon/intron numbering based on the annotation for the canonical transcript ENST00000395169 (Ensembl 75). Distances are not to scale. **(b)** Fusion transcript splicing events. The *E6-MAPK10* fusion transcript is produced from the splicing together of a splice acceptor site within the *E6* ORF [HPV16: 226 bp] with the splice donor site at the end of *MAPK10* intron 11 [chr5: 86,952,584]. The *E2-MAPK10* and *E4-MAPK10* spliceforms are produced from the splicing together of a splice site within the overlapping *E2/E4* ORFs [HPV16: 3,632 bp] with the *MAPK10* intron 11 splice donor site. It is unclear without validation if the *MAPK10* exons downstream of exon 12 are included in the chimaeric fusion transcripts.

The viral-host fusion transcripts were formed from two novel splicing events involving the viral *E2/E4* and *E6* ORFs and exon 12 of *MAPK10* (Fig. 2.15b). No fusion transcripts spanning exon 2 of *MAPK10*, the host exon immediately upstream of the viral integrant, and the viral ORFs were detected. This suggests that the fusion transcripts arose as a result of transcription from the viral early promoter continuing into the *MAPK10* sequence downstream of the viral promoter. Additionally, no novel splice events are detected between *MAPK10* exons 2 and 12, indicating that no mature mRNA is produced from the *MAPK10* promoter on the chromosome 4 in which HPV16 has integrated. The increased *MAPK10* expression observed for clone H is likely due to transcription from the integrated HPV16 early promoter.

2.3 Discussion

Genomic instability and chromosomal abnormalities are hallmarks of many cancers (Hanahan and Weinberg, 2011); unsurprisingly HPV16 and the other oncogenic HR-HPVs are potential sources for these and other cancer hallmarks (Roden and Stern, 2018). It has been shown previously that HR-HPV integration is non-random and there are hotspots associated with CFSs (Thorland et al., 2000) and regions of open chromatin and active transcription (Bodelon et al., 2016; Christiansen et al., 2015). The presence of short, microholomogous sequences at the integration breakpoints suggest that microhomology mediated repair facilitates viral integration (Akagi et al., 2014). Additionally, integration is frequently accompanied by alterations to the expression of nearby host genes (Durst et al., 1987). Evidence suggests

that viral dysregulation of more distant host loci may be achieved via 3D chromatin contacts between the viral integrant and the host locus in question (Adey et al., 2013). The integration and "chromatinisation" of HR-HPV genomes is no doubt facilitated by the fact that papillomavirus dsDNA genomes are already bound by cellular histones (Favre et al., 1977).

By leveraging the CHi-C and gDNA-capture-seq data, we were able to determine the integration breakpoints at single-nucleotide resolution in the panel of five W12 clones. The Hi-C and SCRiBL data revealed that viral integration does not disrupt host genomic TAD structure. Rather, we identified additional contacts within these largely unaltered domains between viral integrants and the host chromatin. An integrated analysis of the Hi-C and RNA-seq data demonstrated that viral integrants have a broad impact on the expression of neighbouring host genes up to at least 2.5 Mb from the integrants; previous studies have observed an impact in narrower regions surrounding viral integrants (Hu et al., 2015).

In the first part of this chapter, we established and validated novel methodologies to capture short genomic regions of interest from Hi-C libraries based on the SCRiBL and capture-seq protocols. Both of which may be useful for investigating any dsDNA virus that integrates into its host genome. The resolution of Hi-C is dependent on cutting frequency of the restriction enzyme used (Lajoie et al., 2015) and while use of the four nucleotide cutter MboI improves the resolution of chromatin contacts detected it also vastly increases the complexity of libraries due to the number of restriction fragments generated. 7.1 million MboI restriction fragments are generated from an *in silico* restriction digest of the GRCh37 reference as compared to 0.8 million for HindIII. From this, ~150 trillion pairwise interactions between fragments are possible in the diploid human genome, $(7.1 \text{ million} * 2)^2$. Although the preponderance of *in cis* interactions in genomes (Lajoie et al., 2015) greatly reduces this theoretical search space, it is vast nonetheless. Therefore, an enrichment step is necessary in order to reliably detect chromatin contacts involving the short genome (~8 kb) of HPV16 integrants and simplify the search space i.e. the library complexity. The SCRiBL protocol developed to capture promoter interactions allows for analysis at the restriction fragment level (Schoenfelder et al., 2015). Modifications to this protocol have enabled us to enrich for virus-host interactions. CHi-C has been used previously to enrich for specific interactions involving cancer/disease risk loci and gene promoters (Dryden et al., 2014; Javierre et al., 2016; Schoenfelder et al., 2015) but this is the first application of this technique to a short, integrated viral genome.

However, it must be acknowledged that the desire for a high resolution comes at the cost of a high di-tag failure rate. The Hi-C and SCRiBL libraries prepared from G2 replicates II and III and D2 replicate II had far higher valid:invalid di-tags than the other libraries (Fig. 2.2). Of the invalid chromatin di-tags detected using the HiCUP (Wingett et al., 2015), the

vast majority fail QC due to being 'dangling ends'. As this affects the same samples in both protocols, it is probable that the issue lies with a step in the Hi-C library preparation. Dangling ends are indicative of failure to remove biotin residues from non-ligated DNA fragments during the Hi-C protocol (Belton et al., 2012; Wingett et al., 2015). In general, most Hi-C libraries consist of approximately 10-45% dangling ends (Belton et al., 2012) but three of six Hi-C and seven of the ten SCRiBL Hi-C libraries have >50% dangling ends. The high frequency of MboI cutting means that there will be some very short restriction fragments generated. These are difficult to ligate (Yaffe and Tanay, 2011). While the SCRiBL protocol was only applied to W12 clones with a single HPV16 integrant it would feasible to further adapt this protocol to look at with multiple HPV integrants in the genome such as the HeLa cell line. The efficiency of this protocol should be increased before is applied to a similar virus capable of integrating into its host genome. This could be achieved by using a restriction enzyme that cuts the genome less frequently and by the proper removal of non-ligated DNA fragments.

The SCRiBL protocol significantly increased the number and proportion of HPV16 contacts sequenced as compared to the standard in nucleus Hi-C (Fig. 2.2). The proportion of HPV16-containing di-tags for the individual SCRiBL replicates ranged from a minimum of 3% (H replicate II) to 27% (D2 replicate I) as compared to a maximum of 0.01% (D2 replicate I) for the Hi-C libraries. This represents a 150-200x fold enrichment for the W12 clones with Hi-C and SCRiBL data: D2, G2 and H. The higher proportion of HPV16-containing di-tags sequenced for the D2 and G2 clones is likely a reflection of the higher viral genome copy number in these clones as compared to the A5, F and H clones.

From the analysis of significant chromatin interactions in the SCRiBL data with GOTHiC (Mifsud et al., 2017) and SeqMonk (Andrews, 2007), it was apparent that the HPV16 integrants interact with two narrowly separated regions on a single host chromosome and that these 3D interactions encompass both short- (<50 kb) and long-range (>1 Mb) contacts (Fig. 2.3). 3D FISH confirmed an ~900 kb interaction between the G2 integrant and the first intron of a downstream host gene, *ARL15* (Fig. 2.8). However, it was also clear from this initial analysis that the viral integration sites characterised by Dall et al. (2008) for the D2, A5 and F clones differed from what we observed. Only the clone H integrant seemed to be on the same chromosome. As a result it was necessary to validate and re-characterise all integration sites using a more accurate means of identifying breakpoint junctions. Techniques used to elucidate the virus-host breakpoints in 2008, namely restriction-site PCR (RS-PCR) or amplification of papillomavirus oncogene transcripts (APOT), were the most feasible at the time. However, they are less sensitive and accurate than the NGS-based capture-seq method that we employed. As Hi-C libraries are limited to detecting interactions between

restriction fragments, it is impossible to determine exact integration breakpoints from this approach. New baits had to be designed for the capture-seq experiment to cover the viral genome because the SCRiBL baits were targeted against the 5′-ends of the HPV16 MboI restriction fragments.

Our investigation revealed that HPV16 had integrated into chromosome 4 for W12 clones A5 and F. We also verified that this was also the case for clone H. However, the site of integration differed slightly from Dall et al. (2008). More surprisingly, we found that the integration loci for clones A5 and F were identical. Although all clones were isolated from the same mixed population of 'episomal' W12 cells, this suggests that integration had already occurred in a sub-population prior to isolation. Given that expression of the oncogenic *E6/E7* ORFs is higher in clone F than clone A5 (Groves et al., 2016), it is tempting to suggest that clone F is further along the path to transformation than clone A5. In contrast to the integration sites determined by Dall et al. (2008), the D2 and G2 integrants are found at distinct and distant loci on chromosome 5. The original and re-characterised integration sites are shown in **Table 2.5**.

Table 2.5: HPV16 Integration Sites in W12 Clones

W12 Clone	Integration Site(s)*	Method	Integration Site†	Breakpoint Coordinates†
A5	8p11.21	RS-PCR	4q13.3	5′ = 74,549,681; 3′ = 74,480,662
F	4q13.3, 8q24.21			
D2	18q21.2	RS-PCR	5q34	5′ = 167,112,984; 3′ = 167,141,612
G2	21q22.1	APOT	5q11.2	5′ = 52,681,626; 3′ = 52,655,805
H	4q21.23	RS-PCR	4q21.3	5′ = 86,983,196; 3′ = 87,153,458

* As characterised by Dall et al. (2008).
† As determined using capture-seq.
APOT = amplification of papillomavirus oncogene transcripts
RS-PCR = restriction-site PCR

In four out of five clones, the integration sites were within host intronic sequence; *RASSF6* (clones A5 & F), *TENM2* (clone D2) and *MAPK10* (clone H). Previous studies evaluating the HPV integration sites in cervical cancer samples have also identified HR-HPV integrants directly upstream of *RASSF6* and *MAPK10* (Holmes et al., 2016; Xu et al., 2013). These findings are consistent with previous observations that virus-host integration breakpoints are significantly closer to genes than predicted by chance (Bodelon et al., 2016). The ORF encoding the viral E2 transcriptional regulator was disrupted in all W12 clones by linearisation

of the viral genome as is frequently seen from clinical samples (zur Hausen, 2002). Deletions within the ORF ranged from 36 bp for clones A5 and F to 2,131 bp for clone H. Disruption and/or truncation of the ORF is sufficient to prevent the production E2. Furthermore, part of the ORF is upstream of the viral early promoter p97 in clones G2, F and A5. These observations confirm that the W12 system is a valid model for recapitulating the early stages of cervical carcinogenesis.

Regardless of the viral genome copy number, analysis of the capture-seq experiment revealed that there was only one virus-host 5′- and 3′-breakpoint per integrant. Therefore, two mechanisms of viral integration are evident from the clones: direct integration of the virus following a viral DSB (clone H) while the 'looping' model proposed by Akagi et al. (2014) can explain the duplication of host sequence flanking the A5/F integrants together with the focal amplification of viral and host flanking sequence observed for clones D2 and G2.

In second part of this chapter we established that HPV16 integration occurred within host TADs in regions containing multiple ENCODE marks of open chromatin and active transcription. This corroborates previous findings from SCCs (Christiansen et al., 2015). Interestingly, comparisons between the TADs of integration for the D2 and G2 clones revealed that although integration does not alter domain structure it reduces the number of intra-TAD contacts when contrasted with the same, integrant-free domain in another clone (Figs. 2.14 & 2.13). TAD boundaries within the 5 Mb regions centred on the D2 and G2 integrants coincide with the boundaries determined by Dixon et al. (2012) from human IMR90 and ESC lines. This supports the observation that TAD boundaries are largely stable across different tissues types Smith et al. (2016). Furthermore, viral interactions with host chromatin generally coincided with annotated CTCF-binding sites. CTCF is essential for delineating the boundaries between chromatin substructures as well as mediating intra- and interchromosomal looping interactions (Ong and Corces, 2014). CTCF-binding sites in a convergent orientation are capable of generating novel looping interactions (Rao et al., 2014). Given that the ectopic insertion of the HPV16 CTCF-binding site does not disrupt host TADs but rather weakens intra-TAD interactions; it is possible that the viral CTCF sites function as insulator elements in these contexts.

The expression of genes in which HPV16 had integrated into, *RASSF6* (clones A5 & F), *TENM2* (clone D2) and *MAPK10* (clone H), were consistently upregulated when compared with their expression in the other four clones. Integration, even within intergenic regions, impacted upon the expression of neighbouring host genes up to at least 2.5 Mb away, including those outside the TAD of integration. Increased expression reflects, in part, the transcriptional activation and enhanced recruitment of RNAP II by the viral early (p97) promoter while decreases were likely the result of E6/E7-mediated activities. The activation of neighbouring

host expression by HR-HPV integrants has been observed on numerous occasions (Durst et al., 1987; Hu et al., 2015). A similar pattern of upregulation was observed by Ojesina et al. (2014) for integrated HR-HPVs in cervical carcinomas.

Although chimaeric viral-host reads were detected in each of the W12 RNA-seq samples by STAR, the two clone H replicates were the only ones in which STAR-Fusion detected fusion transcripts spanning host and viral exon junctions. Clone H was the only instance in which the HPV16 early promoter and the promoter of the gene it had integrated into were in the same orientation. The three viral-host transcripts in clone H, *E6-MAPK10*, *E2-MAPK10* and *E4-MAPK10*, were produced as a result of transcription from the viral early promoter continuing into the truncated *MAPK10* locus (Fig. 2.15). It remains to be seen if the fusion transcripts are functionally relevant to the clone H cells but assuming that the normal splicing of downstream *MAPK10* sequence is intact and that the transcripts are translated, it is possible that fusion proteins containing truncated Serine/Threonine kinase domains are produced. No host-virus transcripts were detected, this indicates that no splicing occurred between the intact second exon of *MAPK10* and the viral ORFs. Fusion transcripts as a result of gross genetic changes are a common phenomenon in cancer and HR-HPV-host transcripts have been documented previously by Bodelon et al. (2016); Burk et al. (2017); Hu et al. (2015).

The primary aim of this project was to develop a method for capturing interactions between integrated copies of the short, dsDNA human papillomavirus genome and host chromatin. I can confidently confirm that we succeeded in this aim and furthermore, we successfully applied this method to a panel of HPV16^{+} *in vitro* models in order to explore the implications of these interactions in early cervical carcinogenesis. In addition to the host transcriptional changes influenced by the activities of viral oncoproteins and integrated copies of the viral promoters, it is apparent from the transcriptional changes to the host gene *ARL15* in the G2 clone that HR-HPV integrants can impact upon host gene expression via novel virus-host chromatin interactions.

3

Transcriptional changes with age in undifferentiated mouse spermatogonia

Declaration This work was a joint effort of the Enright and O'Carroll labs. Dónal O'Carroll designed the study. Anton J. Enright and Dónal O'Carroll supervised the study. Ivalya Ivanova, Dónal O'Carroll and I interpreted the results. I performed the computational analysis for all data displayed in this chapter and generated all figures with the exception of (Fig. 3.1 and Fig. 3.2). The manuscript is undergoing final preparations.

3.1 Introduction

Sexual reproduction is the biological process by which eukaryotic organisms create diploid offspring from the fusion of two unicellular, haploid (*n*) gametes (McDonald et al., 2016). This is the most common reproductive strategy in *eukaryotes* and was present in the last eukaryotic common ancestor (LECA) (Goodenough and Heitman, 2014). Sexual reproduction is adaptive as it allows organisms to escape the consequences of Muller's Ratchet (**Box 3.1**) and facilitates the emergence (and dispersal) of novel genetic combinations more quickly than is possible in asexually reproducing organisms (McDonald et al., 2016). Sexual reproduction requires that organisms cycle between diploid and haploid states via cell fusion and meiosis, a form of reduction division (Goodenough and Heitman, 2014).

> **Box 3.1: Muller's Ratchet** Muller (1964) hypothesised that sexual reproduction evolved as a mechanism to escape the irreversible accumulation of deleterious mutations which could blight asexual organisms. Meiotic recombination facilitates the exchange of genetic material between homologous chromosomes (Holliday, 1964; Szostak et al., 1983) and allows otherwise competing lineages to coalesce. This accelerates adaptation through the emergence of novel genetic combinations (Cooper, 2007). In contrast, an improbable series of sequential mutations would be necessary to replicate the same genetic combination in an asexually-reproducing lineage. Genetic recombination allows eukaryotes to avoid an 'error catastrophe' (Muller, 1964).

Gametogenesis is the mechanism by which haploid gametes are generated from diploid primordial germ cells (PGCs), this occurs in the gonads in multicellular eukaryotes (Magnúsdóttir and Azim Surani, 2014). In the anisogamous animals, male spermatozoa (or sperm for short) and female ova are generated from PGCs during spermatogenesis and oogenesis respectively (Adams and McLaren, 2002; Goodenough and Heitman, 2014). A mature spermatozoön fertilises an ovum and the reproductive-cycle begins anew. Once PGCs are defined in the developing embryo, they migrate to the gonadal ridge, site of the future gonads (Adams and McLaren, 2002).

3.1.1 Mammalian Spermatogenesis

Spermatogenesis is a tightly regulated, highly conserved process that commences in the male gonads, the testes, at puberty (Griswold, 2016). Life-long male fertility is sustained by a populations of spermatogonial stem cells (SSCs) and spermatogonial transit-amplifying cells (TACs) (see **Box 3.2**) (White-Cooper and Bausek, 2010). In humans, the average adult male produces 100 million mature sperm per day from these SSCs (Oatley and Brinster, 2008).

Box 3.2: Transit-amplifying cells TACs are partially committed cells, intermediate between stem cells (SCs) and fully differentiated cells. They undergo a limited number of proliferative cycles before undergoing terminal differentiation (Hsu et al., 2014). Pools of TACs are essential for the maintenance of stem activity and regenerative competence in proliferating tissues (Hsu et al., 2014).

In mammals and mice more specifically, spermatogenesis occurs over twelve stages in the epithelium of the testis seminiferous tubules (Oakberg, 1956). Murine PGCs divide into undifferentiated spermatogonia which give rise to differentiated spermatogonia that further differentiate into primary spermatocytes (De Rooij and Griswold, 2012; Oakberg, 1971). Primary spermatocytes undergo a lengthy meiosis cycle that includes two consecutive cell divisions, meiosis I and II, to produce four haploid round spermatids. These cells undergo spermiogenesis during which they acquire the characteristic sperm morphology including flagella. Elongated spermatids are then released into the seminiferous tubules where they undergo spermiation before entering the epididymis as immobile spermatozoa (O'Donnell et al., 2011).

In humans there are two subtypes of undifferentiated spermatogonia, A Dark (A_d), A Pale (A_p), whereas in mice there are three: A Single (A_s), A Paired (A_{pr}) and A Aligned (A_{al}) (Clermont, 1966; Oakberg, 1971). A_s are single cells that directly emerge from the PGCs, A_{pr} are pairs of interconnected spermatogonia while A_{al} are cysts, chains of 4, 8, or 16 cells (De Rooij and Griswold, 2012). Amongst the heterogeneous, undifferentiated spermatogonia, only a small sub-population of A_s spermatogonia are true SSCs (De Rooij and Griswold, 2012). The rodent SSC niche is maintained by 'nurse' Sertoli cells within the seminiferous tubules that secrete glial cell-derived neurotrophic factor (GDNF) (Oatley and Brinster, 2008). This promotes SSC self-renewal by stimulating GDNF receptor complexes comprised of rearranged during transfection (RET) and GFRα1 co-receptors (Sharma and Braun, 2018). Nakagawa et al. (2010) showed that GFRα1 and a differentiation marker Neurogenin 3 (NGN3) are associated with different degrees of stemness within the undifferentiated spermatogonia. GFR$\alpha1^{pos}$ cells are mainly A_s and A_{pr}, while the NGN3^{pos} cells are mostly A_{al} spermatogonia. From this they suggested that there is a transition within undifferentiated spermatogonia from GFR$\alpha1^{pos}$ to NGN3^{pos} state as part of normal spermatogenesis. Furthermore, they determined that this transition is reversible based on the treatment with the gonadotoxin busulfan and subsequent testis regeneration. They proposed that new A_s cells are generated from the fragmentation of NGN3^{pos} A_{pr}/A_{al} spermatogonia. These subsequently revert to a GFR$\alpha1^{pos}$ state. The presence of transit-amplifying undifferentiated spermatogonia in testis ensures efficient regeneration in the event of testicular injury.

More recently, the lab of Dónal O'Carroll discovered a sub-population of the transit-

amplifying NGN3^{+} spermatogenia that expresses piwi-like RNA-mediated gene silencing 4 (*Piwil4*), encoding the PIWI protein MIWI2 (Carrieri et al., 2017). More importantly, these MIWI2^{pos} NGN3^{pos} c-KIT^{neg} spermatogonia are capable of reconstituting testis depleted by treatment with busulfan (Carrieri et al., 2017). MIWI2 is necessary for the *de novo* DNA methylation of TEs during meiosis I (discussed in 1.1.2) and its loss results in a gradual reduction in spermatozoa numbers and leads to the 'wimpy' testis phenotype characteristic of mutations in the Piwi-family genes (Carmell et al., 2007). c-KIT is a marker of differentiating spermatogonia in adult mice (Rossi et al., 2000). Further analysis of GFRα1 and MIWI2 expression in A_s, A_{pr} and A_{al} spermatogonia revealed that there are three populations of cells distinguishable based on their GFRα1 and MIWI2 expression: GFR$\alpha1^{pos}$ MIWI2^{neg}, GFR$\alpha1^{pos}$ MIWI2^{low} and GFR$\alpha1^{neg}$ MIWI2^{pos} (Carrieri et al., 2017). The majority of A_s cells were GFR$\alpha1^{pos}$ MIWI2^{neg}. All three populations were found amongst the A_{pr} while the A_{al} were largely GFR$\alpha1^{neg}$ MIWI2^{pos}.

3.1.2 Male Fertility and Ageing

Although the SSC sustains life-long male fertility, there is an observable decline in fertility in humans from the age of 25 (Hassan and Killick, 2003). Many men over the age of 40 exhibit fertility abnormalities including lower sperm counts and reduced sperm motility (Kühnert and Nieschlag, 2004). A similar trend is observable in the C57BL strain of mice, males are less fertile from 20 months and the majority infertile by 24 months (Franks and Payne, 1970). Increased life expectancy, better education and changes to the societal norms surrounding marriage and the role of women have increased the age at which people marry in many populations. Unsurprisingly, this and the increasing availability of contraceptive prophylactics have led to a concomitant increases in the age at which people start having children. According to the UK Office for National Statistics (ONS), the average age of fathers in England and Wales has risen from 30.3 in 1964, when records began, to 33.3 in 2016 (ONS, 2017). The fertility rate is now at or below the replacement rate in a number of highly developed countries, no doubt this is due in part to the increasing average paternal age across highly developed countries (Nagase and Brinton, 2017).

Organismal ageing or *senescence* is the time-dependent, progressive decline in biological function and is usually attributed to the accumulation of molecular damage over time (Gems and Partridge, 2013; López-Otín et al., 2013). In addition to the *Hallmarks of Cancer* defined by Hanahan and Weinberg (2000, 2011) (discussed in the context of cervical carcinogenesis in **Chapter 2**), nine *Hallmarks of Ageing* have been defined by López-Otín et al. (2013). Of these, SC exhaustion and the associated reduction in the regenerative capacity of tissues is the most relevant for understanding age-associated male infertility. A number of scenarios have

been put forward to explain the role of SCCs in this phenomenon. The three most prominent of them being: the accumulation of deleterious genetic or epigenetic changes in SSCs and its propagation in their descendants, a compromised SSC niche (e.g. due to fewer Sertoli cells) impairs efficient spermatogenesis and lastly rare, gain-of-function (GOF) mutations that lead to the clonal expansion of certain SSC lineages (Paul and Robaire, 2013). This 'selfish spermatogonial selection' leads to a less diverse pool of mature spermatocytes and what is selectively advantageous for a SSC may be detrimental to the fitness of any eventual progeny (Goriely et al., 2013; Lim et al., 2012).

3.1.3 Overview

In all likelihood, age-associated fertility abnormalities will become increasingly prevalent as the average paternal age continues to grow across the globe. Additionally, many cancer treatments, particularly radiation therapy and alkylating agents, render males infertile by damaging DNA and interrupting spermatogenesis (Dohle, 2010). Under these circumstances, adults may provide sperm samples for cryopreservation prior to treatment. This is not an option available for prepubertal boys. Heterologous transplantation of cryopreserved testis has successfully been performed in mice (Avarbock et al., 1996) but has yet to be carried out in humans. Long-term cryopreservation leads to a deterioration in spermatazoa quality (Dohle, 2010) and it is not known what impact it would have on the SSCs in a testis sample. For those afflicted with testicular germ cell tumours, autologous transplantation of patient testis carries the risk of re-introducing malignant cells (Dohle, 2010). Ultimately, any therapeutic intervention targeting age- or cancer-associated male infertility or subfertility should involve the transplantation of purified SSCs or transit-amplifying undifferentiated spermatogonia.

The male germline has a unique transcriptome that includes many testis-specific transcripts (White-Cooper and Bausek, 2010). A greater understanding of the transcriptional dynamics underpinning SSC identity and fate decisions will provide impetus towards regenerative therapies for male infertility (Oatley and Brinster, 2008). To this end, I have analysed the transcriptional changes that occur in undifferentiated spermatogonia with age and following testicular injury and regeneration in collaboration with the lab of Dónal O'Carroll (MRC Centre for Regenerative Medicine, Edinburgh). We generated bulk RNA-seq and scRNA-seq profiles for the GFRα1^{pos} and MIWI2^{pos} spermatogonia characterised by Carrieri et al. (2017). Furthermore, we profiled transcriptional regulation by assaying chromatin accessibility with assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2013). Ivalya Ivanova (O'Carroll Lab) performed all experiments and I analysed all generated data unless stated otherwise.

3.2 Results

3.2.1 Data generation and processing

In low doses, treatment with the DNA-alkylating agent busulfan damages spermatogonia and can be used to model testicular regeneration (Bucci and Meistrich, 1987). Three cohorts of male mice were established in order to determine the impact of ageing and testis regeneration on the transcriptomes of undifferentiated spermatogonia. The first two cohorts were comprised of adult (3-5 months old) and aged (14-16 months old) mice (Fig. 3.1a). The third group was composed of adults that had been treated with busulfan in order to simulate testis regeneration over a 12 week period (Fig. 3.2a). We performed the following three NGS experiments on fluorescence-activated cell sorting (FACS)-sorted populations of GFRα1 and MIWI2 positive undifferentiated spermatogonia from the testes of these mice:

1. Bulk RNA-seq to analyse differential gene expression from mean RNA expression
2. Bulk ATAC-seq to profile the chromatin accessibility genome-wide
3. Plate-based scRNA-seq to detect differential gene expression at the single-cell level and profile the differences in transcriptional variability.

The QC and analyses steps are explained throughout this chapter.

3.2.2 Cell Preparation and Isolation

In order to isolate the GFRα1 and MIWI2 positive spermatogonia, *Gfra1*$^{\text{GFP/+}}$; *Piwil4*$^{\text{tdTom/+}}$ mice from a mixed 129–C57BL/6 genetic background were generated at the Centre for Regenerative Medicine, Edinburgh using the *Piwil4*-tdTom and *Gfra1*-GFP reporter alleles engineered by Carrieri et al. (2017) and Uesaka et al. (2007). After the removal of tunica albuginea, Ivalya Ivanova prepared single cell suspensions from the testes of adult, aged and regenerated mouse through two consecutive enzymatic digestions with type XI collagenase and trypsin. By applying Magnetic Activated Cell Sorting (MACS) we were able to six-fold enrich for GFRα1 and MIWI2 positive undifferentiated spermatogonia (data not shown) and to significantly reduce the sorting time. We marked CD9$^{\text{pos}}$ cells with Biotin Rat Anti-Mouse CD9 antibody Clone KMC8 (BD Bioscience), followed by streptavidin-coated Magnetic MicroBeads (Mylteni Biotech). The CD9$^{\text{pos}}$ cells were then captured in a column on a magnetic stand while the CD9$^{\text{neg}}$ cells washed away. Eluted, CD9-enriched cells were stained with a c-kit-PECy7 antibody to distinguish them from differentiating spermatogonia. Populations of GFRα1$^{\text{pos}}$, CD9$^{\text{pos}}$ GFRα1-GFP$^{\text{pos}}$ MIWI2-tdTom$^{\text{pos}}$ c-Kit$^{\text{neg}}$ (GFRα1$^{\text{pos}}$

MIWI2^{pos}) and MIWI2^{pos} cells were sorted on a four-laser BD LSRFortessa™ (Fig. 3.1b).

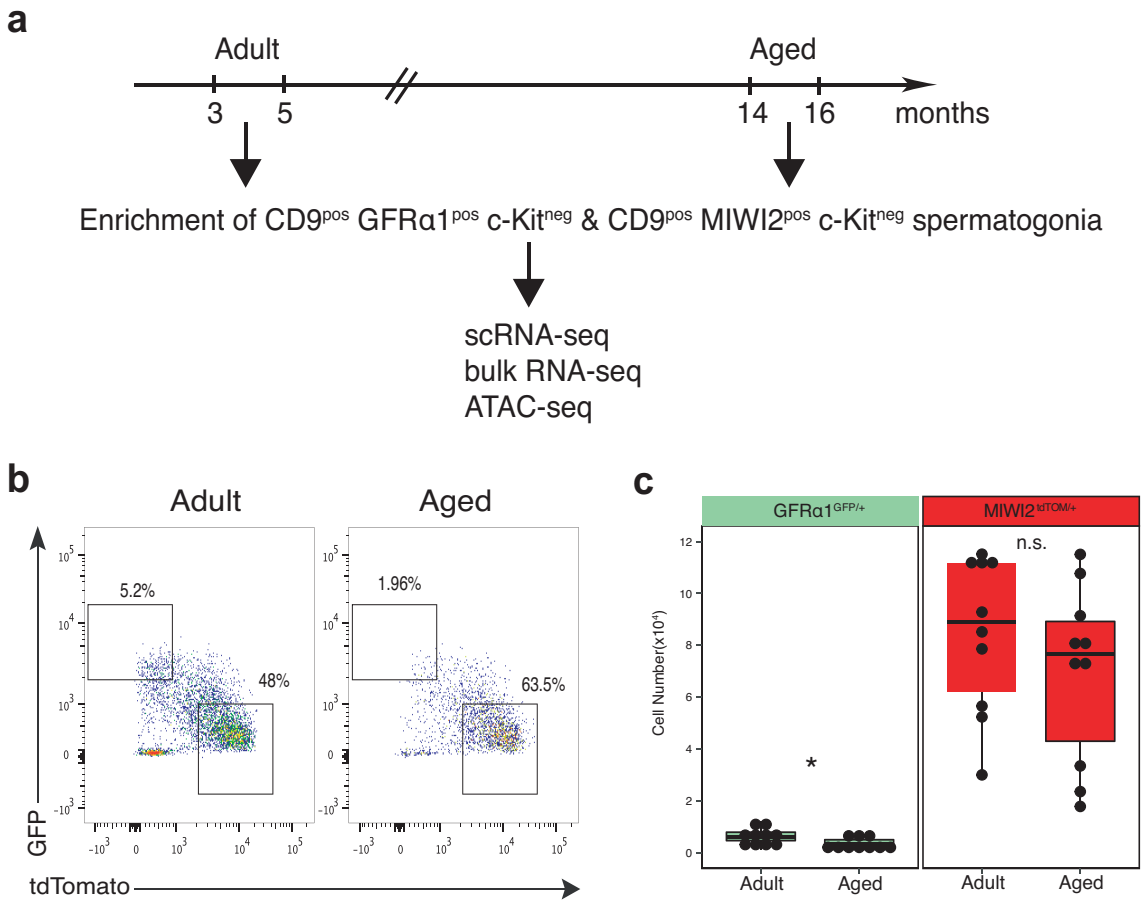


Fig. 3.1: FACS analysis of undifferentiated spermatogonia.

(a) Experimental overview for adult (3-5 months old) and aged (14-16 months old) mice. **(b)** Representative FACS analysis of live GFRα1^{pos} and MIWI2^{pos} gated cells from adult (LHS) and aged (RHS) mice. Numbers indicate the percentages of cells in the defined subpopulations. **(c)** Enumeration of the GFRα1^{pos} (LHS) and MIWI2^{pos} (RHS) populations in adult ($n = 4$) and aged ($n = 3$) mice. The number of GFRα1^{pos} spermatogonia declines with age ($P < 0.05$, unpaired two-tailed Student's t test). n.s., no statistical significance.

The proportions of GFRα1^{pos} and MIWI2^{pos} cells amongst the undifferentiated spermatogonia of adult mice were 5.2% and 48% (Fig. 3.1b). Their proportions in aged mice were 1.96% and 63.5%. Both the proportion and number of GFRα1^{pos} spermatogonia ($P < 0.05$, unpaired two-tailed Student's t test) declined with age (Fig. 3.1c). The number of MIWI2^{pos} spermatogonia did not significantly change.

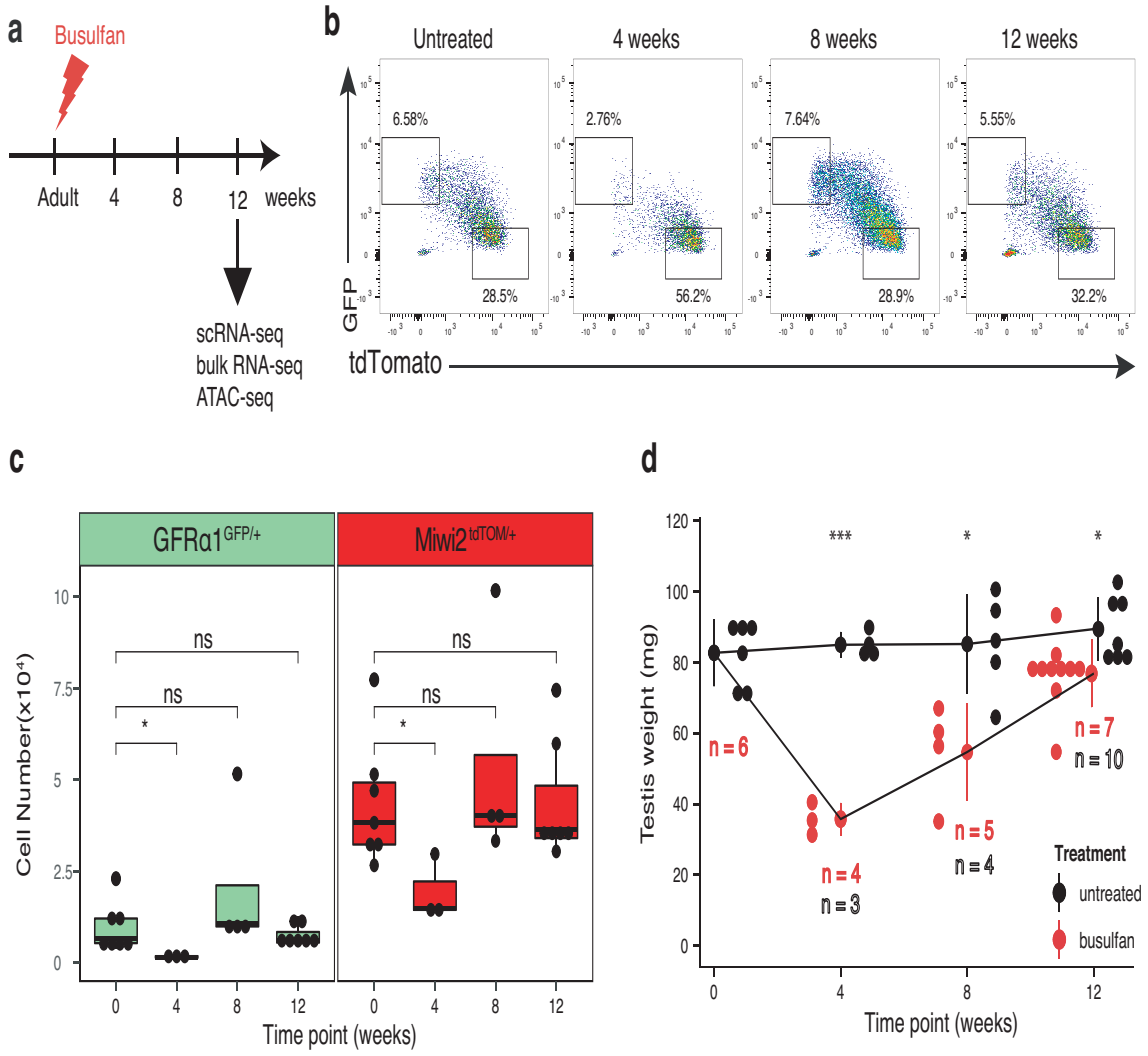


Fig. 3.2: Testis regeneration after busulfan treatment.

(a) Experimental overview for busulfan-treated mice. Adult mice were injected with low concentration busulfan (20 mg/kg of mouse body weight) and their testis allowed recover over a period of twelve weeks. **(b)** Representative FACS analysis of live GFRα1^{pos} and MIWI2^{pos} gated cells from the testes of mice that were untreated or treated with a low concentration busulfan. Analysed at four, eight and twelve weeks after the last injection. **(c)** Enumeration of the GFRα1^{pos} (LHS) and MIWI2^{pos} (RHS) spermatogonial populations in untreated and treated mice. Analysed at four, eight and twelve weeks after the last injection. significance indicated; n.s., no statistical significance. **(d)** Testicular weight for the treated and untreated mice at four time-points. Significance indicated; *n* = number of mice analysed per time point.

Mice from the regenerated cohort were injected with low concentration busulfan (20 mg/kg of mouse body weight) as described in Carrieri et al. (2017). The proportion of GFRα1^{pos}

spermatogonia was lower 4 weeks after the last injection when contrasted with an untreated control but had already started to recover by the second time-point (8 weeks) reaching 7.64% before declining slightly to 5.5% by the final time-point (12 weeks) (Fig. 3.2b). In contrast, the proportion of MIWI2^{pos} spermatogonia increased after busulfan treatment to 56.2% by 4 weeks post-treatment when compared with an untreated control. However the proportion of MIWI2^{pos} cells declined to 28.9% by the second time-point (8 weeks), a level similar to that seen for untreated controls. The absolute number of GFRα1^{pos} and MIWI2^{pos} spermatogonia was also lower 4 weeks after treatment than for untreated controls ($P < 0.05$, unpaired two-tailed Student's t test) (Fig. 3.2c). The numbers increased thereafter and return to physiologically normal levels by the final time point (12 weeks). Mirroring the pattern of decline and recovery in GFRα1^{pos} spermatogonia, the average testicular weight is significantly lower 4 weeks after treatment when compared to untreated controls ($P < 0.001$, unpaired two-tailed Student's t test), see **Fig. 3.2d**. Average testicular weight recovers by the second time-point but is still lower than that of the untreated controls ($P < 0.05$, unpaired two-tailed Student's t test). It continues to increase thereafter, with the average testicular weight of the treated mice approaching that of the untreated controls by the final time-point (12 weeks); however it is lower nonetheless ($P < 0.05$, unpaired two-tailed Student's t test).

3.2.3 scRNA-Seq using the plate-based Smart-seq2 protocol

In order to profile the transcriptional changes that occur with ageing and regeneration at a single-cell level, Ivalya Ivanova (O'Carroll Lab) prepared 75bp single-end scRNA-seq cDNA libraries for undifferentiated spermatogonia from adult, aged and regenerated mouse testis using the Smart-seq2 protocol (Picelli et al., 2014). Seven 96-well plates were prepared from FACS-sorted GFRα1^{pos}, MIWI2^{pos} and GFRα1^{pos} MIWI2^{pos} spermatogonia from adult, aged and regenerated mouse testes. Individual plates contained cells from a single animal. Each plate contained equal numbers of GFRα1^{pos}, MIWI2^{pos} and GFRα1^{pos} MIWI2^{pos} undifferentiated spermatogonia. Non-stranded cDNA libraries were prepared from cell lysates with the Nextera™ XT Kit (Illumina) and the 75bp single-end cDNA libraries sequenced on an Illumina™ NextSeq (Nuffield Division of Clinical Laboratory Sciences, Oxford). Adult and aged samples were in duplicate while regenerated samples were in triplicate. Adult and aged plates were sequenced together while the regenerated plates were sequenced separately.

Raw single-end FASTQs were mapped against the *Mus musculus* genome (mm10) and GENCODE transcript annotation version M17 (Mudge and Harrow, 2015) using *STAR* (v2.7.0f) (Dobin et al., 2013) with its default parameters. Non strand-specific gene counts were quantified from the read alignments using *htseq-count* (v0.9.1) (Anders et al., 2015) by

setting the stranded parameter to "no". QC and cell- and gene-filtering were performed with the R/Bioconductor packages *scater* (v1.10.1) (McCarthy et al., 2017) and *scran* (v1.10.2) (Lun et al., 2016). QC for the two batches (i.e. adult + aged and regenerated) of spermatogonia were performed separately before merging them and performing a final round of QC on the combined dataset. Lowly and un-expressed genes, those with a mean normalised expression < 1 across all cells in a batch and those expressed in a single plate and/or in < 10 % cells were removed. Cells with small library sizes, those whose $\log_{10}$(library size) was 3 median absolute deviations (MADs) lower than the median value, and low library complexity, those whose $\log_{10}$(total features by counts) were 3 MADs lower than the median value, were filtered out. The numbers of cells per remaining after each round of QC are contained in **Table 3.1**.

Table 3.1: Single-cell RNA sequencing Quality Control and Filtering.

Condition	**Adult**	**Aged**	**Regenerated**	**Total**
Marker	GFRα1^{pos}	GFRα1^{pos}	GFRα1^{pos}	
Raw	64 cells	64 cells	96 cells	224
By Lib. size	63 cells	60 cells	95 cells	218
By Genes	60 cells	57 cells	75 cells	192
By Diversity	51 cells	41 cells	71 cells	163
Final	51 cells	41 cells	62 cells	154
Marker	MIWI2^{pos}	MIWI2^{pos}	MIWI2^{pos}	
Raw	64 cells	64 cells	96 cells	224
By Lib. size	61 cells	55 cells	92 cells	208
By Genes	56 cells	51 cells	74 cells	181
By Diversity	48 cells	42 cells	68 cells	158
Final	48 cells	42 cells	61 cells	151
Marker	GFRα1^{pos} MIWI2^{pos}	GFRα1^{pos} MIWI2^{pos}	GFRα1^{pos} MIWI2^{pos}	
Raw	64 cells	64 cells	96 cells	224
By Lib. size	62 cells	62 cells	94 cells	218
By Genes	57 cells	62 cells	92 cells	211
By Diversity	44 cells	54 cells	91 cells	189
Final	44 cells	54 cells	61 cells	159

The single-cell expression was normalised using scran's *computeSumFactors* function that uses a deconvolution method that borrows information from neighbouring (i.e. transcriptionally similar) cells in order to calculate sum factors for count normalisation. All further analyses looking at single-cell RNA expression use the $\log_2$-transformed, normalised counts (including a pseudocount) i.e. $\log_2$(norm exprs + 1). The spermatogonia that passed the QC filtering per batch are visualised with the t-Distributed Stochastic Neighbour Embedding (t-SNE) dimensionality reduction technique in **Fig. 3.3**. Adult, aged and regenerated spermatogonia are separated based on their GFRα1 and MIWI2 expression status (Fig. 3.3a & d). The GFRα1^{pos} MIWI2^{pos} cells are transcriptionally intermediate between but also interspersed with the GFRα1^{pos} and MIWI2^{pos} spermatogonia.

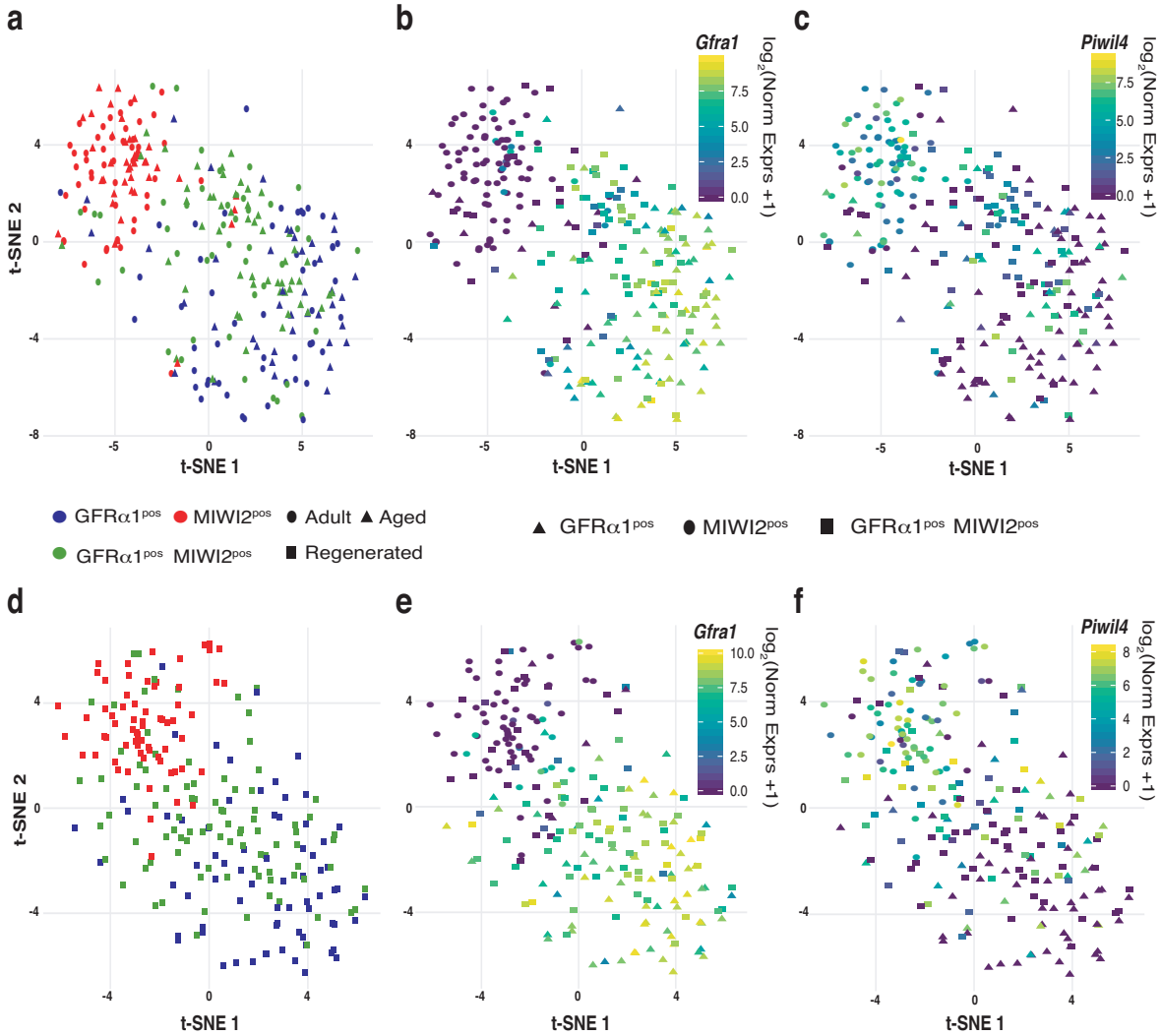


Fig. 3.3: Visualisation of filtered single spermatogonia.

Fig. 3.3: Visualisation of filtered single spermatogonia (continued).
(a) t-SNE dimensionality reduction for undifferentiated spermatogonia isolated from the testes of adult and aged mice. Spermatogonia are coloured based on their GFRα1 and MIWI2 expression. Their shape is based on the testes of origin, adult = circles, aged = triangles. The cells are coloured based on their $\log_2$(Normalised *Gfra1* Expression) and $\log_2$(Normalised *Piwil4* Expression) in **(b)** and **(c)**. **(d)** t-SNE dimensionality reduction for undifferentiated spermatogonia isolated from the regenerated testes of adult mice. Spermatogonia are coloured based on their GFRα1 and MIWI2 expression. Their shape is based on the testes of origin, regenerated = squares. The cells are coloured based on their $\log_2$(Normalised *Gfra1* Expression) and $\log_2$(Normalised *Piwil4* Expression) in **(e)** and **(f)**.

Generally, a cell's GFRα1 and MIWI2 status is reflected in the expression of the genes encoding them, *Gfra1* (Fig. 3.3b & e) and *Piwil4* (Fig. 3.3c & f). *Gfra1* is expressed in GFR$\alpha1^{pos}$ and GFR$\alpha1^{pos}$ MIWI2^{pos} spermatogonia and is largely absent from MIWI2^{pos} spermatogonia. Its expression is higher in GFR$\alpha1^{pos}$ than GFR$\alpha1^{pos}$ MIWI2^{pos} spermatogonia. Similarly, *Piwil4* is expressed in MIWI2^{pos} and most GFR$\alpha1^{pos}$ MIWI2^{pos} spermatogonia but is absent from majority of GFR$\alpha1^{pos}$ spermatogonia. Its expression is highest in MIWI2^{pos} spermatogonia.

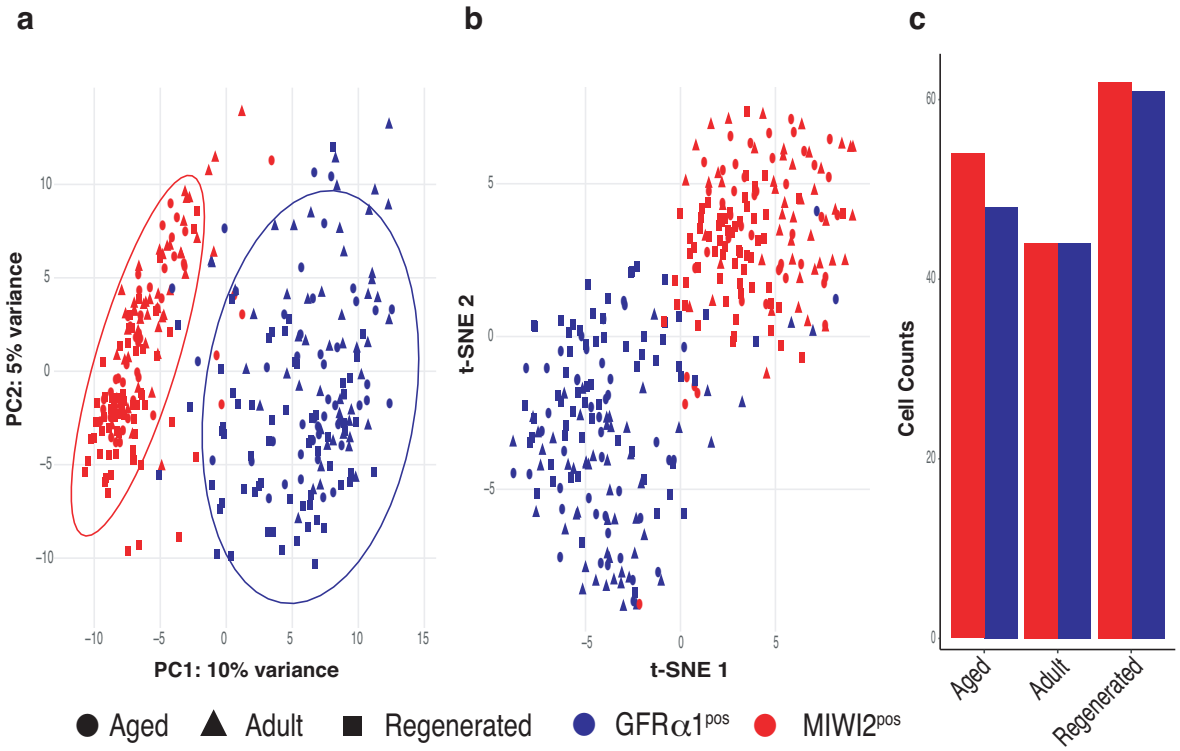


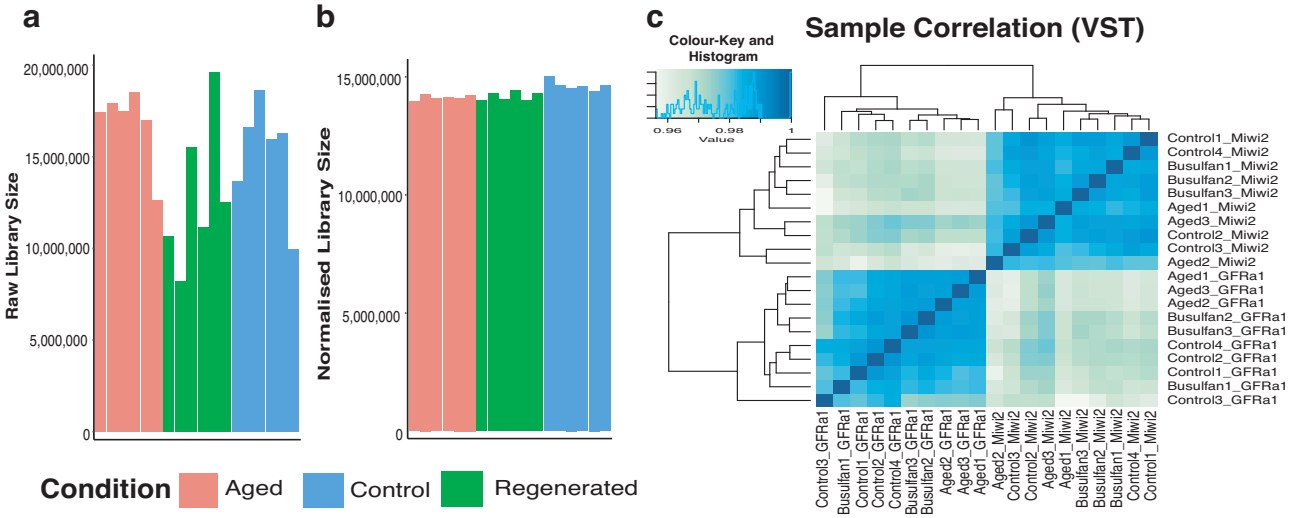
Fig. 3.4: Visualisation of filtered GFR$\alpha1^{pos}$ and MIWI2^{pos} spermatogonia.

Fig. 3.4: Visualisation of filtered GFRα1^{pos} and MIWI2^{pos} spermatogonia (continued).
(a) PCA and **(b)** t-SNE dimensionality reduction for the final filtered GFRα1^{pos} and MIWI2^{pos} spermatogonia isolated from the adult, aged and regenerated mouse testes. Cells are coloured based on their GFRα1 and MIWI2 expression status, GFRα1^{pos} = blue and MIWI2^{pos} = red. Their shape reflects the testes of origin; adult = circle , aged = triangle and regenerated = square. PC1 accounts for 10% of the variance in gene expression and separates out the GFRα1^{pos} and MIWI2^{pos} spermatogonia. PC2 accounts for 5% of the variance in gene expression. Centroids show the average positions of the GFRα1^{pos} and MIWI2^{pos} spermatogonia in the PCA. **(c)** Final cell numbers after QC and cell filtering. 313 cells remained, of which 153 are GFRα1^{pos} and 160 MIWI2^{pos}. 102 cells are from adult testes, 88 from aged testes and 123 from regenerated testes.

The GFRα1^{pos} MIWI2^{pos} spermatogonia were excluded from subsequent analyses because they are not as distinctive a subpopulation and as the GFRα1^{pos} and MIWI2^{pos} spermatogonia. All cells from the Busulfan2 plate were excluded due to the low number of cells (21) remaining when the GFRα1^{pos} MIWI2^{pos} were excluded. A total of 313 cells remained after this final filtering step, the filtered cells and cell numbers are displayed in **Fig. 3.4**. There did not seem to be any major batch- or plate-effect, all cells are found within two major GFRα1^{pos} and MIWI2^{pos} clusters (Fig. 3.4 a & b).
The final set of 313 cells and 11,168 genes was used as input for single-cell differential gene expression and variability testing analyses. All GFRα1^{pos} and MIWI2^{pos} gene counts were summed per plate to create six 'pseudo-bulk' samples that were used to perform standard differential gene expression analysis with *DESeq2* (v1.22.2) (Love et al., 2014) using a "local" fit. Summing counts per plate minimises plate-specific effects and overcomes any confounding relationship between plate and biological groupings (Lun and Marioni, 2017) i.e. the testis types. It is robust against differing cell library sizes and unequal numbers of cells per plate. The R/Bioconductor package *BASiCS* (v1.4.7) (Eling et al., 2018; Vallejos et al., 2015, 2016) was used to detect genes that exhibit greater cell-to-cell heterogeneity, as measured by over-dispersion, and perform differential variability testing in a pairwise fashion. In the absence of spike-in controls from single-cell library preparation, the BASiCS regression model borrows information between plates to estimate the technical sources of variance. This is based on the assumption that biological sources of variability will be shared across cells on the different plates but technical sources will be plate specific. This allows for the estimation of a residual over-dispersion value per gene per condition that is not confounded by its mean expression. Only genes that are expressed in at least 2 cells (in both groups) are tested. The BASiCS regression model was run for 20,000 iterations with 10,000 burn-in iterations and a thinning value of 20. *goseq* (v1.34.1) (Young et al., 2010) was used for functional enrichment analyses in differentially expressed genes.

3.2.4 Bulk RNA-sequencing of undifferentiated spermatogonia

Ivalya Ivanova (O'Carroll Lab) prepared 75bp single-end cDNA libraries for GFRα1^{pos} and MIWI2^{pos} spermatogonia from adult, aged and regenerated mouse testes. Samples were generated by pooling 100 FACS-sorted cells from the testes of a single animal. Non-stranded cDNA libraries were prepared with the Nextera™ XT Kit (Illumina) and the 75bp single-end cDNA libraries sequenced on an Illumina™ NextSeq (Nuffield Division of Clinical Laboratory Sciences, Oxford). Aged and regenerated samples were in triplicate while the control, adult samples were in quadruplicate. There were two technical replicates per sample. Raw single-end FASTQs were mapped against the mouse genome (mm10) and GENCODE transcript annotation version M17 using STAR (v2.7.0f) with its default parameters. Non strand-specific gene counts were quantified from the read alignments using htseq-count (v0.9.1) by setting the stranded parameter to "no". Counts from the technical replicates were combined. Size factor normalisation and differential gene expression analyses were performed with DESeq2 (v1.22.2) as described previously. Sample outliers were detected by performing Principal component analysis (PCA) and sample-to-sample correlation with gene expression counts normalised by the DESeq2 variance-stabilising transformation (VST), see **Fig. 3.5**. The GFRα1^{pos} sample from the Control3 mouse was an outlier in both analyses (Fig. 3.5c & d); this and the corresponding MIWI2^{pos} sample from the same animal were removed from further analyses. The results of the PCA for the filtered samples are in **Fig. 3.5e**. A PCA loadings plot with the top 30 genes contributing to the first two components is shown in **Fig. 3.5f**. *Dusp6* and contribute the most to PC1 while *Hprt* makes the largest contribution to PC2.



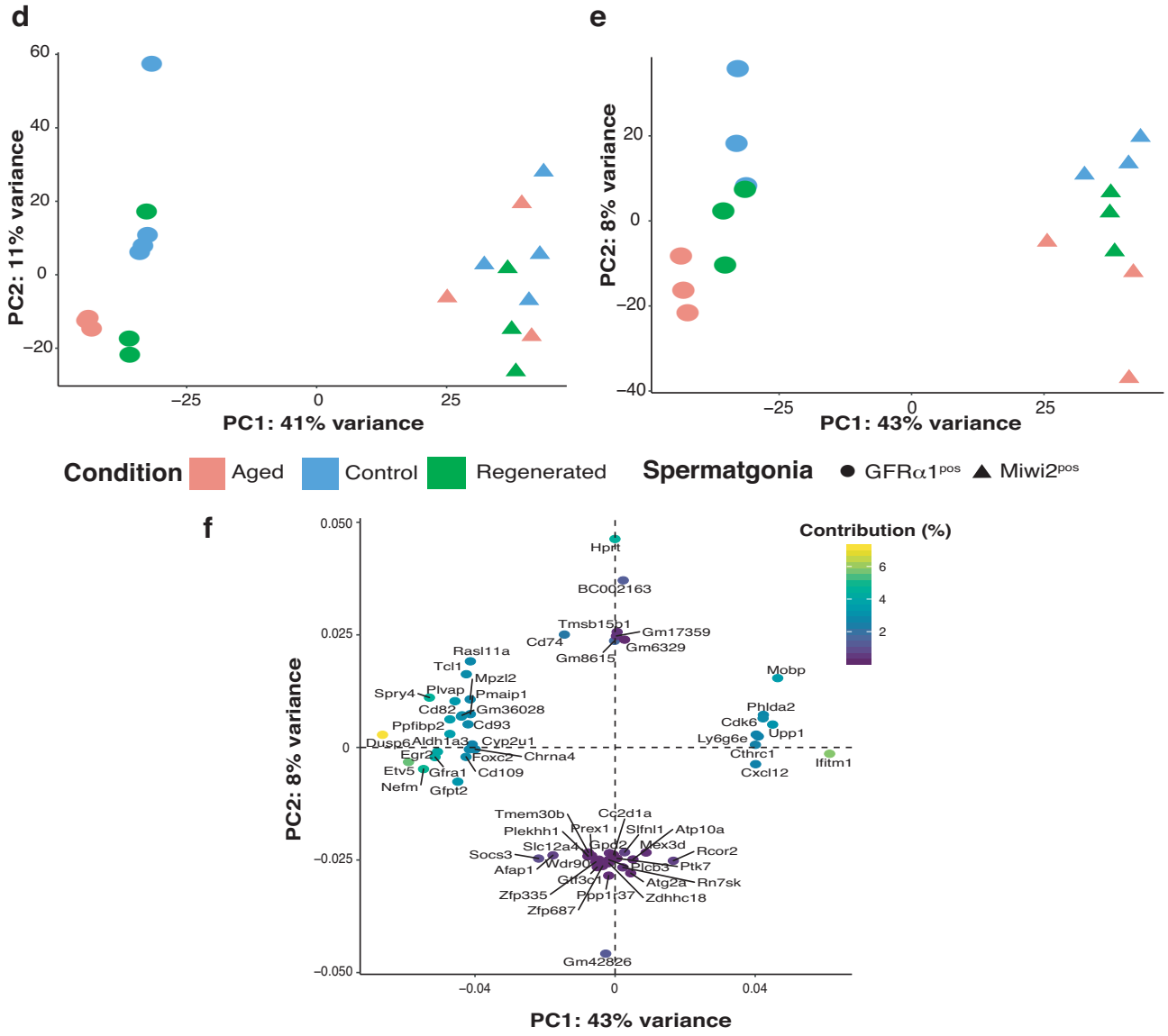


Fig. 3.5: Quality Control for bulk RNA-seq data.

(a) Raw and **(b)** DESeq2-normalised sequencing library sizes. **(c)** Hierarchical clustering of bulk RNA-seq samples based on the sample-to-sample Pearson correlation of their VST counts. The main bifurcation separates GFRα1^{pos} and MIWI2^{pos} samples. Within these two branches, samples tend to cluster based on whether they are derived from adult, aged or regenerated testis. The GFRα1^{pos} adult sample derived from the Control3 mouse appeared to be an outlier compared to the rest of the GFRα1^{pos} samples. **(d)** PCA of raw and **(e)** filtered bulk RNA-seq samples based on VST counts. Samples are coloured based on the testis types: blue = adult, red = aged, green = regenerated. Circles are GFRα1^{pos} spermatogonia and triangles MIWI2^{pos} spermatogonia. All GFRα1^{pos} samples are on the LHS while all MIWI2^{pos} samples on the RHS. PC1 defines an axis based on the marker expressed and is responsible for 43% of sample-to-sample variation in the filtered samples. PC2 defines an axis based on the testis type of origin and is responsible for 8% of sample-to-sample variation in the filtered samples. **(f)** Plot of PCA loadings for the filtered samples. Top 30 genes contributing to the first two components. (GFRα1^{pos}, *n* adult = 4, *n* aged = 3, *n* regenerated = 3; MIWI2^{pos}, *n* adult = 4, *n* aged = 3, *n* regenerated = 3).

Again, *goseq* (v1.34.1) was used for functional enrichment analyses in differentially expressed genes. Additionally, the expression of TEs in the undifferentiated spermatogonia was determined by mapping the raw FASTQs against *RepBase* rodent repeats (rodrep) (Bao et al., 2015) using *Bowtie 2* (v2.3.5) (Langmead and Salzberg, 2012) in single-end mode. Uniquely mapping reads were used to generate repeat family and subfamily counts for each sample. These were summed to generate a master table of repeat hits per sample. Repeats with a count of ≤ 1 across all samples were excluded. This table of counts was processed for differential repeat expression as before using DESeq2.

3.2.5 | Chromatin Accessibility from ATAC-seq data

As discussed in **Chapter 1** (1.1.1), chromatin accessibility at/around TSSs is a major determinant of whether or not transcription is permissible. Open chromatin is more accessible to the binding of transcriptional regulators. We used ATAC-seq (Buenrostro et al., 2015a) to study the euchromatin and heterochromatin composition in undifferentiated spermatogonia. ATAC-seq provides a genome-wide view of DNA accessibility based on the frequency at which purified Tn5 transposase 'tags' open chromatin with sequencing adaptors (Buenrostro et al., 2013). PCR amplification of the transposed adaptor sequence produces narrow peaks centred on regions of accessible chromatin when sequenced with NGS. Ivalya Ivanova (O'Carroll Lab) prepared 40bp paired-end ATAC-seq DNA libraries from the nuclei of FACS-sorted GFRα1^{pos} and MIWI2^{pos} single-cell suspensions from adult, aged and regenerated mouse testes. Briefly, Ivalya collected 20,000 sorted cells per population in a 1.5 ml eppendorf tubes. Cells were spun down and washed with PBS. Nuclei were extracted with lysis buffer (10mM Tris pH7.4, 10mM NaCl, 3mM $MgCl_2$, 0.1% Igepal CA-630 and 1x protease inhibitor). Pelleted nuclei were re-suspended in a Tn5 mix (10mM Tris pH8, 5mM MgCl2, 10% di-methyl formamide and Tn5 transposase) and incubated at 37^{o}C for 30 minutes shaking at 500rpm. Transposed DNA was then purified with MinElute kit (QIAGEN) and was amplified with low-plex sequencing primers for 5 cycles. A quantitative polymerase chain reaction (qPCR) determined that an additional 7 to 10 cycles were required per sample. Sequencing libraries were purified with Agencourt™ AMPure beads and the quality checked on a Tape station (DNA5000HS) and Qubit. The libraries were sequenced on an Illumina™ NextSeq (Nuffield Division of Clinical Laboratory Sciences, Oxford). All samples were generated in biological triplicate.

Accessible regions of the mouse genome were detected using the ENCODE DCC ATAC-seq analysis pipeline developed by the Kundaje Lab (Koh et al., 2016). Briefly, read adaptors are trimmed from the FASTQs and mapped against the mm10 reference using Bowtie 2. Unmapped reads are filtered out using *samtools* (Li et al., 2009) and duplicate reads identified

using *Picard* (http://broadinstitute.github.io/picard). Peak calling is performed using *MACS2* (v2.1.1.20160309) (Zhang et al., 2008) and those peaks overlapping ENCODE-blacklisted genomic regions (Amemiya et al., 2019) are removed with *bedtools* (v2.25.0) (Quinlan and Hall, 2010). A list of concordant 'optimal' and 'conservative' ATAC-seq peaks was derived from the peaks present in 2 out of 3 replicates per condition. The lists of optimal and conservative ATAC-seq peaks were merged and sorted with bedtools to yield a final list of peaks per sample. These peaks were annotated with genomic-features using the R/Bioconductor library *ChIPseeker* (Yu et al., 2015). Analysis of Reactome Pathway enrichment for the genes overlapping ATAC-seq peaks was performed using *clusterProfiler* (Yu et al., 2012) and *ReactomePA* (Yu and He, 2016). Bedtools was used to filter peaks based on GENCODE transcript annotation M17 and obtain sequences for genic and intergenic peaks in FASTA format. Analysis of TF binding motifs in these cohorts was performed with *MEME-ChIP* (v5.0.2) (Machanick and Bailey, 2011) using the HOCOMOCO database of mouse TF motifs (v11) (Kulakovskiy et al., 2018). The results of the ATAC-seq QC are shown in **Fig. 3.6**.

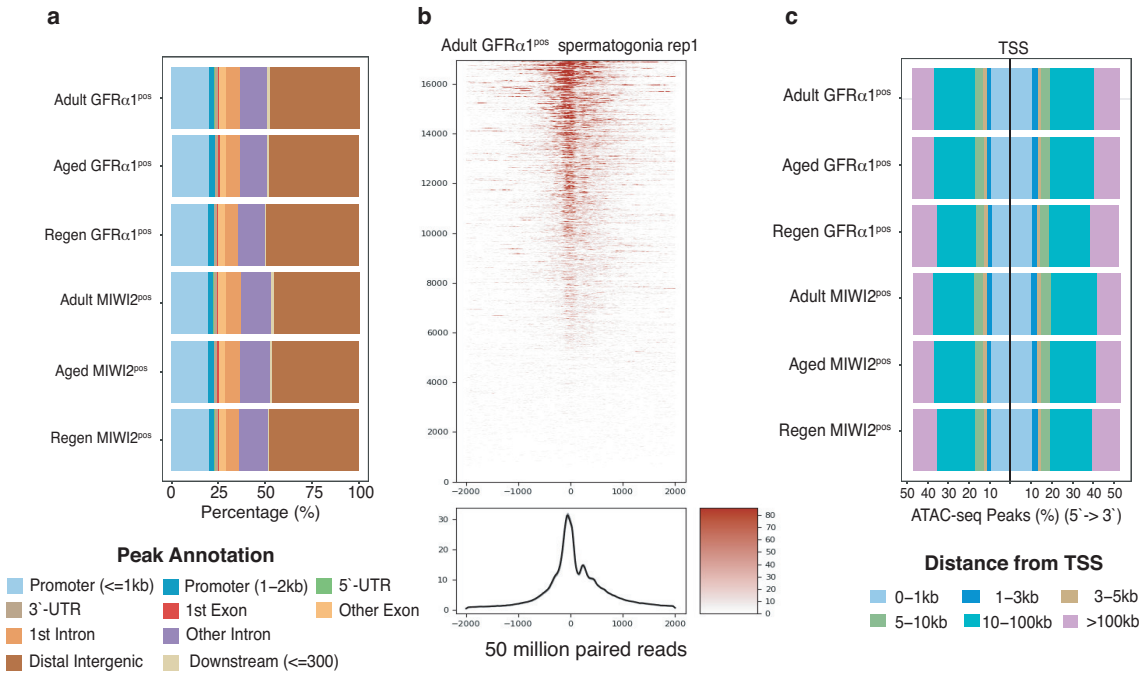


Fig. 3.6: Quality Control for ATAC-seq data.

(a) Genomic feature annotation for all ATAC-seq peaks. Cyan: $\leq$1 kb from a promoter, light brown: 3′ UTRs, orange: first exon, etc. **(b)** Representative aggregation plot for GFRα1^{pos} replicate 1 showing the aggregated ATAC-seq signal for all TSSs throughout the mouse genome in a 4 kb window centred on GENCODE-annotated TSSs. TSSs are sorted based on ATAC-seq peak signal intensity.

Fig. 3.6: Quality Control for ATAC-seq data (continued).
(c) Distance of ATAC-seq peaks from TSSs. Cyan: $\leq$ 1 kb, blue: 1-3 kb, orange: 3-5 kb, etc. (GFRα1^{pos}, *n* adult = 3, *n* aged = 3, *n* regenerated = 3; MIWI2^{pos}, *n* adult = 3, *n* aged = 3, *n* regenerated = 3).

The genome-wide, metagene plot in **Fig. 3.6** shows a characteristic ATAC-seq bimodal peak 'footprint' centred on the TSSs. The final number of ATAC-seq peaks per condition was as follows: 106,589 peaks from adult GFRα1^{pos} spermatogonia, 106,921 from aged GFRα1^{pos} spermatogonia, 104,881 from regenerated GFRα1^{pos} spermatogonia, 106,291 from adult MIWI2^{pos}, 104,804 from aged MIWI2^{pos} and 97,589 from regenerated MIWI2^{pos} spermatogonia. While the majority of peaks are found in distal intergenic regions, an average of 20% are within 1 kb of a TSS (Fig. 3.6a). The TSS-associated peaks exhibit the narrow distribution centred on TSSs characteristic of ATAC-seq (Fig. 3.6b). Although only 20% of peaks are within 1 kb of a GENCODE-annotated TSS, 45% are within 10 kb of a TSS (Fig. 3.6c).

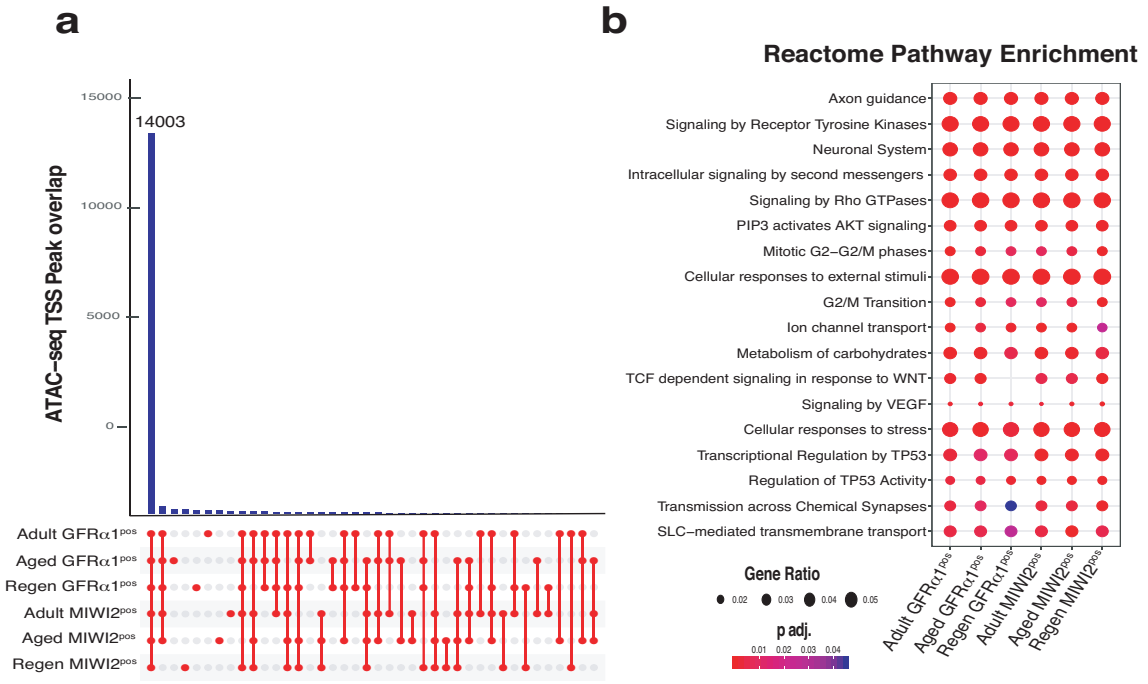


Fig. 3.7: Chromatin accessibility at transcription start sites.
(a) Upset plot showing the overlap between genes that have ATAC-seq peaks within 1 kb of their TSSs in each condition. The histogram shows the number of genes with TSS-peaks and the red lines below show which conditions are share them. **(b)** Reactome pathway enrichment for genes with ATAC-seq peaks. (GFRα1^{pos}, *n* adult = 3, *n* aged = 3, *n* regenerated = 3; MIWI2^{pos}, *n* adult = 3, *n* aged = 3, *n* regenerated = 3).

Of the 16,653 genes that have an ATAC-seq peak within 1 kb of their TSS in any condition, 14,503 (87%) have a peak in all conditions (Fig. 3.7a). Furthermore, these ATAC-seq-associated genes are involved in broadly similar biological pathways in all conditions (Fig. 3.7b). There is a consistent enrichment of Reactome cell-signalling, transcriptional regulation and mitotic pathways. MEME-ChiP identified a number of enriched TF-binding sites amongst the genic and intergenic ATAC-seq peaks. A motif similar to the SP1/2/3 motif was detected from the genic ATAC-seq peaks in all conditions, see **Fig. 3.8a** for the motif in adult GFRα1^{pos} spermatogonia. Another genic motif similar to a combined Signal transducer and activator of transcription 1 (STAT1) forkhead box protein J3 isoform X2 (FOXJ3) motif was found in all conditions with the exception of adult MIWI2^{pos} spermatogonia (Fig. 3.8b). Likewise a motif similar to the CTCF CCCTC-binding factor like (CTCFL) motif was identified from the intergenic ATAC-seq peaks in all conditions(Fig. 3.8c & d). Another intergenic motif similar to a combined *doublesex* and *mab-3*-related transcription factor 1 (DMRT1) *doublesex* and *mab-3*-related transcription factor B1 (DMRTB) SRY-Box 15 (SOX15) motif was detected in all samples with the exception of regenerated MIWI2^{pos} spermatogonia.

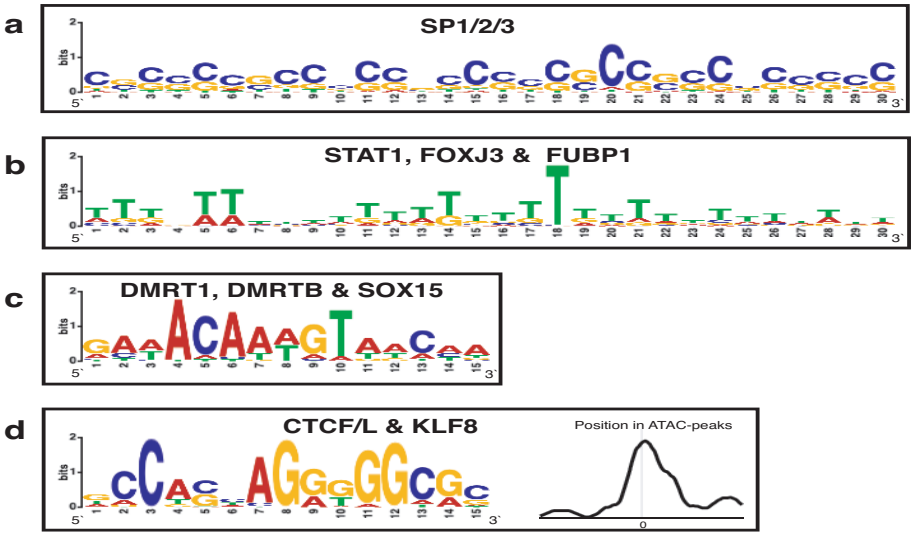


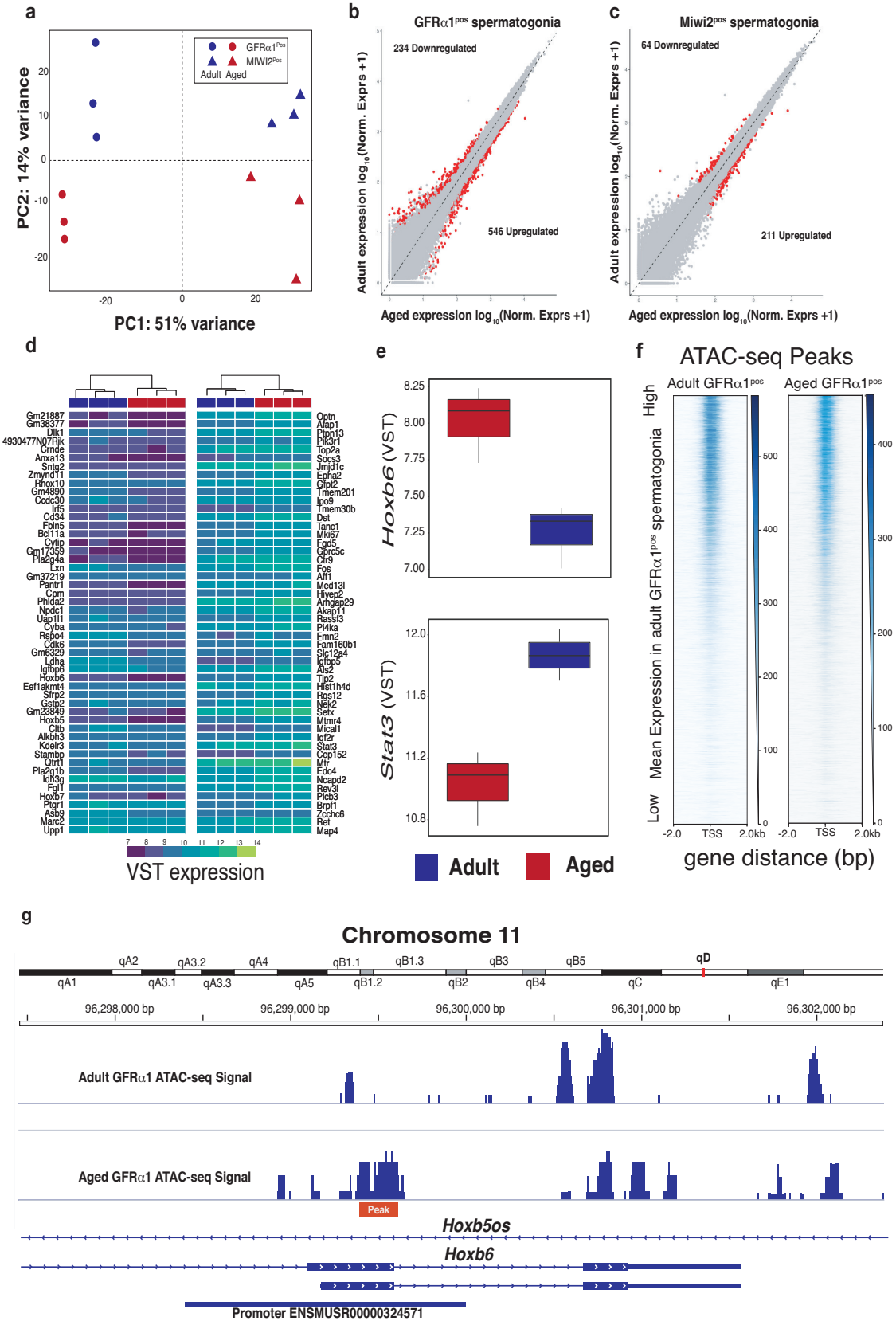
Fig. 3.8: Transcription factor binding motifs in regions of accessible chromatin for adult GFRα1^{pos} spermatogonia.

The two most significant binding motifs identified from genic **(a)** & **(b)** and intergenic **(c)** & **(d)** ATAC-seq peaks in adult GFRα1^{pos} spermatogonia. **(a)** is most similar to the SP1, SP2 and SP3 binding motif. **(b)** most resembles a combined STAT1, FOXJ3, FUBP1 binding motif while **(c)** is most like to a combined DMRT1, DMRTB and SOX15 binding motif. **(d)** is most similar to a combined CTCF/L and KLF8 binding motif and is enriched at the ATAC-seq peaks themselves. (GFRα1^{pos}, *n* adult = 3, *n* aged = 3, *n* regenerated = 3; MIWI2^{pos}, *n* adult = 3, *n* aged = 3, *n* regenerated = 3).

3.2.6 Mean gene expression is altered with age

Analysis of the bulk RNA-seq data from adult and aged GFRα1^{pos} and MIWI2^{pos} spermatogonia revealed that the expression of GFRα1 and MIWI2 is the main source of difference between samples. From PCA of VST gene expression, PC1 separates samples based on their GFRα1 MIWI2 expression status and is responsible for 51% of the variance in gene expression (Fig. 3.9a). In contrast PC2, which coincides with the age of the mouse of origin, is responsible for 14% of the variance in gene expression.

Ageing had a subtle effect on mean expression. An absolute fold change of 1.7 was used as the threshold for differential gene expression analyses as a result. 546 genes are significantly up- and 234 down-regulated between aged and adult GFRα1^{pos} spermatogonial samples, greater than 1.7-fold increase (adj. $P < 0.05$, Wald test) (Fig. 3.9b). VST gene expression for the top 50 most down- and up-regulated genes is shown in **Fig. 3.9d**. *Hoxb6* and *Stat3* are differentially expressed genes whose expression has previously been shown to alter with ageing (Fig. 3.9e). 211 genes are up- and 64 down-regulated when the same contrast is made for the MIWI2^{pos} spermatogonial samples (Fig. 3.9c). Chromatin is more accessible at the TSSs of genes that are highly expressed in adult GFRα1^{pos} spermatogonia than those of less abundant transcripts (Fig. 3.9f). ATAC-seq peaks from adult and aged GFRα1^{pos} samples are both more frequent at the TSSs of these highly expressed genes. While the ATAC-seq peak profiles for adult and aged GFRα1^{pos} spermatogonia are very similar there are examples of differences such as at the promoter of the differentially expressed gene *Hoxb6* on chromsome 11 (Fig. 3.9g). We detected enrichment of gene ontology (GO) Molecular Function categories and Reactome pathways associated with ribosome assembly and translation amongst the genes that are differentially expressed with age in GFRα1^{pos} undifferentiated spermatogonia (adjusted $p < 0.05$) (Fig. 3.9h).


a
GFRα1Pos
MIWI2Pos
Adult Aged
PC2: 14% variance
PC1: 51% variance
b
GFRα1pos spermatogonia
234 Downregulated
546 Upregulated
Adult expression log10(Norm. Exprs +1)
Aged expression log10(Norm. Exprs +1)
c
Miwi2pos spermatogonia
64 Downregulated
211 Upregulated
Adult expression log10(Norm. Exprs +1)
Aged expression log10(Norm. Exprs +1)
d
VST expression
e
Hoxb6 (VST)
Stat3 (VST)
Adult
Aged
f
ATAC-seq Peaks
Adult GFRα1pos
Aged GFRα1pos
High
Low
Mean Expression in adult GFRα1pos spermatogonia
TSS
gene distance (bp)
g
Chromosome 11
Adult GFRα1 ATAC-seq Signal
Aged GFRα1 ATAC-seq Signal
Peak
Hoxb5os
Hoxb6
Promoter ENSMUSR00000324571

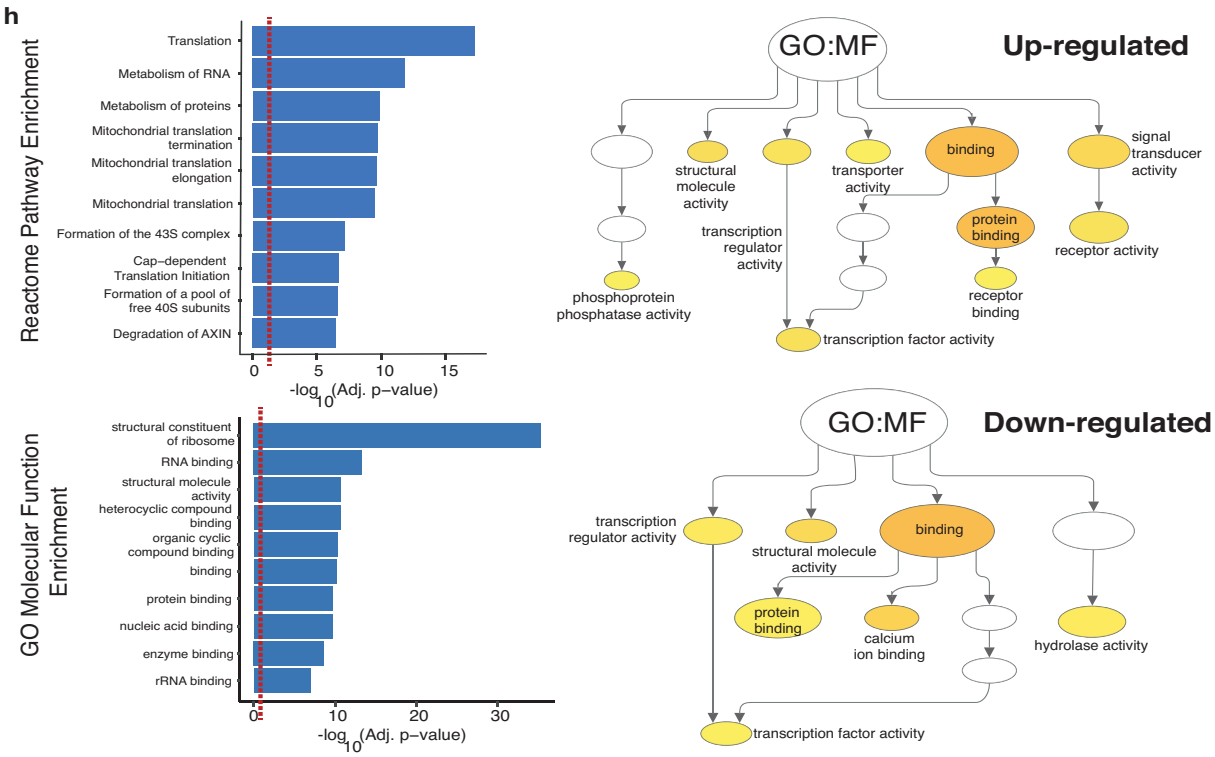


Fig. 3.9: Impact of age on spermatogonial mean expression.

(a) PCA based on mean VST expression levels in bulk RNA-seq samples. GFRα1^{pos} = circles and MIWI2^{pos} = triangles. Blue = adult and red = aged spermatogonia. **(b)** Scatterplot of mean normalised expression levels in adult and aged GFRα1^{pos} spermatogonia. Significantly changing genes, greater than 1.7-fold (adj. $P < 0.05$, Wald test), are highlighted in red. **(c)** Scatter plot of mean normalised expression levels in adult and aged MIWI2^{pos} spermatogonia. Significantly changing genes are highlighted in red. **(d)** VST expression for the top 50 most down- (LHS) and up-regulated (RHS) genes between aged and adult GFRα1^{pos} spermatogonia. Rows are genes and columns are GFRα1^{pos} spermatogonia bulk RNA-seq samples. **(e)** Boxplots of *Hoxb6*, down-regulated in aged GFRα1^{pos} spermatogonia, and *Stat3*, up-regulated in aged GFRα1^{pos} spermatogonia, VST expression. **(f)** Adult and aged GFRα1^{pos} ATAC-seq peak intensities across a 4 kb window centred on the GENCODE-annotated TSSs of adult GFRα1^{pos}-expressed genes. Peaks sorted by mean VST expression in adult GFRα1^{pos} bulk RNA-seq samples. **(g)** Differential chromatin accessibility peak overlapping the promoter of a *Hoxb6* isoform in GFRα1^{pos} spermatogonia. The peak present in aged GFRα1^{pos} spermatogonia but is absent from younger, adult cells. **(h)** Reactome Pathway and GO enrichment analyses for aged vs. adult GFRα1^{pos} spermatogonia. Benjamini–Hochberg corrected p-values (adjusted p-values) were used to visualise GO and Reactome enrichment. The statistical significance threshold was set at adjusted p-value = 0.05 (red line). (Adult GFRα1^{pos}, n = 3; aged GFRα1^{pos}, n = 3 ; adult MIWI2^{pos}, n = 3; aged MIWI2^{pos}, n = 3).

As is already apparent from **Fig. 3.9a**, GFRα1^{pos} and MIWI2^{pos} undifferentiated spermatogonia are very distinct populations. More specifically, they are distinguishable based on the differences in mean expression between them. There are 1,738 genes upregulated between adult GFRα1^{pos} and MIWI2^{pos} undifferentiated cells and 1,980 downregulated, greater than 1.7-fold increase (adj. $P < 0.05$, Wald test).
Likewise, 2,062 genes are upregulated between aged GFRα1^{pos} and MIWI2^{pos} cells and 2,540 downregulated. There is strong overlap between the results of these contrasts, similar sets of genes are differentially expressed (Fig. 3.10a & b) and many of the same molecular functions are affected in both (Fig. 3.10c vs. d). Common functions include RNAP II promoter- and DNA-binding as well as TF and signal receptor binding.

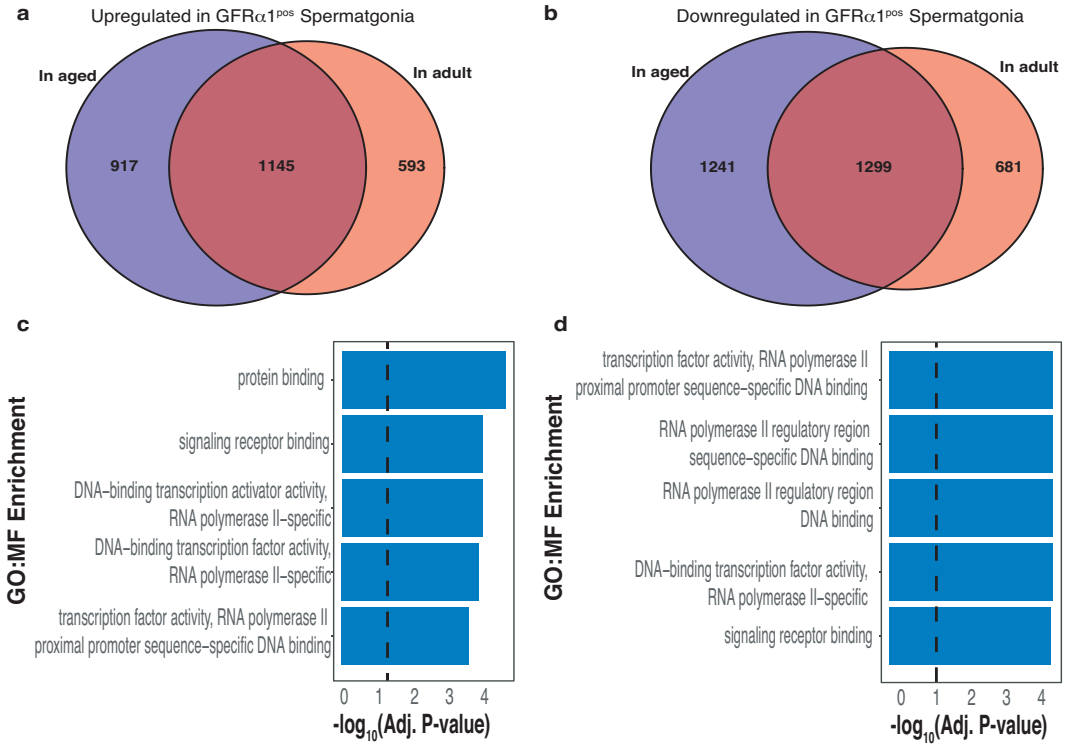


Fig. 3.10: GFRα1 and MIWI2 are expressed in distinct spermatogonial sub-populations. **(a)** and **(b)** show the overlap between differentially genes from adult GFRα1^{pos} and MIWI2^{pos} undifferentiated spermatogonia and between aged GFRα1^{pos} and MIWI2^{pos} cells. GO enrichment analyses for adult **(c)** and aged **(d)** GFRα1^{pos} and MIWI2^{pos} undifferentiated spermatogonia. Benjamini–Hochberg corrected p-values (adjusted p-values) were used to visualise enrichment the top 5 GO molecular function categories. The statistical significance threshold was set at adjusted p-value = 0.05 (dashed line). (Adult GFRα1^{pos}, $n = 3$; aged GFRα1^{pos}, $n = 3$; adult MIWI2^{pos}, $n = 3$; aged MIWI2^{pos}, $n = 3$).

3.2.7 Changes in single-cell expression with age

Similar to the results of the bulk RNA-seq analyses, t-SNE dimensionality of the scRNA-seq data shows that there is a clear separation of undifferentiated spermatogonia based on their GFRα1 and MIWI2 state (Fig. **Fig. 3.11a**). In contrast with the bulk RNA-seq samples, there is no separation based on the age of the animals. *Gfra1* is expressed in the vast majority of adult and aged GFRα1^{pos} spermatogonia and largely absent from adult and aged MIWI2^{pos} spermatogonia (**Fig. 3.11b**).

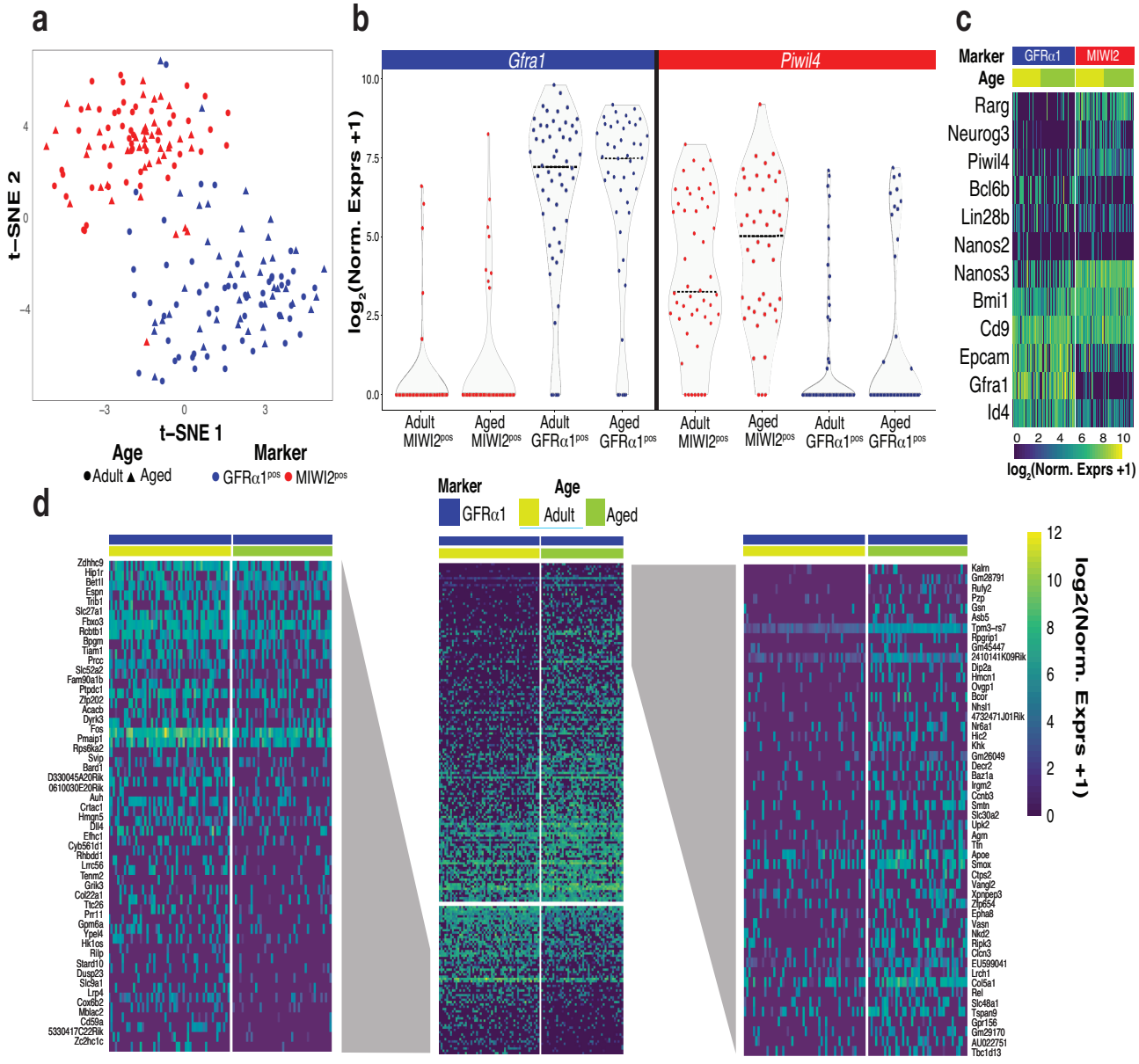


Fig. 3.11: Impact of age on individual undifferentiated spermatogonia.

Fig. 3.11: Impact of age on individual undifferentiated spermatogonia (continued).
(a) t-SNE representation of undifferentiated spermatogonia isolated from the testes of adult and aged mice. GFRα1^{pos} spermatogonia are in blue and MIWI2^{pos} in red. Their shape is based on the testes type of origin, adult = circles, aged = triangles. **(b)** Violin plots based on normalised *Gfra1* (LHS) and *Piwil4* (RHS) gene expression in single GFRα1^{pos} and MIWI2^{pos} spermatogonia. **(c)** Heat map of normalised gene expression in single GFRα1^{pos} and MIWI2^{pos} spermatogonia. Cells are columns and genes rows. The genes selected are known to be involved in spermatogenesis. Cells are annotated at the top of the heat map based on their marker expression and testes of origin. GFRα1^{pos} spermatogonia are in blue and MIWI2^{pos} in red. Cells from adult testes are in yellow and aged in green.**(d)** Heat map of normalised gene expression for differentially expressed genes in aged vs. adult GFRα1^{pos} spermatogonia. Significantly changing genes, greater than 1.7-fold (adj. $P < 0.05$, Wald test). 146 genes up-regulated and 64 down-regulated with age. (Adult GFRα1^{pos} cells, $n = 54$; aged GFRα1^{pos} cells, $n = 44$).

The inverse pattern is observed for*Piwil4*, it is expressed in the vast majority of adult and aged MIWI2^{pos} spermatogonia but is largely absent from the GFRα1^{pos} spermatogonia. GFRα1^{pos} and MIWI2^{pos} have distinctive expression patterns for a number of known spermatogenic marker genes (Fig. 3.11c). *Rarg*, *Neurog3*, *Nanos3* and *Bmi1* expression is higher in MIWI2^{pos} cells while *Id4* and *Epcam* are more highly expressed in GFRα1^{pos} cells. *Cd9* is ubiquitously expressed in both populations. Analysis of the pooled expression of single GFRα1^{pos} spermatogonia revealed that 146 genes are upregulated with age and 64 downregulated (greater than 1.7-fold increase (adj. $P < 0.05$, Wald test).

3.2.8 Spermatogonial transcriptional variability declines with age

We next profiled changes in expression variability that occur with age in GFRα1^{pos} and MIWI2^{pos} undifferentiated spermatogonia. When comparing the over-dispersion parameter for genes expressed in GFRα1^{pos} and MIWI2^{pos} spermatgonia, we observed a significant reduction in cell-to-cell transcriptional variability (Mann-Whitney-Wilcoxon test; ***, $p < 10^{-76}$) in MIWI2^{pos} spermatogonia; this is consistent with a transition to a more committed state (Fig. 3.12a). Furthermore we observed that cell-to-cell transcriptional variability in GFRα1^{pos} (Fig. 3.12b) (Mann-Whitney-Wilcoxon test; ***, $p < 10^{-19}$) and MIWI2^{pos} (Mann-Whitney-Wilcoxon test; ***, $p < 10^{-30}$) decreases with age. 196 genes become more variable with age and 988 less variable (Fig. 3.12c) in GFRα1^{pos} spermatogonia. Similarly, 149 genes become significantly more variable with age and 1342 less variable in MIWI2^{pos} spermatogonia (Fig. 3.12d). Significantly variable genes, 2-fold higher overdispersion (expected false discovery rate (EFDR) < 0.1). Examples of genes that more variable and less with age in GFRα1^{pos} spermatogonia are shown in **Fig. 3.12e** and **f**.

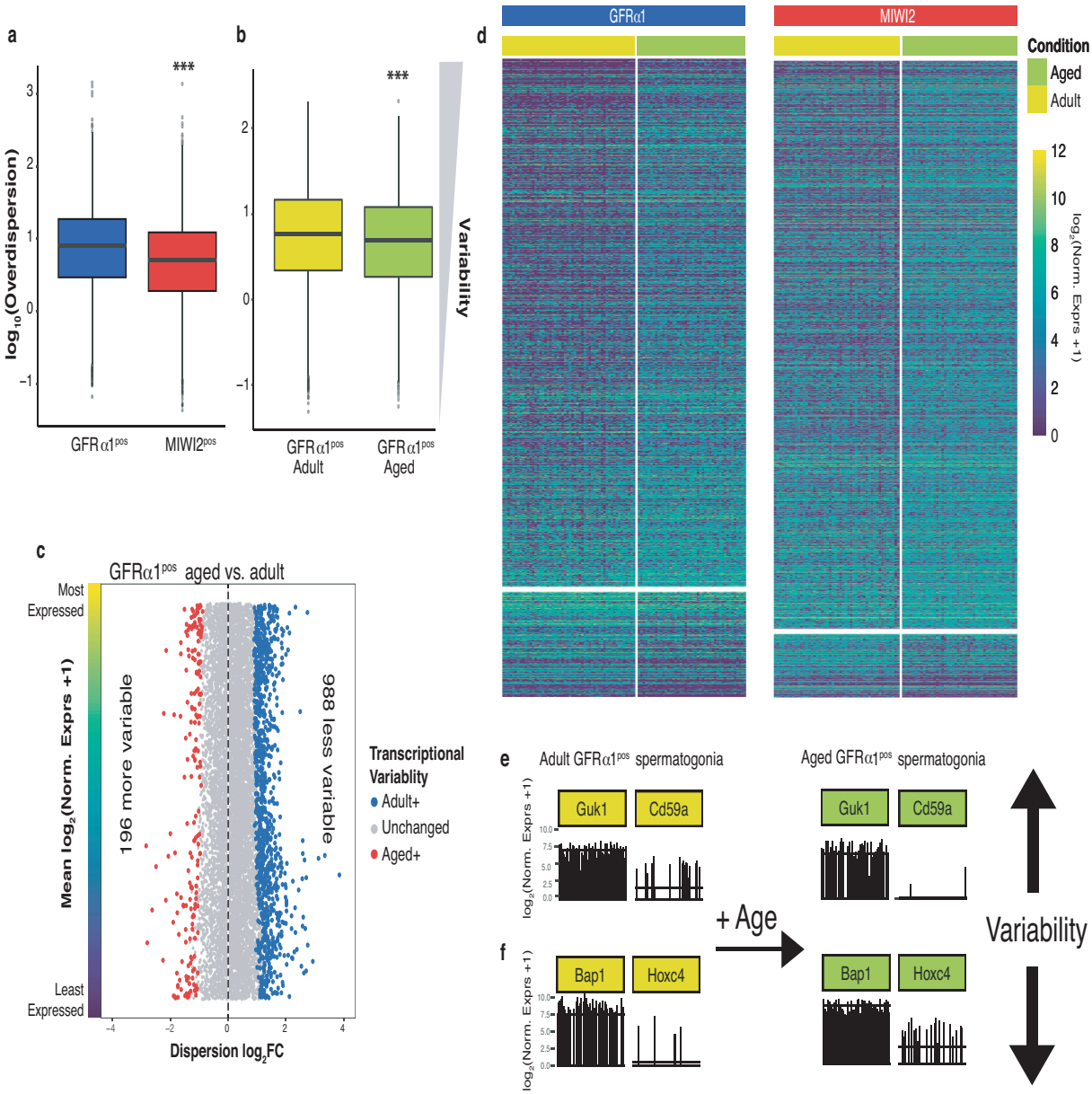


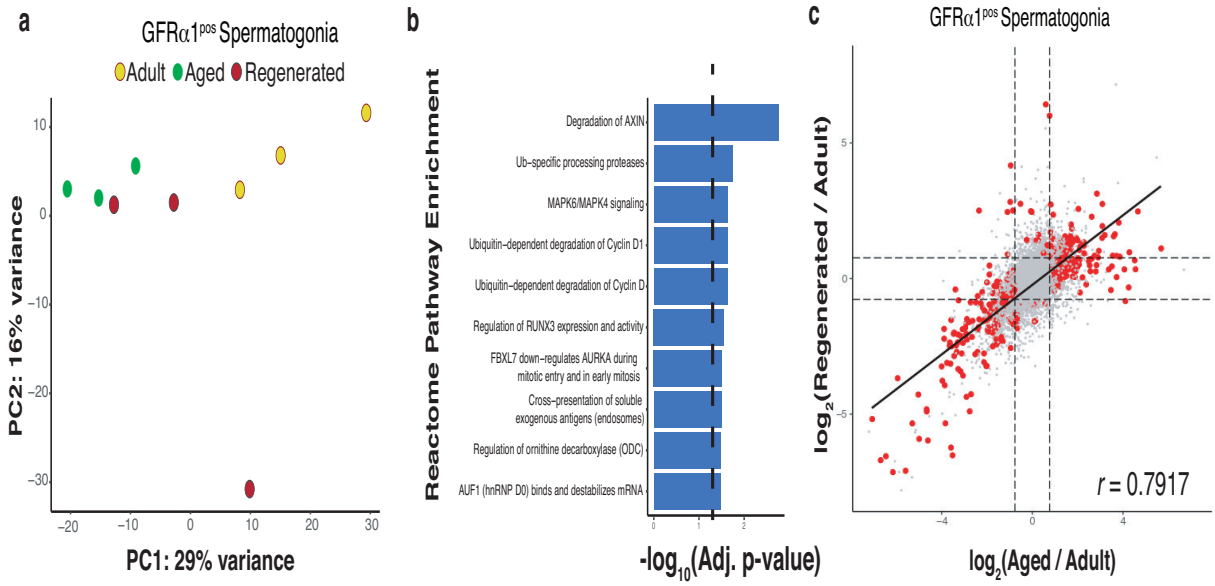
Fig. 3.12: Spermatogonial transcriptional variability declines with age.

(a) Cell-to-cell transcriptional variability is higher in GFRα1^{pos} than MIWI2^{pos} spermatogonia (Mann-Whitney-Wilcoxon test; ***, $p < 10^{-76}$). **(b)** Cell-to-cell transcriptional variability in GFRα1^{pos} spermatogonia decreases with age (Mann-Whitney-Wilcoxon test; ***, $p<10^{-19}$). **(c)** The vast majority of the genes expressed by GFRα1^{pos} spermatogonia exhibit less cell-to-cell expression variability with age. 196 genes become more variable and 988 less variable with age. Significantly variable genes, 2-fold higher overdispersion (EFDR < 0.1). Genes are sorted by their mean expression. Genes highlighted in red exhibit increased variability while those in blue reduced variability.

Fig. 3.12: Spermatogonial transcriptional variability declines with age (continued).
(d) Heatmaps showing the single-cell expression of genes that have altered cell-to-cell variability with age in GFRα1^{pos} (LHS) and MIWI2^{pos} (RHS) spermatogonia. Each column is a cell and each row a gene whose cell-to-cell expression variability, as measured by over-dispersion, changed with age. Genes in the top half become less variable with age and those in the bottom half less variable. 196 genes become more variable and 988 less variable with age in MIWI2^{pos} spermatogonia. **(e)** Example genes that exhibit decreasing cell-to-cell expression variability with age in GFRα1^{pos} spermatogonia. **(f)** Example genes that exhibit increasing cell-to-cell expression variability with age in GFRα1^{pos} spermatogonia. (GFRα1^{pos} cells, *n* adult = 54, *n* aged = 44; MIWI2^{pos} cells, *n* adult = 48, *n* aged = 44).

3.2.9 Impact of testicular tissue regeneration on spermatogonial transcriptomes

We next profiled the impact of testicular injury and subsequent regeneration on the transcriptomes of GFRα1^{pos} spermatogonia. Analysis of the bulk RNA-seq data from adult, aged and regenerated GFRα1^{pos} spermatogonia revealed that the testis condition is detectable in a PCA of their VST gene expression (Fig 3.13a). PC1 is responsible for 29% of the variance in gene expression. The GFRα1^{pos} samples are arranged along PC1 based on the testis condition. Regenerated samples appear to be transcriptionally intermediate between adult and aged samples.



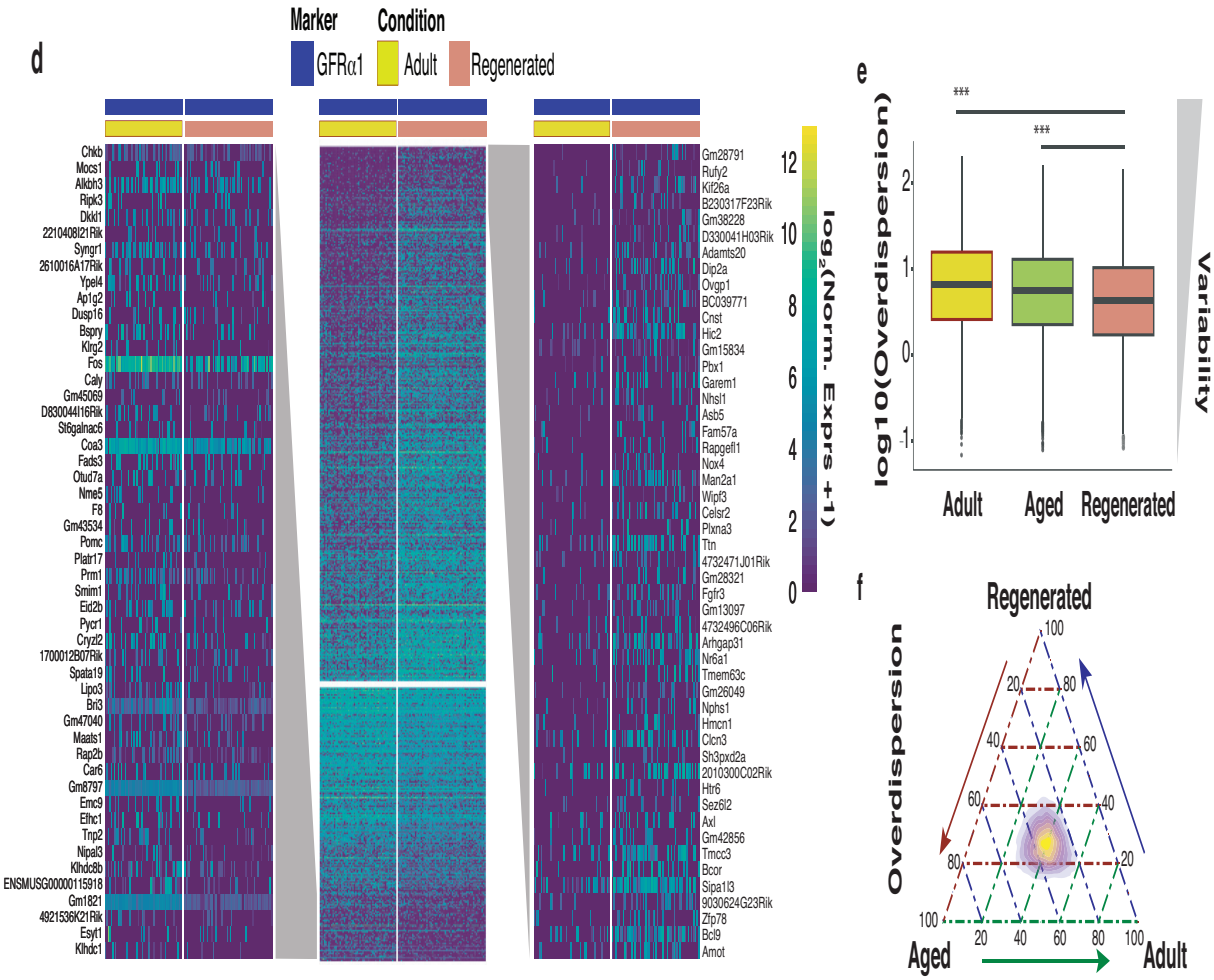


Fig. 3.13: Transcriptional variability decreases following testicular injury and regeneration. **(a)** PCA based on VST expression levels in GFRα1^{pos} bulk RNA-seq samples. Yellow = adult, green = aged and red = regenerated spermatogonia. PC1 accounts for 29% of the variance and separates out samples based on the testis type. PC2 accounts for 16% of the variance. **(b)** Reactome Pathway enrichment analysis for regenerated vs. adult GFRα1^{pos} spermatogonia. 173 genes upregulated and 45 downregulated with testis injury and regeneration. Significantly changing genes, greater than 1.7-fold (adj. $P < 0.05$, Wald test). **(c)** The impacts of ageing and testis regeneration on mean expression in GFRα1^{pos} spermatogonia are highly correlated (Pearson's $r = 0.7917$). Genes that are differentially expressed in either of the contrasts are highlighted in red. A linear regression fitted to the log2(fold-changes) of these genes is indicated with a black diagonal line. (Adult GFRα1^{pos}, $n = 3$; aged GFRα1^{pos}, $n = 3$; regenerated GFRα1^{pos}, $n = 3$). **(d)** Heat map of normalised gene expression for differentially expressed genes in regenerated vs. adult GFRα1^{pos} single spermatogonia. 282 genes downregulated and 551 upregulated with testis injury and regeneration. **(e)** Box plots of cell-to-cell transcriptional variability in GFRα1^{pos} spermatogonia. Variability decreases following testicular injury and regeneration when compared to spermatogonia from the testes of adult (Mann-Whitney-Wilcoxon test; ***, $p < 10^{-96}$) and aged mice (Mann-Whitney-Wilcoxon test; ***, $p < 10^{-35}$). **(f)** Ternary density contour plot of gene overdispersion values in GFRα1^{pos} spermatogonia. (Adult GFRα1^{pos} cells, $n = 54$; aged GFRα1^{pos} cells, $n = 44$; regenerated GFRα1^{pos} cells, $n = 62$).

Testicular injury and regeneration is responsible for small, subtle differences in mean expression, 173 genes are upregulated and 45 downregulated in GFRα1^{pos} spermatogonia following testis regeneration, greater than 1.7-fold ($P < 0.05$, Wald test). We detected enrichment of Reactome pathways associated with protein degradation, cell cycle regulation and cell signalling (adjusted $p < 0.05$) (Fig 3.13b). Although fewer genes are differentially expressed in GFRα1^{pos} spermatogonia due to testis regeneration, the fold-changes of genes that are differentially expressed as a result of ageing or testis regeneration are strongly correlated (Pearson's $r = 0.7917$) (Fig 3.13c).

Analysis of the pooled single-cell expression revealed that 551 genes are upregulated in GFRα1^{pos} spermatogonia with testis injury and regeneration while 282 genes are downregulated, normalised expression for the top 50 up- and downregulated genes is shown in **Fig 3.13d**. Given the more numerous sources of technical noise affecting in scRNA-seq experiments when compared to bulk RNA-seq (Lun and Marioni, 2017), it is surprising that the number of differentially expressed genes detected from the analysis of the scRNA-seq is higher than for the same contrast in the bulk RNA-seq data. In parallel to the changes in transcriptional variability with age, cell-to-cell transcriptional heterogenity, as quantified by gene overdispersion, in GFRα1^{pos} spermatogonia declines following testicular injury and regeneration when compared to cells from the testes of adult (Mann-Whitney-Wilcoxon test; ***, $p < 10^{-96}$) and aged mice (Mann-Whitney-Wilcoxon test; ***, $p < 10^{-35}$) (Fig 3.13e). A ternary density contour plot of the gene overdispersion values in adult, aged and regenerated GFRα1^{pos} spermatogonia shows an appreciable shift away from the centre towards the adult pole (Fig 3.13f). This decline in transcriptional diversity is all the more stark given that 2,420 genes are significantly less variable in GFRα1^{pos} spermatogonia following busulfan treatment and testis regeneration while a mere 52 genes are more variable. Significantly variable genes, 2-fold higher overdispersion (EFDR < 0.1).

3.2.10 | Increased expression of transposable elements with age

Finally we profiled the expression of *RepBase* rodent repeats in GFRα1^{pos} and MIWI2^{pos} undifferentiated spermatogonial bulk RNA-seq samples. From PCA of VST repeat expression it is possible to distinguish GFRα1^{pos} and MIWI2^{pos} samples based on their expression of specific repeats (Fig. 3.14a). Similar to the results of the gene expression analysis, PC1 separates samples based on their of the GFRα1/MIWI2 state and is responsible for 22% of the variance in repeat expression. PC2 separates samples based on the mouse cohort of origin and is responsible for 14% of the variance in repeat expression. 16 repeats, most of which are ERVs and short interspersed elements (SINEs), are upregulated and 1 downregulated between aged MIWI2^{pos} and GFRα1^{pos} undifferentiated spermatogonia, greater than 1.7-fold

increase ($P < 0.05$, Wald test) (Fig. 3.14b). While in contrast, 1 repeat is upregulated and 2 downregulated in adult spermatogonia and 2 repeats upregulated and 3 downregulated when the same contrast is made for the regenerated samples. Additionally, 8 repeats, most of which are ERVs and SINEs, are upregulated between adult and aged MIWI2^{pos} spermatogonia (Fig. 3.14c) and between regenerated and aged MIWI2^{pos} spermatogonia (Fig. 3.14d). No repeats are differentially expressed between adult and regenerated MIWI2^{pos} spermatogonia. Similarly, no repeats are differentially expressed between any of the GFRα1^{pos} testis conditions.

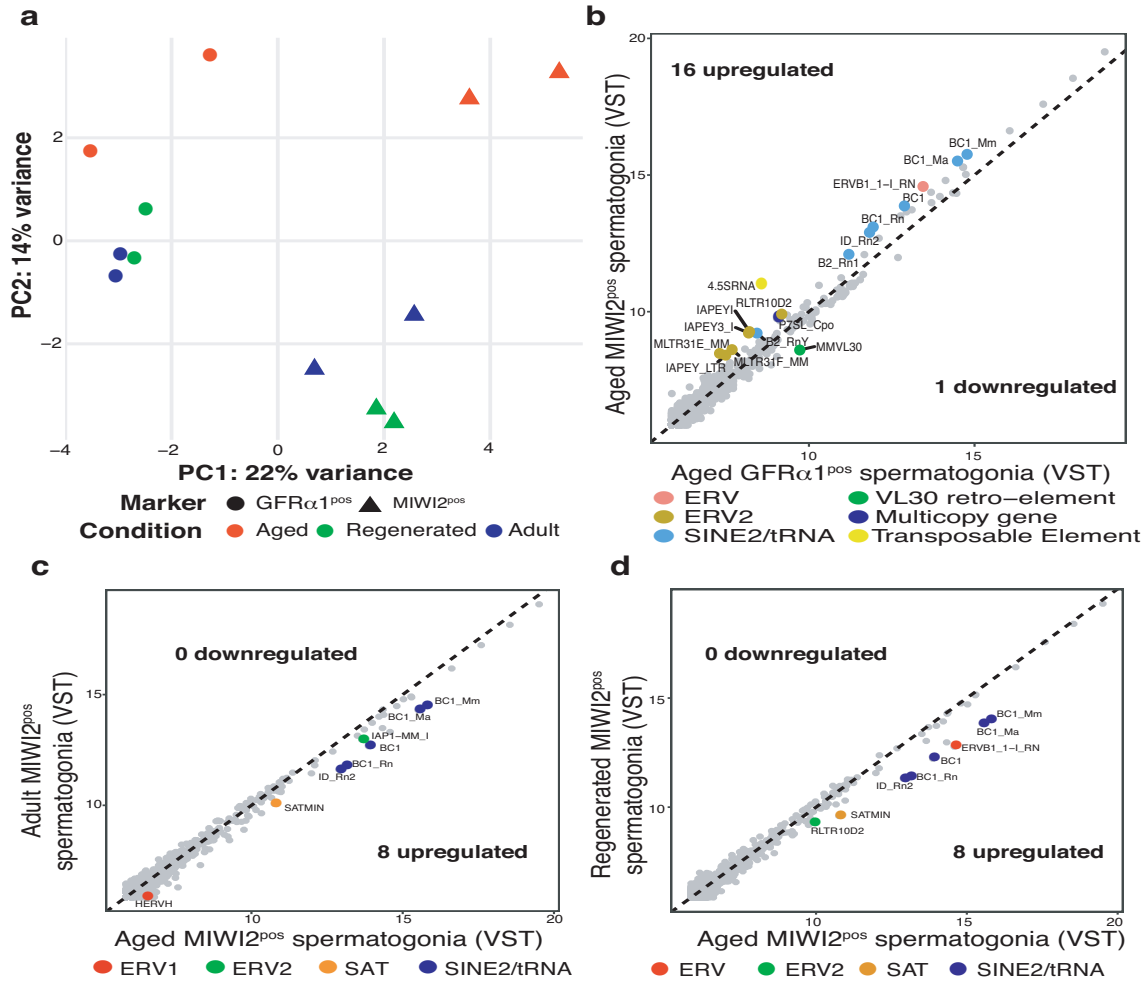


Fig. 3.14: Expression of transposable elements in undifferentiated spermatogonia.
(a) PCA of GFRα1^{pos} and MIWI2^{pos} undifferentiated spermatogonia based on the VST repeat expression. Blue = adult, red = aged and green = regenerated samples. GFRα1^{pos} = circles and MIWI2^{pos} = triangles. **(b)** Expression scatterplot showing relative average expression (VST counts) of TEs between aged GFRα1^{pos} and MIWI2^{pos} undifferentiated spermatogonia. Significantly deregulated repeats ($P < 0.05$, Wald test) with a greater than 1.7-fold change are highlighted.

Fig. 3.14: Expression of transposable elements in undifferentiated spermatogonia (continued). **(c)** Expression scatterplot showing relative average expression (VST counts) of TEs between aged and adult MIWI2^{pos} undifferentiated spermatogonia. **(d)** Expression scatterplot showing relative average expression (VST counts) of TEs between aged and regenerated MIWI2^{pos} undifferentiated spermatogonia. (Adult n = 3; aged n = 2; regenerated n = 2).

3.3 Discussion

Senescence is the time-dependent, progressive decline in biological function and is usually attributed to the accumulation of molecular changes with time (Gems and Partridge, 2013; López-Otín et al., 2013). Both genetic and epigenetic factors have been implicated in this senescence-associated dysregulation. A relatively small proportion of an organism's transcriptome changes with increasing age; the genes whose expression is altered with age are largely tissue- and cell-type specific (Stegeman and Weake, 2017). Most studies investigating transcriptional changes have only considered alterations to mean expression with age. Numerous studies have investigated the transcriptional changes associated with ageing, especially in haematopoietic cells, (Chen et al., 2013; de Magalhães et al., 2009; Kowalczyk et al., 2015) but thus far no one has profiled the transcriptome-wide impact of ageing on mammalian spermatogenesis. By combining bulk RNA-seq, scRNA-seq and ATAC-seq data we were able to assay the changes to mean gene expression, cell-to-cell transcriptional variability and chromatin accessibility that occur with ageing and testicular regeneration in FACS-sorted, mouse undifferentiated spermatogonia.

Using FACS-sorted, purified populations of GFRα1^{pos} and MIWI2^{pos} undifferentiated spermatogonia we were able to confirm that the MIWI2^{pos} sub-population characterised by Carrieri et al. (2017) is transcriptionally distinct from the more stem-like GFRα1^{pos} sub-population. These populations can be distinguished from each other based on differences in their mean expression of genes (Fig. 3.10) and TEs (Fig. 3.14). Furthermore, these populations are distinguishable based on differences in their cell-to-cell transcriptional heterogeneity. Overall gene expression is more variable in GFRα1^{pos} than MIWI2^{pos} undifferentiated spermatogonia (Fig. 3.12a). A far larger number of individual genes exhibit greater cell-to-cell transcriptional heterogeneity in GFRα1^{pos} undifferentiated spermatogonia. As discussed in 1.3, biological noise is an inherent feature of biological processes involving molecular interactions (Ecker et al., 2017). When components in a system are present in large quantities, random fluctuations have little impact on the overall system (Swain et al., 2002). However in a small system like a biological cell, molecules are generally found in relatively low

abundances. It has been suggested that transcriptional variability in pluri- and multipotent pools of cells facilitates the dynamic decision between self-renewal and differentiation based on external stimuli secreted from the niche (Dueck et al., 2016; Kumar et al., 2014; Macarthur and Lemischka, 2013). This confers developmental robustness (Torres-Padilla and Chambers, 2014). Therefore it is unsurprising that the more stem-like GFRα1^{pos} undifferentiated spermatogonia exhibit more cell-to-cell transcriptional variability than the transit-amplifying MIWI2^{pos} sub-population. Moreover, it has been suggested in a recent *bioRχiv* preprint that transcriptional diversity is a hallmark of developmental potential (Gulati et al., 2019).

FACS analysis confirmed that ageing affects the spermatogonial compartment, it impacts upon the proportion and number of GFRα1^{pos} and MIWI2^{pos} undifferentiated spermatogonia in the mouse testes. Although an age-effect is detectable from changes in mean expression for the bulk RNA-seq samples, only a small number of genes are differentially expressed between young and aged GFRα1^{pos} or MIWI2^{pos} samples. Members of the homeobox (HOX) B cluster (*Hoxb5*, *Hoxb6* and *Hoxb7*) are amongst the genes down-regulated with age in GFRα1^{pos} spermatogonia (Fig. 3.9d). Wagner et al. (2009) have shown that while the human orthologue of *Hoxb7* is down-regulated with age in mesenchymal cells, it and the *Hoxb5* and *Hoxb6* orthologues are induced by ageing in HSPCs.

Amongst the genes that are up-regulated in GFRα1^{pos} undifferentiated spermatogonia with ageing, the proto-oncogenes *Stat3*, *Fos* and *Ret* are associated with SSC self-renewal and proliferation (He et al., 2008; Oatley and Brinster, 2008; Oatley et al., 2010). Interestingly, RET and GFRα1 are the co-receptors for GDNF secreted from the spermatogonial niche (Sharma and Braun, 2018). It has been previously shown that the expression of *Fos* in SSCs is up-regulated by GDNF signalling (He et al., 2008). This suggests that the up-regulation of *Ret* in aged GFRα1^{pos} spermatogonia leads to a concomitant change in *Fos*. As a result, SSC ageing may alter the balance between self-renewal and proliferation. A shift towards greater proliferation with increased age has been documented for mouse haematopoietic stem cells (HSCs) (Kirschner et al., 2017). Excessive proliferation of progenitors is responsible for the increased occurrence of stem-cell exhaustion with age (López-Otín et al., 2013). Even with the small effect size, enrichment of Reactome pathways and GO molecular function categories associated with ribosome assembly and translation are detectable amongst the genes that are differentially expressed between aged and young GFRα1^{pos} spermatogonia. Aberrant proteostasis is yet another Hallmark of Ageing (López-Otín et al., 2013). Although different pathways are affected by testicular injury and regeneration, there was a strong correlation between the genes that are a differentially expressed as a result of ageing or testicular injury and regeneration.

There was considerable overlap between the ATAC-seq peak TSSs across the conditions

(Figs. 3.7). Similar TFs motifs and pathways are enriched. Few ATAC-seq promoter peaks change with age (Fig. 3.9f) which supports previous observations from human $CD8^+$ T cells (Moskowitz et al., 2017; Ucar et al., 2017). Intriguingly, *Hoxb6* is one of the few genes with an age-specific ATAC-seq peak that is also differentially expressed. However, it might be worth broadening our search outside the promoter regions and giving more consideration to genic peaks downstream of TSSs as well as intergenic peaks (e.g. at enhancer CREs) if we are to better understand the epigenetic differences between adult and aged and adult and regenerated spermatogonial samples. It would be possible to profile mouse enhancers by integrating comparative genomics data from humans with the results of our ATAC-seq experiments and combining these data with information about known spermatogenic DNA TF binding motifs.

Notably, differentially expressed TEs were only detected for the contrasts involving aged $MIWI2^{pos}$ spermatogonia. Of the TEs that are upregulated in aged $MIWI2^{pos}$ spermatogonia as compared to $GFR\alpha1^{pos}$ cells, expression of the intracisternal-A particle (IAP) family of ERVs has been shown to be repressed by MIWI2 in undifferentiated spermatogonia (Vasiliauskaité et al., 2018). IAPs drive the expression of neighbouring genes in the absence of MIWI2-mediated DNA methylation. Previous studies in mice and *Drosophila* have shown that ageing leads to the increased expression and mobilisation of TEs due to epigenetic and chromatin changes (De Cecco et al., 2013; Wood and Helfand, 2013; Wood et al., 2016).

Finally, testis ageing and regeneration both lead to decreased overall cell-to-cell transcriptional heterogeneity in $GFR\alpha1^{pos}$ and $MIWI2^{pos}$ spermatogonia (Fig. 3.12b & c). Many more individual genes become less variable with age than more variable. Notably, amongst the top genes that become less variable with age, the HOX gene *Hoxc4* and *Bap1*, encoding a Polycomb de-ubiquitinating enzyme, have been implicated in the regulation of self-renewal (Dey et al., 2012; Schmidt et al., 2009). *Hoxc4* expression in SSCs is directly regulated by GDNF-signalling (Schmidt et al., 2009). *Cd59a* encoding an inhibitor of the complement membrane attack complex (MAC) (Harris et al., 2003) becomes more variable with age. This decreasing transcriptional noise with age is surprising given the increased transcriptional heterogeneity with age observed for mouse $CD4^+$ T cells, human pancreas and mouse lung (Angelidis et al., 2019; Enge et al., 2017; Martinez-Jimenez et al., 2017). While the increased transcriptional variability with age in these tissues likely reflects a functional decline in the transcriptional noise buffering pathways, the decreasing variability with age observed in this study may instead be the result of the "selfish spermatogonial selection" that leads to clonal expansion of certain spermatogonial sub-lineages with increased paternal age. The large decrease in transcriptional heterogeneity seen for the spermatogonia from busulfan-treated, regenerated testes lends some credence to this theory. It is not difficult to envisage that some

spermatogonia will be more resistant to busulfan. Regeneration would lead to the clonal expansion of the surviving SSCs and/or transit-amplifying undifferentiated spermatogonia. It stands to reason that this could represent a genetic bottleneck that reduces genomic diversity and potentially transcriptomic variability in the surviving lineages. Regardless of the origin of the decline in variability, what is true for somatic tissues is not necessarily true for gametes. Altered, rather than increased, cell-to-cell transcriptional variability may be a hallmark of ageing in mammalian tissues.

4

YTHDF2-mediated buffering of transcriptional noise in murine zygotes

4.1 Introduction

Stochasticity is an inherent feature of biological processes involving molecular interactions (Ecker et al., 2017). When components in a system are present in large quantities, random fluctuations have little impact on the overall system (Swain et al., 2002). However in small systems like a biological cell, components i.e. molecules are generally found at relatively low abundances. Even amongst isogenic populations of cells a small, transient fluctuation in the quantity of an RNA molecule has the potential to introduce non-genetic phenotypic variation (Dong et al., 2011; Elowitz et al., 2002). This *transcriptional noise*, as discussed in **Chapter 1** (1.3), manifests itself as cell-to-cell variability in transcript abundances and reflects the contributions of intrinsic and extrinsic factors (Swain et al., 2002). The former includes gene-specific factors like the rates of transcription and RNA degradation (Elowitz et al., 2002) while the latter encompasses cell-specific features like the cell size and the availability of upstream regulator molecules including RNA polymerases and TFs (Raser and O'Shea, 2005; Sherman et al., 2015; Valadares Barroso et al., 2018). Intrinsic noise is a more significant source of transcriptional variability than extrinsic noise in mammalian cells (Levesque and Raj, 2013; Raj et al., 2006). Transcriptional noise is a necessary and regulated feature of cell populations (Antolović et al., 2017; Dueck et al., 2016). It provides phenotypic plasticity and allows a population to react more dynamically to external stimuli (Ecker et al., 2017). However, excessive heterogeneity in transcript abundance may compromise homoeostasis. To overcome deleterious effects of heterogeneity, transcriptional noise must either be regulated or integrated into the cell system (Dueck et al., 2016).

The steady-state abundance of a transcript in a cell is lower than the amount predicted purely from its rate of transcription and also reflects its rate of turnover (Maekawa et al., 2015). Though transcription is the main source of intrinsic cell-to-cell transcriptional heterogeneity, degradation also has a bearing on it. The stability of eukaryotic mRNAs, as measured by their half-lives, is influenced by many general and transcript-specific decay factors as discussed in **Chapter 1** (1.2.1). However, amongst the most important determinants of cytoplasmic mRNA stability is $3'$ polyadenylation (Chang et al., 2014; Lim et al., 2014; Norbury, 2013). Poly(A)-tails facilitate nuclear export (Fuke and Ohno, 2008) but the main purpose of mRNA poly(A)-tails is to stabilise the rate of translation initiation (Wakiyama et al., 2000) by protecting transcripts from premature degradation (Lim et al., 2014). Due to the involvement of poly(A)-tails in protecting transcripts from decay, most mRNA degradation pathways involve a de-adenylation step (Beilharz et al., 2009; Mukherjee et al., 2002; Yamashita et al., 2005). Notably it has been shown that the miRMD pathway impacts upon cell-to-cell heterogeneity (Gambardella et al., 2017; Schmiedel et al., 2015). miRNAs buffer expression

variability during the noisy transcriptional changes that facilitate the differentiation processes involved in metazoan development (Posadas and Carthew, 2014). Schmiedel et al. (2015) have shown that miRNAs reduce expression noise for lowly expressed genes; this effect is enhanced when multiple miRNAs target the same transcript.

As discussed in Chapter 1 (1.2.2), m^6A is the most abundant non-terminal modification to vertebrate mRNA molecules (Desrosiers et al., 1974). m^6A is found along the length of mRNAs but it is particularly abundant towards their $3'$-ends (Dominissini et al., 2012; Meyer et al., 2012). Most m^6A is deposited co-transcriptionally in the nucleus by a complex containing the RNA methyltransferases METTL3 and METTL14 (Ke et al., 2017; Wang et al., 2016a,b). m^6A is a erased by the ALKBH5 RNA demethylase (Zheng et al., 2013). mRNA m^6A is usually found in the context of a DRACH motif and although at least 25% of human transcripts are modified, the vast majority of motif sites are un-methylated (Dominissini et al., 2012; Meyer et al., 2012). Single-stranded m^6A is specifically recognised by members of the YTH superfamily via the YTH domain (Xu et al., 2014; Zhu et al., 2014). As a result of interactions mediated by YTH proteins, m^6A is involved at many stages of the mRNA life-cycle. These include splicing (Liu et al., 2015; Louloupi et al., 2018; Xiao et al., 2016), APA (Kasowitz et al., 2018), sub-cellular localisation (Roundtree et al., 2017; Wang et al., 2013), translation (Coots et al., 2017; Li et al., 2017; Wang et al., 2015) and RNA stability (Du et al., 2016; Geula et al., 2015; Wang et al., 2013, 2014).

In common with the miRMD pathway, m^6AMDm^6A-mediated decay occurs in P-bodies where the CCR4-NOT deadenylase complex is recruited by YTHDF2 (Wang et al., 2013). CCR4-NOT recruitment is likely synergistic; the greater the number of m^6A sites per transcript the shorter its half-life (Ke et al., 2017). m^6A-modified mRNAs with rapid turnover rates are associated with regulatory functions while unmodified, more stable transcripts are involved in cellular housekeeping (Ke et al., 2017). It has been shown that m^6AMD regulates the transcriptional shifts necessary for the vertebrate MZT (Ivanova et al., 2017; Zhao et al., 2017), the murine EHT (Lv et al., 2018) and facilitates the transition from a pluripotent to differentiating state in mESCs (Geula et al., 2015). Given the parallels between the transcriptional programmes underpinning vertebrate embryonic development and tumourigenesis (Aiello and Stanger, 2016; Youssef et al., 2012), it is unsurprising that the expression of METTL3, ALKBH5 and YTHDF2 are clinically relevant in certain cancers (Chen et al., 2017, 2018; Paris et al., 2019; Yang et al., 2017).

Zhao et al. (2017) observed a striking overlap between the targets of zygotically-expressed miR-430 and maternally-supplied Ythdf2 in the zebrafish MZT, . Taken together with the observation that some miRNAs enhance METTL3 binding (Chen et al., 2015), this suggests

that there is interplay between the miRMD and m^{6}AMD pathways. It is possible that m^{6}AMD also modulates transcriptional noise.

4.1.1 Overview

Gene knockout (KO) approaches combined with single-cell RNA-sequencing techniques present us with an opportunity to explore the relationship, if any, between m^{6}AMD and transcriptional noise control. In the absence of m^{6}AMD, mRNA decay will still occur via other RNA decay pathways but not to the same extent. Turnover will be slower and this will increase transcriptional noise. More rapid rates of decay are associated with reduced noise (Swain, 2004). The abolition of m^{6}AMD could potentially lead to *unstructured* transcriptional heterogeneity. In contrast with the developmental delay observed by Zhao et al. (2017) for zebrafish embryos from a *ythdf2*$^{-/-}$ maternal background, Ivanova et al. (2017) determined that the development of mouse embryos beyond the 2-cell stage is blocked with an oocyte-specific YTHDF2 conditional knock out. To this end, I have analysed in collaboration with the lab of Dónal O'Carroll (MRC Centre for Regenerative Medicine, Edinburgh) the transcriptional changes that occur in mouse zygotes and 2-cell stage embryos in the absence of maternally-supplied YTHDF2 using the Smart-seq2 scRNA-seq protocol (Picelli et al., 2014). Ivalya Ivanova (O'Carroll Lab) performed all experiments and I analysed all generated data unless stated otherwise.

As will be discussed in due course, the limitations of mouse zygotes and embryo systems meant that it was also necessary to profile the impact of YTHDF2 transcriptional regulation in another context. The labs of Kamil Kranc and Dónal O'Carroll derived *Ythdf2*$^{-/-}$ and *Ythdf2*$^{+/+}$ pre-LSCs from mouse foetal liver HSPCs in order to investigate the role of YTHDF2 in acute myeloid leukaemia (AML) initiation and progression Paris et al. (2019). In collaboration with Nils Eling and Christina Ernst from the lab of John Marioni (EMBL-EBI / CRUK Institute, Cambridge), we performed droplet-based scRNA-seq using the 10x Genomics™ technology (Zheng et al., 2017) for three replicates of the *Ythdf2*$^{-/-}$ and *Ythdf2*$^{+/+}$ pre-LSCs obtained from the Kranc Lab. Furthermore, we generated bulk RNA-seq libraries from these samples in order to obtain marker genes to identify the cells in each sequenced single-cell library. Nils Eling, Christina Ernst and I designed this part of the study. I analysed all generated data unless stated otherwise.

4.2 Results

4.2.1 Zygote and 2-cell embryo scRNA-seq using the plate-based Smart-seq2 protocol

In order to profile the transcriptional changes that occur in mouse MZT with the depletion of maternal YTHDF2, Ivalya Ivanova (O'Carroll Lab) prepared single-end scRNA-seq cDNA libraries for zygotes and 2-cell stage embryos from *Ythdf2*^{mCKO} and *Ythdf2*$^{+/+}$ *Zp3Cre Tg*$^{+}$ (*Ythdf2*^{CTRL}) backgrounds. All mice used in this study were on a mixed or C57BL/6 genetic background and were bred and maintained at the MRC Centre for Regenerative Medicine, Edinburgh. The *Ythdf2*$^{HA\text{-}Fl}$ allele used to generate the maternal oocyte-specific conditional knockout *Ythdf2*^{mCKO} and control *Ythdf2*^{CTRL} was described previously by Ivanova et al. (2017). The absence of sequence downstream from exon 1 of *Ythdf2* means that no functional YTHDF2 protein is expressed in the *Ythdf2*^{mCKO} zygotes or embryos from the maternal *Ythdf2* locus or maternally-supplied *Ythdf2* transcripts. All dames were crossed with wild-type C57BL/6 sires. Three 96-well plates, two from zygotes and one from 2-cell embryos were prepared using the Smart-seq2 protocol (Picelli et al., 2014). Single-cells were distributed across the plates such that maternal YTHDF2 expression and the plate of origin were not confounded. Ambion® External RNA Control Consortium (ERCC) Spike-In Control mix (1:500000 dilution) was added to the cell lysates as an internal control. Non-stranded cDNA libraries were prepared from cell lysates with the Nextera™ XT Kit (Illumina) and the 75 bp single-end cDNA libraries sequenced on an Illumina™ NextSeq 500 (Genome Core Facility, EMBL Heidelberg).

Raw single-end FASTQs from the 177 successfully sequenced cells were mapped against the *Mus musculus genome* (mm10) and GENCODE transcript annotation version M14 (Mudge and Harrow, 2015) using STAR (v2.5.3) (Dobin et al., 2013) with its default parameters. Non strand-specific gene counts were quantified from the read alignments using *htseq* (Anders et al., 2015) by setting the htseq-count stranded parameter to "no".

QC and cell- and gene-filtering were performed with the R/Bioconductor packages *scater* (McCarthy et al., 2017) and *scran* (Lun et al., 2016). The raw numbers of genes sequenced per cell are shown in **Fig. 4.1a**. The log(total counts) per cell are displayed in **Fig. 4.1b**. The proportions of ERCC spike-ins and mitochondrial genes among sequenced reads are shown in **Fig. 4.1c** and **d**. The proportion of reads originating form the mouse mitochondrial genome is dramatically higher in the cells obtained from the 2-cell embryos than zygotes. This was regardless of the maternal genetic background.

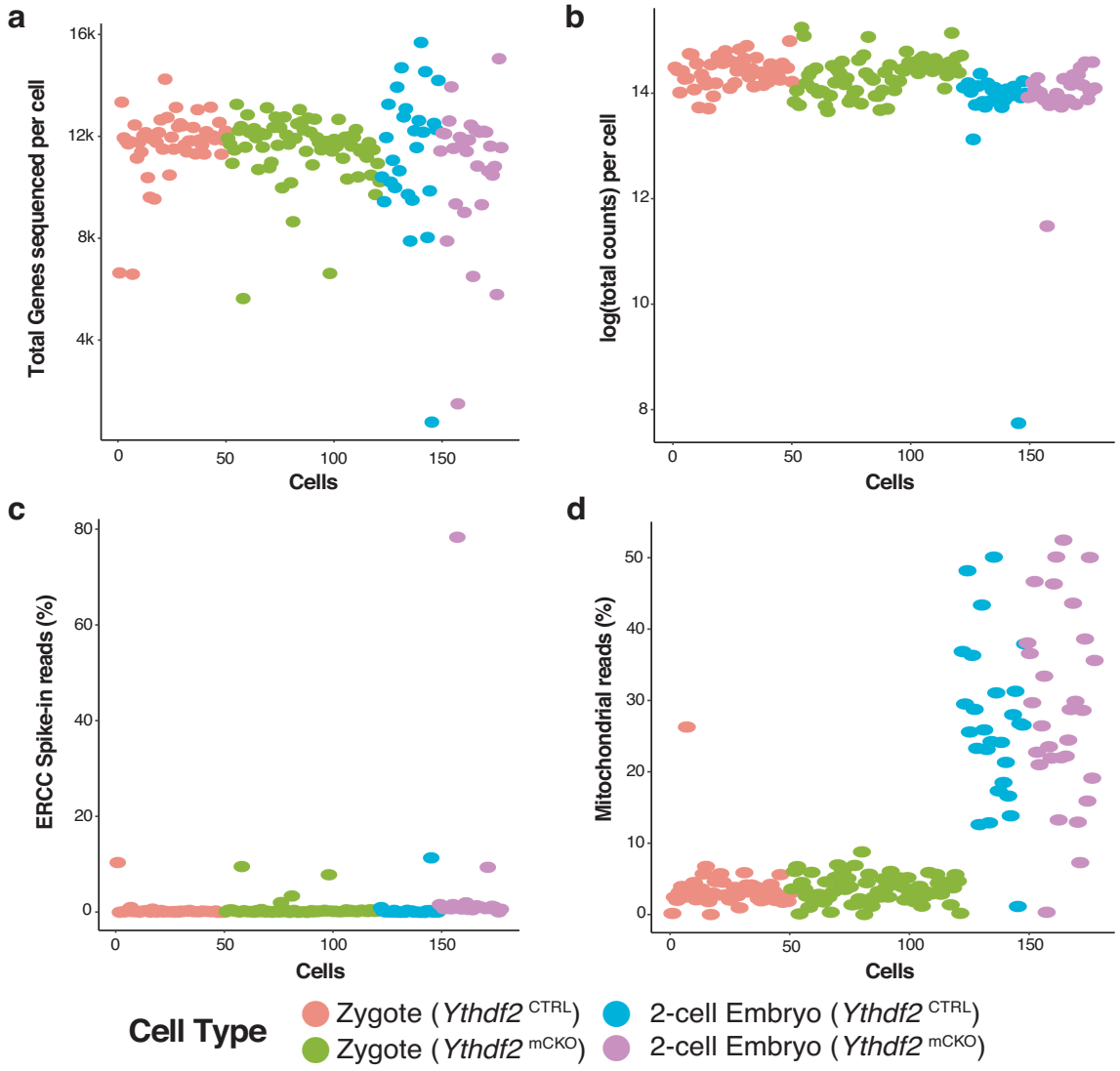


Fig. 4.1: Quality Control of single-cell transcriptional profiles from zygotic and embryonic cells.

(a) Total genes expressed per cell. **(b)** Total read counts per cell. **(c)** Proportions of reads per cell that are derived from ERCC spike-in molecules. **(d)** Proportions of sequenced reads per cell that originate from the mitochondrial genome.

Cells with small library sizes, those whose $\log_{10}$(library size) was 3 MADs lower than the median value, and low library complexity, those whose $\log_{10}$(total features by counts) was 3 MADs lower than the median value, were filtered out. Cells with proportionately high ERCC and mitochondrial expression, those whose expression values were 3 MADs higher than the median value, were also filtered out. The numbers of cells that passed each stage of QC are contained in **Table 4.1**.

Table 4.1: Cell quality control and filtering for mouse zygotes and 2-cell embryos.

Condition	***Ythdf2*$^{\text{CTRL}}$**	***Ythdf2*$^{\text{mCKO}}$**	**Total**
Stage	**Zygote**	**Zygote**	
Raw	50 cells	71 cells	121 cells
By Lib. size	50 cells	71 cells	121 cells
By No. Genes Exprsd.	46 cells	68 cells	114 cells
By Spike Exprs.	46 cells	67 cells	113 cells
By Mito Exprs.	46 cells	67 cells	113 cells
Final	46 cells	67 cells	113 cells
Stage	**2-cell Embryo**	**2-cell Embryo**	**Total**
Raw	27 cells	29 cells	56 cells
By Lib. size	25 cells	28 cells	53 cells
By No. Genes Exprsd.	21 cells	22 cells	43 cells
By Spike Exprs.	20 cells	10 cells	30 cells
By Mito Exprs.	3 cells	2 cells	5 cells
Final	3 cells	2 cells	5 cells

Most cells from the 2-cell embryos, both *Ythdf2*$^{\text{CTRL}}$ and *Ythdf2*$^{\text{mCKO}}$, were filtered out owing to their high mitochondrial gene expression (see Table 4.1). This aberrant mitochondrial gene expression is apparent from **Fig. 4.1d**. An over-representation of reads originating from the mitochondria is indicative of apoptotic or broken cells (Ilicic et al., 2016). Cells with perforated outer membranes will easily lose cytoplasmic RNAs but those enclosed in mitochondria are more likely to be retained.

The single-cell expression was normalised using *scran*'s *computeSumFactors* function that uses a deconvolution method that borrows information from neighbouring (i.e. transcriptionally similar) cells in order to calculate sum factors for count normalisation. All subsequent analyses looking at single-cell RNA expression use this $\log_2$-transformed, normalised counts (including a pseudocount) i.e. $\log_2$(normalised expression + 1). PCA for cells that passed the QC filtering is visualised with a scatter plot in **Fig. 4.2e**.

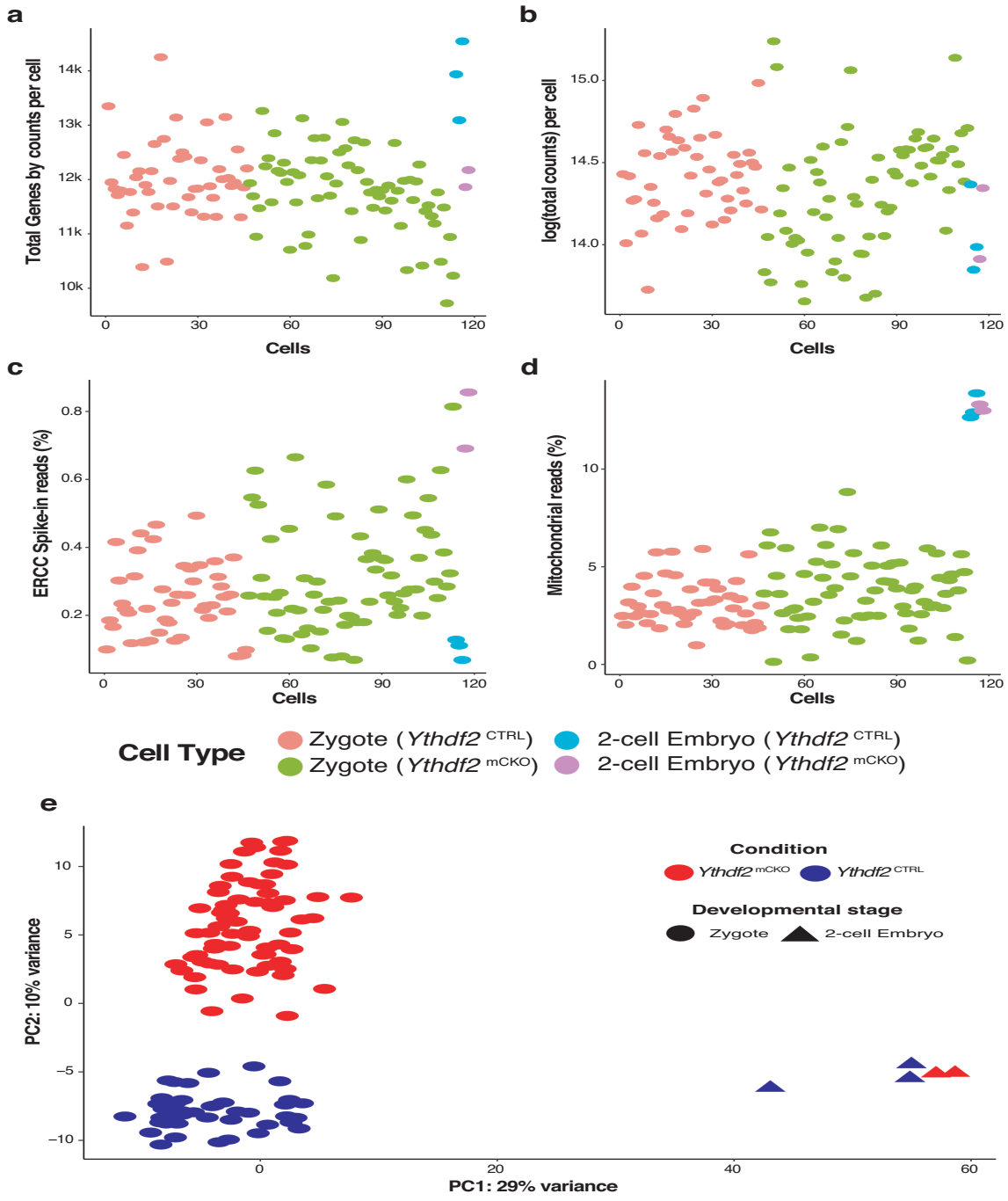


Fig. 4.2: Filtered cells remaining after quality control.

(a) Total genes expressed per cell in the filtered cells. **(b)** Total read counts per cell in the filtered cells. **(c)** Proportions of reads per cell that are derived from ERCC spike-in molecules in the filtered cells. **(d)** Proportions of sequenced reads per cell that originate from the mitochondrial genome in the filtered cells. **(e)** PCA dimensionality reduction for mouse zygotes and 2-cell stage embryos after QC and filtering. Single zygotes are visualised as circles and two-cell embryos as triangles. Cells are coloured based on their maternal *Ythdf2* background; blue = *Ythdf2*$^{\text{CTRL}}$ and red = *Ythdf2*$^{\text{mCKO}}$. Filtered zygote single-cells, *n Ythdf2*$^{\text{CTRL}}$ = 46, *n Ythdf2*$^{\text{mCKO}}$ = 67; filtered two-cell embryos single-cells, *n Ythdf2*$^{\text{CTRL}}$ = 3, *n Ythdf2*$^{\text{mCKO}}$ = 2.

Of the 118 cells that passed the QC, only 5 were from 2-cell stage embryos. As a result, all subsequent analyses were restricted to the zygotic cells. The remaining 113 zygotes are visualised with the t-SNE dimensionality reduction technique in **Fig. 4.3**. The t-SNE separates zygotes based on their YTHDF2 maternal background (Fig. 4.3a). Furthermore, *Ythdf2* expression is higher in *Ythdf2*$^{\text{CTRL}}$ zygotes than *Ythdf2*$^{\text{mCKO}}$ cells (Fig. 4.3b). Lowly and un-expressed genes, those with a mean normalised expression < 1 across all cells in batch and those expressed in a single plate and/or in < 10 % cells were removed after cell filtering.

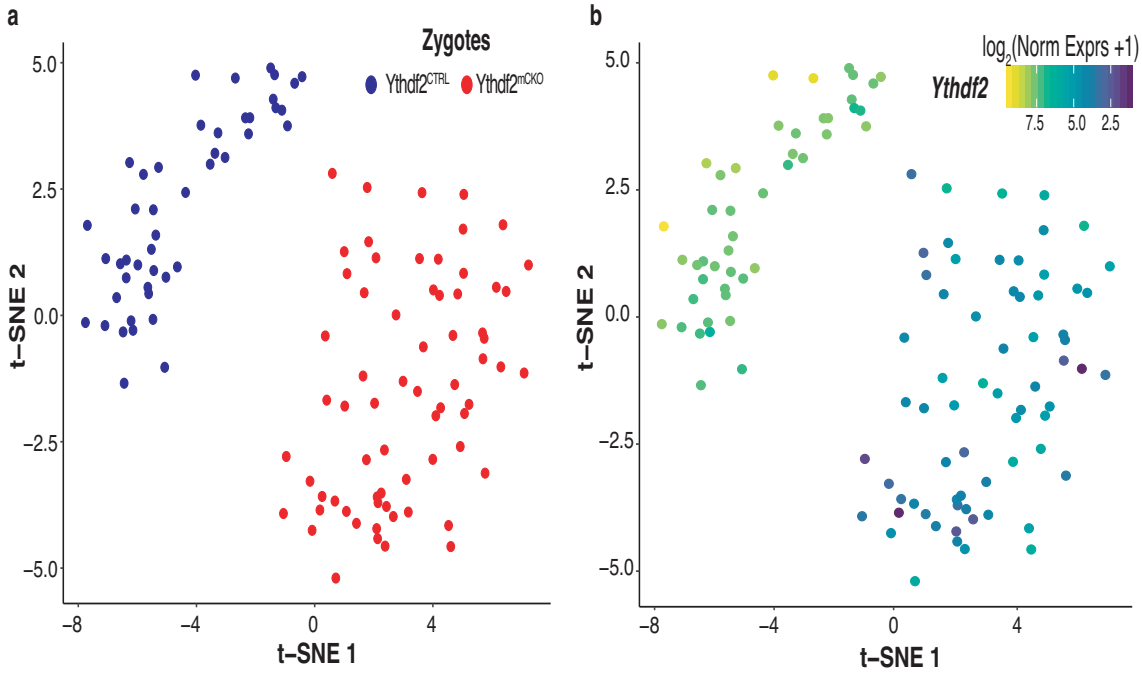


Fig. 4.3: Visualisation of filtered *Ythdf2*$^{\text{mCKO}}$ and *Ythdf2*$^{\text{CTRL}}$ zygotes.
(a) t-SNE dimensionality reduction for the final set of filtered mouse zygotes. Zygotes are coloured based on their *Ythdf2* background; blue = *Ythdf2*$^{\text{CTRL}}$ and red = *Ythdf2*$^{\text{mCKO}}$. **(b)** t-SNE of zygotes coloured based on their *Ythdf2* expression ($\log_2$(normalised expression + 1)). Final zygote numbers after QC and cell filtering. 113 cells remained, of which 67 are from a *Ythdf2*$^{\text{mCKO}}$ background and 46 from a *Ythdf2*$^{\text{CTRL}}$ background.

The final dataset of 113 zygotic cells and 12,689 genes was used as input for single-cell differential gene expression and variability testing analyses. Differences in mean gene expression and transcriptional variability between the *Ythdf2*$^{\text{mCKO}}$ and *Ythdf2*$^{\text{CTRL}}$ zygotes were tested using the R/Bioconductor package BASiCS (Eling et al., 2018; Vallejos et al., 2015, 2016). The use of spike-in controls allows BASiCS to better estimate plate-specific

sources of technical noise. Transcriptional variability was initially tested using the normal BASiCS model which due the dependence of the variance on the mean only assesses changes in gene over-dispersion (i.e. transcriptional variability) for the genes that are not differentially expressed. The BASiCS regression model was subsequently developed by Eling et al. (2018) and overcomes this dependence by inferring a regression trend between the gene over-dispersion and mean expression. This allows changes in mean expression and transcriptional variability (estimated by the residual over-dispersion) to be tested simultaneously. Only genes that are expressed in at least 2 cells (in both groups) are tested by the BASiCS regression model. The BASiCS Markov chain Monte Carlo (MCMC) was run for 40,000 iterations with 20,000 burn-in iterations and a thinning value of 20.
The R/Bioconductor packages *goseq* (Young et al., 2010) and *fgsea* (Sergushichev, 2016) were used for functional enrichment analyses. m^6A peak data from methylated RNA immunoprecipitation sequencing (meRIP-seq) and miCLIP datasets in the *RMBase* (Xuan et al., 2018) were used to determine which genes were likely to be under YTHDF2 regulation. This was necessary due to the failure of an meRIP-seq experiment in germinal vesicle (GV) oocytes (data not shown). Only m^6A peaks identified in five or more independent experiments were considered. *Deeptools* (Ramírez et al., 2014) was used to plot the number of m^6A peaks in 10nt intervals across the gene bodies. *RNAfold* (Zuker and Stiegler, 1981) was used to calculate the minimum free energy (MFE) from $3'$-UTR sequences. To account for the positive relationship between sequence length and 'structuredness', normalised MFE density values were calculated from the raw MFE values as described by Trotta (2014). *GraphClust* (Heyne et al., 2012) was used to detect clusters of RNA secondary structural motifs in the $3'$-UTRs of differentially expressed and variable genes. The GENCODE vM14 APPRIS-annotated primary transcripts or failing that, the longest spliceforms were used in all analyses of RNA secondary structure.

4.2.2 Removal of maternal YTHDF2 increases transcript abundances

Analysis of the bulk scRNA-seq data for the mouse zygotes revealed that the absence of maternal YTHDF2 had a significant impact on mean gene expression in *Ythdf2*$^{\text{mCKO}}$ zygotes. 1,304 genes are differentially expressed between *Ythdf2*$^{\text{CTRL}}$ and *Ythdf2*$^{\text{mCKO}}$ zygotes, 2-fold higher mean expression (EFDR < 0.05, BASiCS) (Fig. 4.4a). Furthermore, we observed a slight but significant increase in overall mean expression for *Ythdf2*$^{\text{mCKO}}$ zygotes (Mann-Whitney-Wilcoxon test; **, $p < 4 \times 10^{-3}$) (Fig. 4.4d). Gene set enrichment analysis (GSEA) identified 4 Reactome pathway gene sets enriched amongst the differentially expressed genes, the two most significant are shown in **Fig. 4.4e**. According to the GSEA, removal of maternal YTHDF2 leads to concordant increases in the expression of genes involved in

nonsense-mediated mRNA decay (NMD) independent of exon junction complexes (EJCs) as well as genes engaged in the citric acid (TCA) cycle and respiratory electron transport.

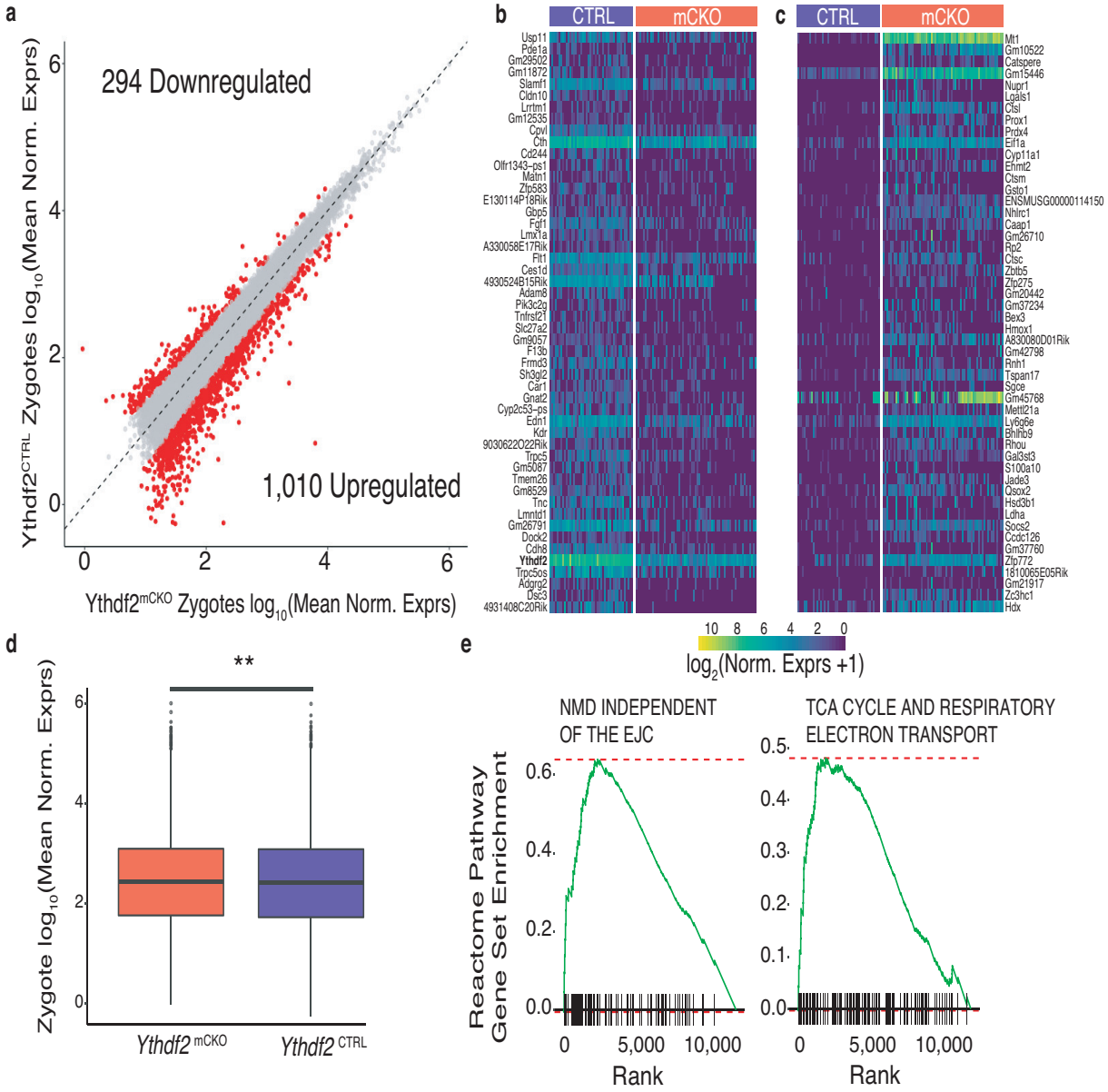


Fig. 4.4: Ythdf2^{mCKO} increases transcript abundances.

(a) Scatterplot of $\log_{10}$(mean normalised expression) levels in *Ythdf2*^{mCKO} and *Ythdf2*^{CTRL} zygotes. Significantly changing genes, greater than 2-fold (EFDR < 0.05, BASiCS), are highlighted in red. 1,010 genes are upregulated in the absence of maternal YTHDF2 and 294 downregulated. **(b)** Heatmap of the normalised expression for the 50 most downregulated genes. **(c)** Heatmap of the normalised expression for the 50 most upregulated genes.

Fig. 4.4: Ythdf2^{mCKO} increases transcript abundances (continued).
(d) There is a significant increase in overall mean expression in *Ythdf2*^{mCKO} zygotes compared to *Ythdf2*^{CTRL} zygotes (Mann-Whitney-Wilcoxon test; **, $p < 4 \times 10^{-3}$). **(e)** GSEA of Reactome pathways in *Ythdf2*^{mCKO} and *Ythdf2*^{CTRL} zygotes. Gene list sorted in descending order based on the mean expression $\log_2$(fold-changes). NMD = Nonsense-mediated decay, EJC = exon junction complex, TCA = citric acid cycle. *n Ythdf2*^{CTRL} zygotes = 46; *n Ythdf2*^{mCKO} zygotes = 67.

Furthermore, m^{6}A peaks from publicly available datasets are more enriched around the stop codons of upregulated genes than those of downregulated genes (Fisher's Exact test; ***, $p < 1 \times 10^{-9}$) (Fig. 4.5a). An earlier m6A-seq experiment by the O'Carroll Lab on oocytes from *Ythdf2*^{ko} and *Ythdf2*^{ctrl} mice had failed due to low library sequence complexity (data not shown). This was not repeated due to the large number of oocytes and dames that would be necessary. Interestingly, while there does not appear to be any significant difference in the lengths of 3′-UTRs between the *Ythdf2*^{mCKO} upregulated and downregulated genes (Fig. 4.5b), the 3′-UTRs of upregulated genes appear to be significantly more structured than those of downregulated genes (Mann-Whitney-Wilcoxon test; **, $p < 3 \times 10^{-3}$) (Fig. 4.5c). Furthermore the m^{6}A consensus sequence is present in structural motifs predicted from 3′-UTRs of upregulated genes (Fig. 4.5d).

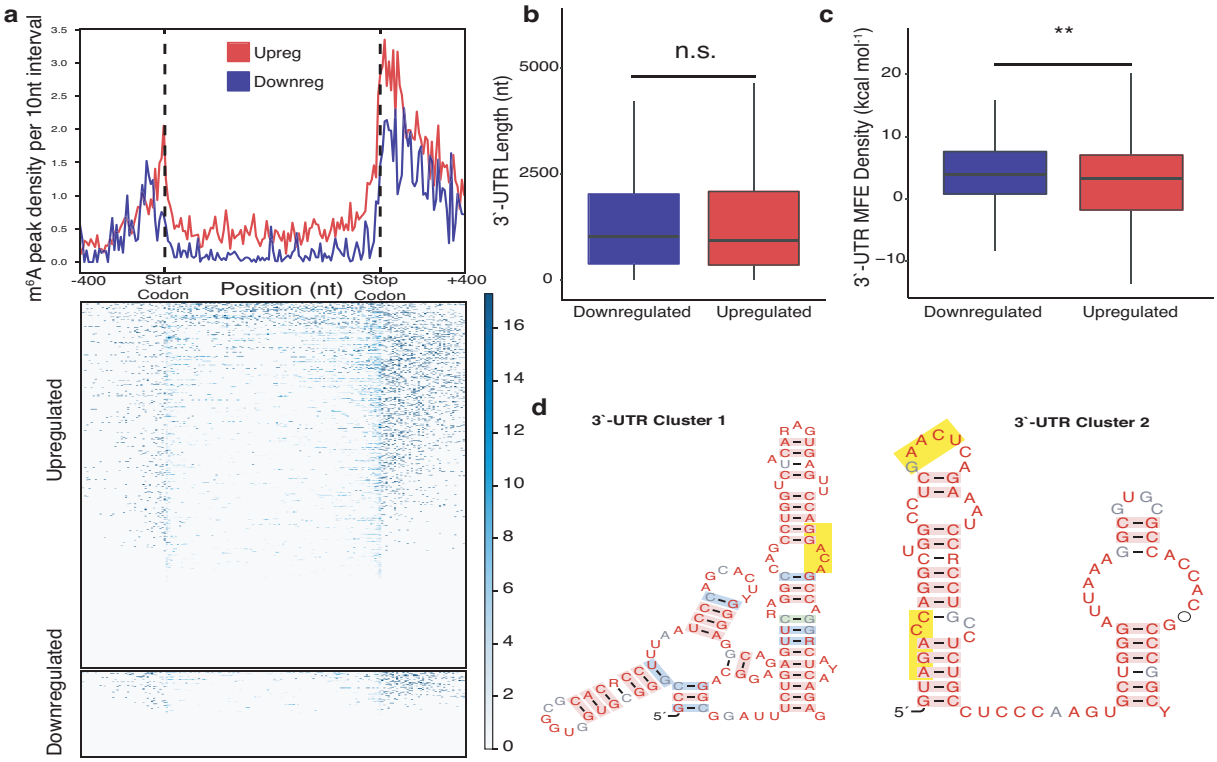


Fig. 4.5: m^{6}A enriched near the 3′-ends of *Ythdf2*^{mCKO} upregulated transcripts.

Fig. 4.5: m^{6}A enriched near the 3′-ends of *Ythdf2*^{mCKO} upregulated transcripts (continued). **(a)** Prevalence of RMBase m^{6}A peaks from public mouse datasets around the gene bodies of upregulated and downregulated transcripts in *Ythdf2*^{mCKO} zygotes. Gene bodies are scaled to the same length in each case. m^{6}A is more enriched in 3′ regions flanking the stop codons of upregulated genes than similar regions in downregulated genes (Fisher's Exact test; ***, $p < 1 \times 10^{-9}$). 3′ regions defined as stop codons ± 400nt. **(b)** 3′-UTRs lengths are not significantly different between the genes upregulated and downregulated with *Ythdf2*^{mCKO} (Mann-Whitney-Wilcoxon test; n.s.). **(c)** 3′-UTRs of upregulated genes are more structured. The median MFE density for the 3′-UTRs of upregulated genes is lower than that of the downregulated genes (Mann-Whitney-Wilcoxon test; **, $p < 3 \times 10^{-3}$). **(d)** Two of the secondary structural motifs predicted by GraphClust from the 3′-UTRs of upregulated genes. Instances of the m^{6}A RRACH consensus sequence in these structures are highlighted in yellow. *n Ythdf2*^{CTRL} zygotes = 46; *n Ythdf2*^{mCKO} zygotes = 67.

4.2.3 Transcript abundances increase heterogeneously in the absence of maternal YTHDF2

We assessed zygote-to-zygote transcriptional heterogeneity for both conditions by testing changes in over-dispersion estimates for genes that don't change in mean expression. Overall zygote-to-zygote transcriptional variability (over-dispersion) increases when maternally-supplied YTHDF2 is absent (Mann-Whitney-Wilcoxon test; ***, $p < 5^{-75}$) (Fig. 4.6a). Of the 1,724 genes that have significant changes in over-dispersion between the *Ythdf2*^{CTRL} and *Ythdf2*^{mCKO} zygotes, 1,628 genes becomes significantly more variable and 96 less variable with *Ythdf2*^{mCKO} (Fig. 4.6b). Significantly differentially variable genes are those that with a 1.5-fold increase in over-dispersion (EFDR < 0.05). Similar to the genes that become more abundant with *Ythdf2*^{mCKO}, this set of genes also exhibits an enrichment of m^{6}A peaks, from publicly available datasets, around their stop codons (Fig. 4.6c). A greater proportion of these genes have at least one m^{6}A peak within 400 nt of their stop codon (905 / 1,442) than do not (537 / 1,442) (Binomial Test, ***, $p < 1 \times 10^{-22}$).

As an extension of the analysis of cell-to-cell heterogeneity using the gene over-dispersion values we used the residual over-dispersion estimates generated by the BASiCS regression model to assess the transcriptional variability in the differentially expressed genes. This allowed us to evaluate the relationship, if any, between the effects of YTHDF2 on transcriptional variability and mean expression. 1,309 genes had significant changes in residual over-dispersion, a greater than 1.5-fold increase in the distance between residual over-dispersion estimates for the conditions (EFDR < 0.05). Taking the genes that were differentially expressed and differentially variable according to the residual over-dispersion estimates revealed four categories of genes: those with (i) higher abundances and increased

variability (mCKO$^+$ mCKO$^+$) with *Ythdf2*^{mCKO}, (ii) higher abundances and reduced variability (mCKO$^+$ CTRL$^+$), (iii) lower abundances and increased variability (CTRL$^+$ mCKO$^+$), and (iv) lower abundances and reduced variability (CTRL$^+$ CTRL$^+$) (Fig. 4.6d). Interestingly, this showed that for those genes upregulated with *Ythdf2*^{mCKO}, a greater proportion of these exhibit a concomitant increase in transcriptional variability than a reduction when compared to those that are downregulated (Fisher's Exact test; ***, $p < 2 \times 10^{-6}$). This suggests that zygotic transcript abundances have increased heterogeneously in the absence of maternal YTHDF2 and that m^{6}AMD may have a regulatory role in buffering transcriptional noise in the context of the mammalian MZT.

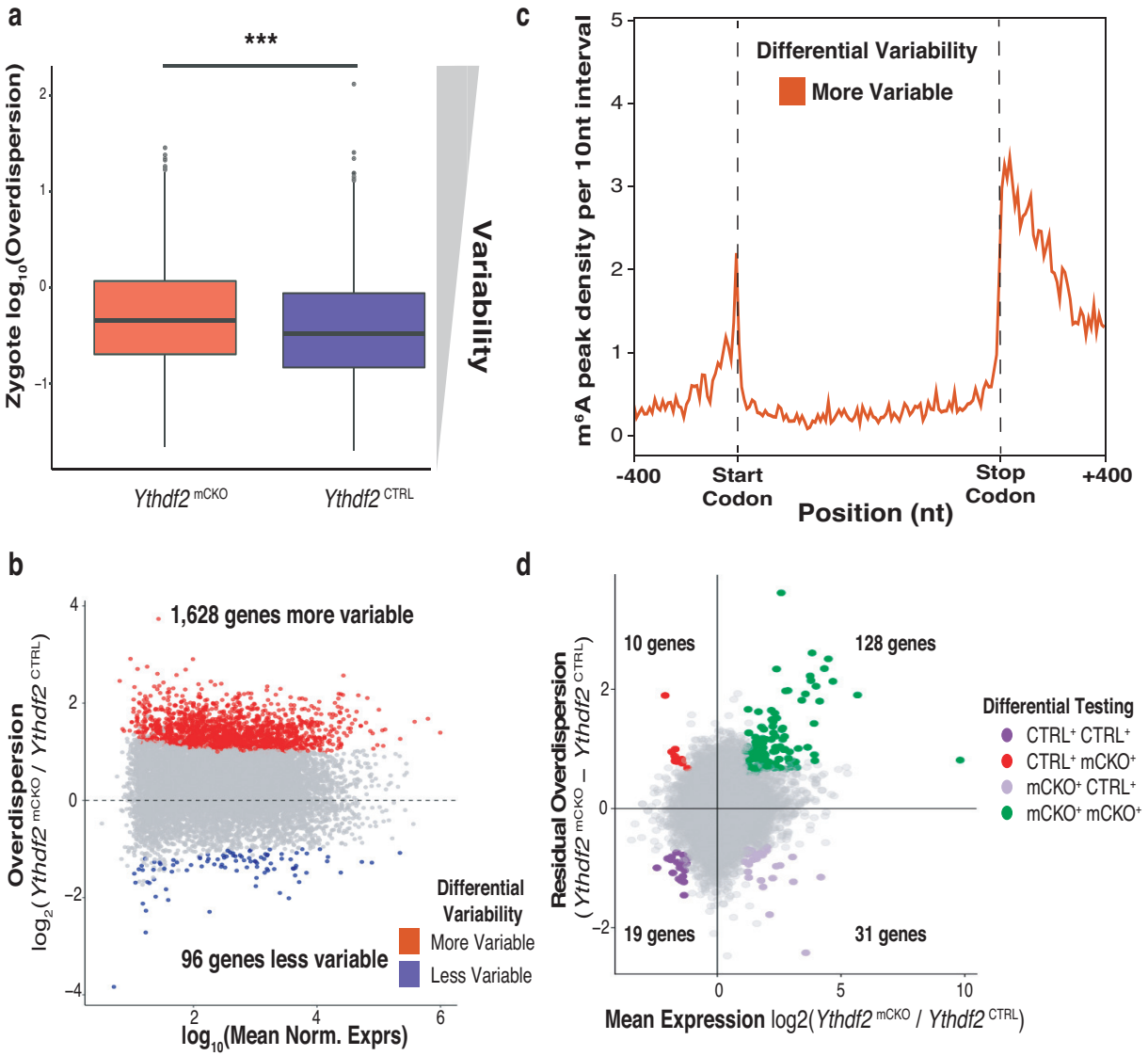


Fig. 4.6: *Ythdf2*^{mCKO} increases transcript abundances heterogeneously.

Fig. 4.6: *Ythdf2*$^{\text{mCKO}}$ increases transcript abundances heterogeneously (continued).
(a) Overall zygote-to-zygote transcriptional variability increases in the absence of maternal YTHDF2 (Mann-Whitney-Wilcoxon test; ***, $p < 5^{-75}$). **(b)** Scatterplot of mean expression and overdisperion fold-changes for genes that are not differentially expressed in response to *Ythdf2*$^{\text{mCKO}}$. The expression of 1,628 genes becomes significantly more variable and 96 less variable with *Ythdf2*$^{\text{mCKO}}$. Significantly differentially variable genes exhibit 1.5-fold increase in cell-to-cell variability (EFDR < 0.05). Significantly more variable genes are highlighted in red and less variable genes in blue. **(c)** Prevalence of RMBase m^{6}A peaks from public mouse datasets across the gene bodies of transcripts that are more variable in *Ythdf2*$^{\text{mCKO}}$ zygotes. Gene bodies are scaled to the same length in each case. A greater proportion of the genes have at least one m^{6}A peak within 400 nt of their stop codon (905/1,442) than do not (537/1,442) (Binomial Test, ***, $p < 1 \times 10^{-22}$). **(d)** Scatter plot of mean expression and residual over-dispersion differences between *Ythdf2*$^{\text{CTRL}}$ and *Ythdf2*$^{\text{mCKO}}$ zygotes. 1,309 genes had significant changes in residual over-dispersion, greater than 1.5-fold increase in the distance between residual over-dispersion estimates (EFDR < 0.05). Genes with higher abundances and increased variability with *Ythdf2*$^{\text{mCKO}}$ are highlighted in green (mCKO$^+$ mCKO$^+$; 128 genes), higher abundance and reduced variability in pink (mCKO$^+$ CTRL$^+$; 31 genes), (iii) lower abundance and increased variability in red (CTRL$^+$ mCKO$^+$; 10 genes) and (iv) lower abundance and reduced variability in blue (CTRL$^+$ CTRL$^+$; 19 genes). A signification proportion of upregulated genes become more variable with *Ythdf2*$^{\text{mCKO}}$ (Fisher's Exact test; ***, $p < 2 \times 10^{-6}$). (*n Ythdf2*$^{\text{CTRL}}$ zygotes = 46; *n Ythdf2*$^{\text{mCKO}}$ zygotes = 67.)

4.2.4 Mouse pre-Leukaemic Stem Cell scRNA-Seq using the droplet-based 10x Genomics™ system

The absence of matching m^{6}A data for the mouse zygotes meant that if we wanted to establish a direct regulatory relationship between m^{6}A and transcriptional variability we needed to use a system for which there was m^{6}A data available and YTHDF2 was known to affect transcript abundances. The labs of Kamil Kranc and Dónal O'Carroll derived YTHDF2 KO in order to investigate the role YTHDF2 in AML initiation and progression Paris et al. (2019). They isolated c-Kit$^+$ HSPCs from the livers of Ythdf2$^{\text{HA-Fl/HA-Fl}}$ (*Ythdf2*$^{\text{WT}}$) and Ythdf2$^{\text{HA-Fl/HA-Fl}}$; Vav-iCre (*Ythdf2*$^{\text{CKO}}$) foetuses and co-transduced these with MSCV-Meis1a-puro and MSCV-Hoxa9-neo retroviruses and passaged them to generate pre-leukaemic stem cells (pre-LSCs). In collaboration with Nils Eling and Christina Ernst from the lab of John Marioni (EMBL-EBI / CRUK Institute, Cambridge), we performed droplet-based scRNA-seq using the 10x Genomics™ technology for three replicates of the *Ythdf2*$^{\text{WT}}$ and *Ythdf2*$^{\text{CKO}}$ pre-LSCs. This technique captures cell-specific transcriptomes by counting the number of barcoded mRNA 3′-ends (Zheng et al., 2017). Furthermore, we generated duplicate bulk RNA-seq libraries from these samples in order to obtain marker genes to identify the cells in each sequenced single-cell library. pre-LSCs were cultured at the CRUK Institute, Cambridge by Christina

Ernst according to the instructions provided by the Kranc Lab (see Paris et al. (2019)). Single-cell suspensions were loaded into individual channels of the Chromium2^{TM} Single Cell A Chip (10x Genomics™) with the aim of recovering 4,000-5,000 high-quality cells per library. The Chromium Single Cell 3′ Library and Gel Bead Kit v2 (10x Genomics™) was used for single-cell barcoding, cDNA synthesis and library preparation, following the manufacturer's instructions in the Single Cell 3′ Reagent Kits User Guide. Reverse transcription takes place within each droplet and the barcoded cell cDNA libraries are amplified together. Oligonucleotide barcodes consist of sequencing adapters and primers, a 14 bp cell-specific barcode, a 10 bp unique molecular identifier (UMI) randomer that is unique to individual cDNA molecules, and a 30 bp oligo-dT adapter (Zheng et al., 2017). Libraries were sequenced together on an Illumina HiSeq 2500 using 75 bp paired-end sequencing. The samples included in each cDNA library and the number of cells captured are listed in **Table 4.2**. The experiment was designed such that three of the four 10x libraries contained pairs of *Ythdf2*^{WT} and *Ythdf2*^{CKO} samples while the fourth was comprised of cells from all six samples. This was done to ensure that YTHDF2 status would not be confounded with batch. Furthermore the fourth library would allow us to detect any major technical variation. Library-specific batch effects would have been detected if the cells from other library had radically diverged from this batch.

Nils Eling (Marioni Lab) processed the scRNA-seq data using the 10X Genomics™ *Cell Ranger* pipeline (Zheng et al., 2017). We obtained gene-specific transcript counts for 23,397 cells using the Cell Ranger *count* function with default settings (**Table 4.2**). This pipeline aligned reads against the *Mus musculus* reference genome (GRCm38) and Ensembl transcript annotation (version 89) (Cunningham et al., 2019) using STAR and quantifies the UMIs per gene and cell. Cell Ranger only retains cells with similar UMI distributions (Zheng et al., 2017).

Table 4.2: pre-LSC Samples in each 10x Single Cell Library.

Samples	Library	Cell Ranger	QC Filtering	Cluster Filtering
KO5 & WT2	do26181	5,897 cells	5,645 cells	4,881 cells
KO6 & WT4	do26182	5,155 cells	4,673 cells	2,153 cells
KO7 & WT16	do26183	5,552 cells	5,261 cells	4,765 cells
All cKO & WT	do26184	6,793 cells	6,204 cells	5,721 cells
Total		23,397 cells	21,783 cells	17,520 cells

We used the default Cell Ranger thresholds in order to obtain high-quality cells with a large number of UMIs. The R/Bioconductor package *DropletUtils* (Lun et al., 2019) was used to

remove any swapped barcodes that resulted from technological sequencing errors. Cell and gene QC was performed with the R/Bioconductor packages *scater* and *scran*. The number of cells remaining after QC for each library can be seen in **Table 4.2**. To avoid biases due to difference in library sizes, we down-sampled the UMI counts prior to quality filtering. We removed cells that had too few or too many UMIs (Fig. 4.7a), less than 8.25 or greater than 9.75 log(total UMI counts per cell), and removed cells expressing too few or too many features, less than 1,500 or greater than 4,000 genes (Fig. 4.7b). Furthermore, we excluded cells with $\geq$3% of UMIs mapping to the mitochondrial genome (Fig. 4.7c).

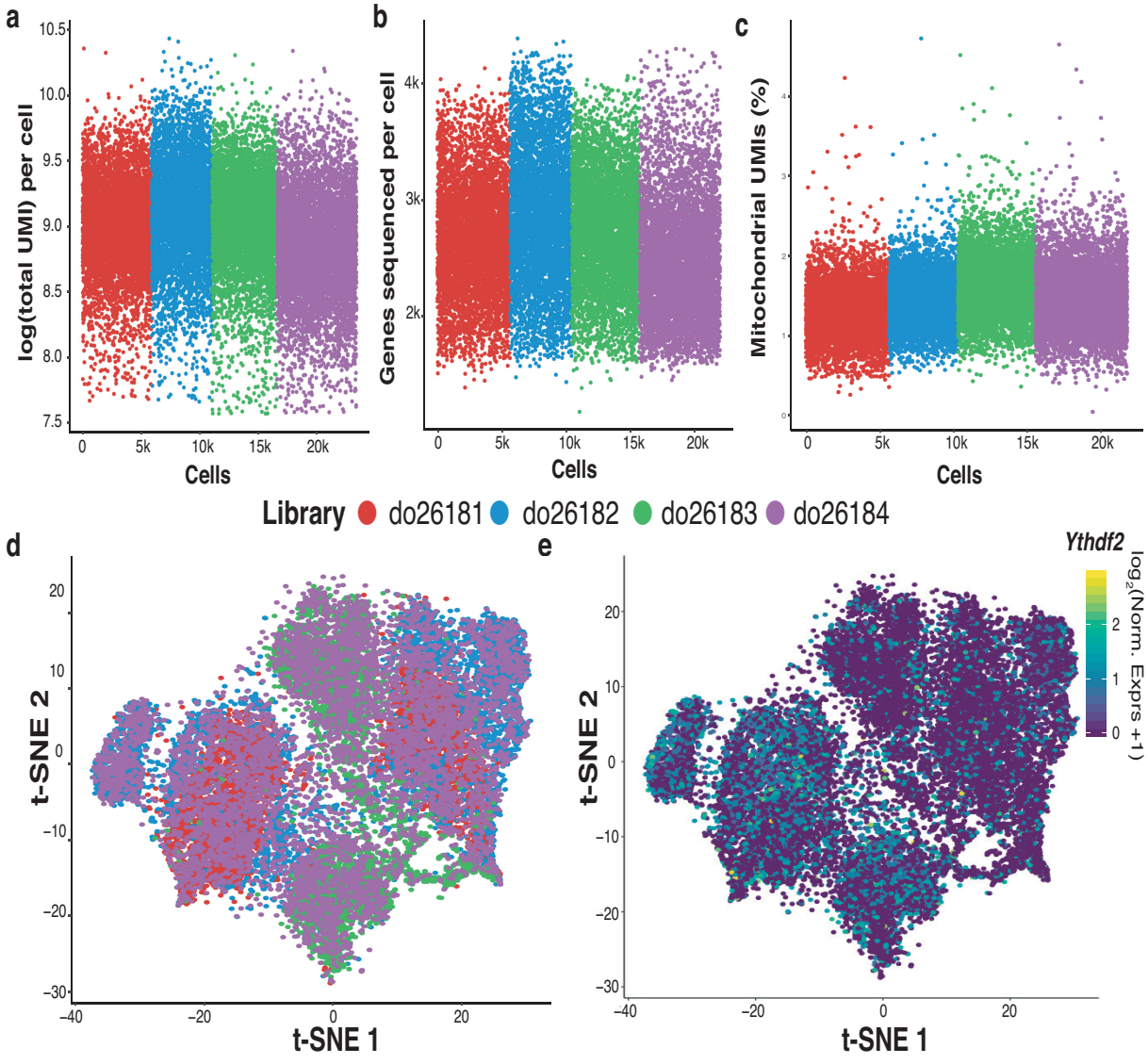


Fig. 4.7: Quality Control for droplet-based single-cell RNA-seq data.

Fig. 4.7: Quality Control for droplet-based single-cell RNA-seq data (continued).
(a) Raw UMI counts per cell for the four libraries sequenced. do26183 had the smallest library size. **(b)** Raw number of genes expressed per cell. **(c)** Raw percentage reads mapping to the mouse mitochondrial genome. t-SNE dimensionality reduction based on cells from the pre-LSCs samples after QC. Cells are coloured by their library of origin in **(d)** and their $\log_2$(Normalised *Ythdf2* Expression) in **(e)**. do26181 library, n = 5,645 cells; do26182 library, n = 4,673 cells; do26183 library n = 5,261 cells; do26184 library n = 6,204 cells.

The transcript counts of quality filtered cells were normalised using *scran*. For this, cells with similar transcriptomic complexity were clustered together using a graph-based method implemented in the *scran quickCluster* function with a maximum cluster size of 2,000 cells. Size factors were calculated for each of the seven clusters before being scaled between clusters using the *computeSumFactors* function. All further analyses looking at single-cell RNA expression use these $\log_2$-transformed, normalised counts (including a pseudocount) i.e. $\log_2$(normalised expression + 1). The pre-LSCs that passed QC filtering are visualised with the t-SNE dimensionality reduction technique in **Fig. 4.7d**. It is apparent from the dimensionality analysis that the droplet-based scRNA-seq did not introduce much technical transcriptional variation. No individual library clusters separately from the others. Most importantly, cells from the library do26184, which contained cells from all 6 samples, localised with cells from the other three libraries. However, there is not as obvious a structure to the data as for the zygote scRNA-seq data, *ythdf2* is expressed across many cells but is potentially more abundant in cells in the lower half of the t-SNE (Fig. 4.7e). Worryingly, rather than there being two clusters (or even one cluster) of cells as one might expect if YTHDF2 had significant impact upon pre-LSCs transcriptome, there seem to one cluster per sample as evidenced by the fact that there are two clusters of cells for libraries do26181, do26182 and do26183 (Fig. 4.7d). As a result, further cell filtering was necessary. Lowly expressed genes, with an average $\log_2$-transformed, normalised expression < 0.1 per cluster were excluded. Furthermore, it was determined that cluster 6 contained a heterogeneous set of cells that had only been clustered together because they had to be included in a cluster. After removing this cluster, we retained more than 17,000 high-quality single cells expressing 5,468 genes **(Table 4.2)**. The final set of filtered cells are visualised with the t-SNE dimensionality reduction technique in **Fig. 4.8**. Although *Ythdf2* is expressed in all six of the remaining clusters, it was most highly expressed in clusters 3,4, and 7 and as a result these were annotated as KO1, KO2, and KO3. The other three remaining clusters were annotated as WT1, WT2 and WT3 (Fig. 4.8b). The UMI counts per cell per cluster were pooled in order to detect differentially expressed genes between the *Ythdf2*$^{\mathrm{WT}}$ and *Ythdf2*$^{\mathrm{CKO}}$ pre-LSCs. The pooled counts were normalised with *edgeR* McCarthy et al. (2012) and differential expression analysis performed using the edgeR Quasi-Likelihood Test. Only

two significantly differentially expressed genes, *Ythdf2* and *Cebpe*, were detected with a 2-fold increase in mean expression (FDR < 0.1).

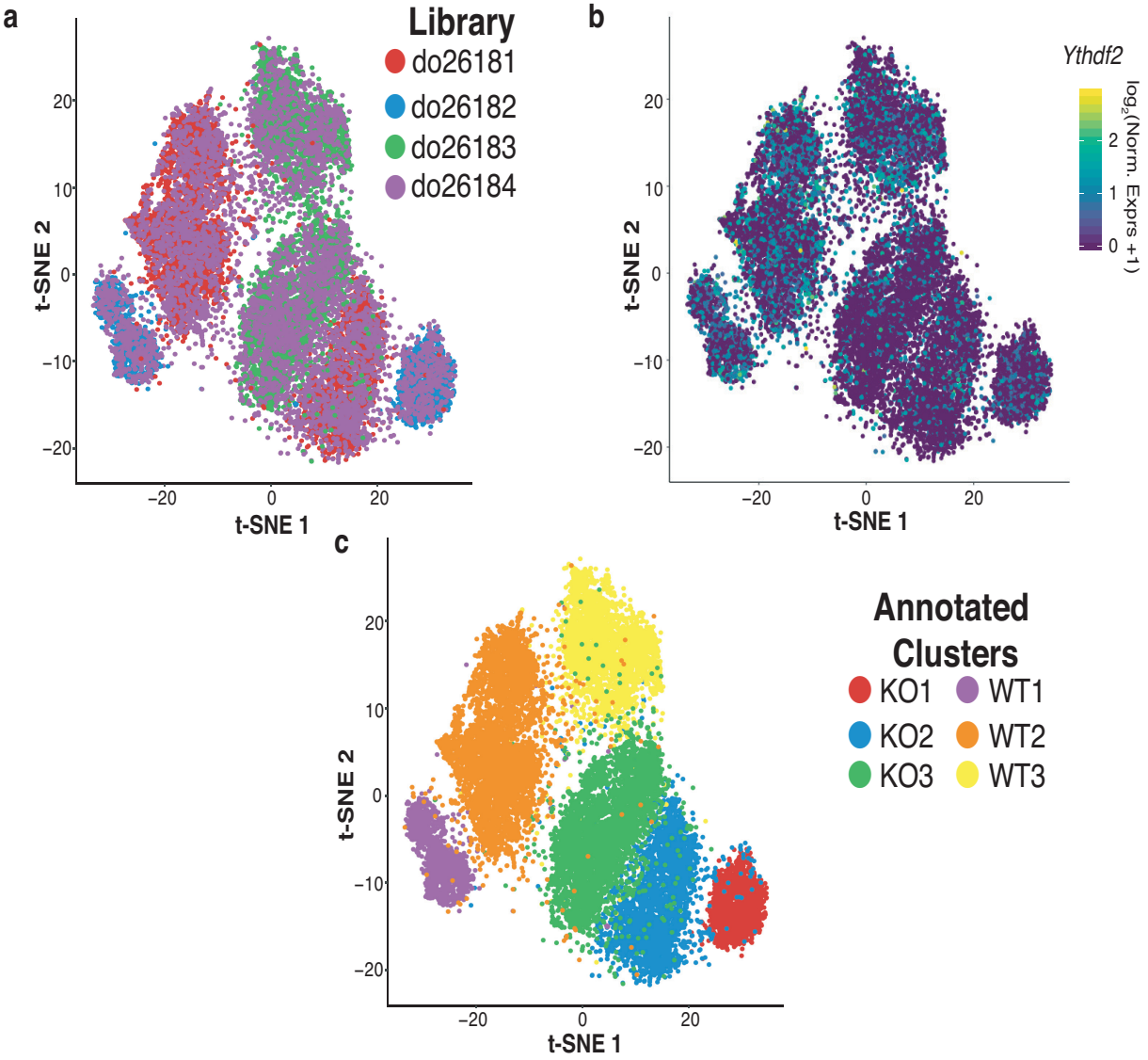


Fig. 4.8: Visualisation of filtered pre-LSCs.

t-SNE dimensionality reduction for the final filtered set of pre-LSCs. Cells are coloured based on their library of origin in **(a)**, their $\log_2$(Normalised *Ythdf2* Expression) in **(b)** and their annotated cluster in **(c)**. *n* cells do26181 library = 4881; *n* cells do26182 library = 2153; *n* cells do26183 library = 4765 ; *n* cells do26184 library = 5721.

We made a final attempt at improving the cell sample assignment by performing bulk RNA-seq in duplicate for the six pre-LSCs samples. Christina Ernst (Marioni Lab) prepared cDNA libraries from pre-LSC RNA that had been DNA- and ribo-depleted with Ribo-Zero Gold

according to manufacturer's instructions (Illumina, RS-122-2303) using the TruSeq Stranded Total kit 2. Libraries were sequenced on an Illumina HiSeq 2500 (Genome Core Facility, CRUK Institute, Cambridge) using a paired-end 125 bp run. Nils Eling mapped paired-end FASTQs against the *Mus musculus genome* (GRCm38) and Ensembl transcript annotation (version 89) using STAR (v2.5.3) with its default parameters. Strand-specific gene counts were quantified from the read alignments by setting the htseq-count stranded parameter to "reverse". We visualised several features of the aligned and counted data (number of intronic/exonic reads, number of multi-mapping reads, low-quality reads and total library size) and did not detect any low-quality bulk RNA-seq libraries (data not shown). Lowly expressed genes (average count < 10 across all samples) were excluded from downstream analysis and visualisation (Fig. 4.9).

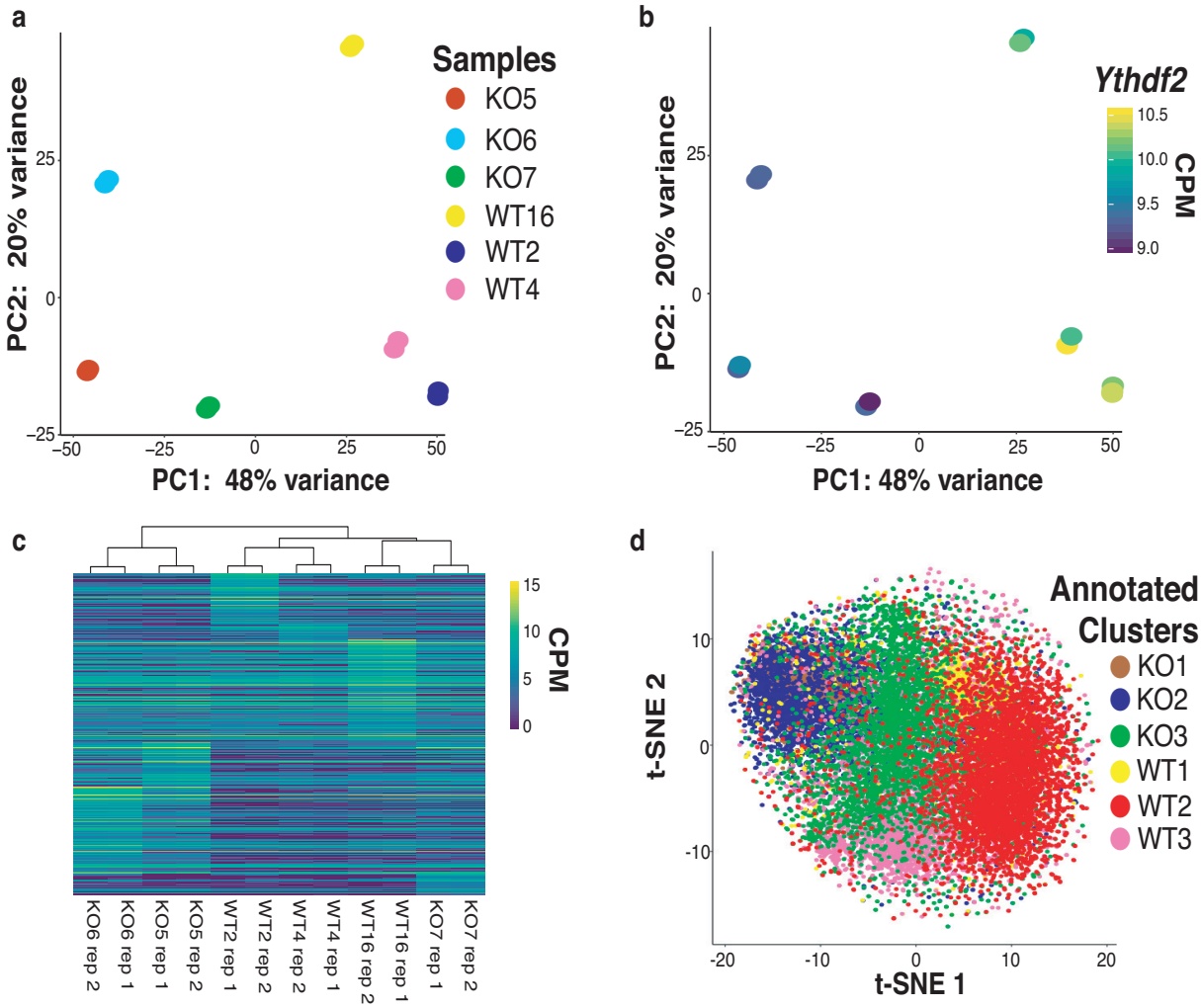


Fig. 4.9: Analysis of pre-LSC bulk RNA-seq marker genes.

Fig. 4.9: Analysis of pre-LSC bulk RNA-seq marker genes (continued).
PCA based on filtered bulk RNA-seq counts for pre-LSC samples. Samples coloured based on the sample of origin in **(a)** and coloured their *Ythdf2* expression in counts per million (CPM) in **(b)**. **(c)** Unsupervised clustering of bulk RNA-seq samples according to marker gene expression. **(d)** t-SNE dimensionality reduction analysis for filtered scRNA-seq data using the marker genes detected in the bulk RNA-seq data. Cells are coloured according to their annotated cluster. *n* WT2 = 2; *n* WT4 = 2; *n* WT16 = 2; *n* KO5 = 2; *n* KO6 = 2; *n* KO7 = 2.

PCA for the filtered bulk RNA-seq samples is shown in **Fig. 4.9a**. PC1, which explains 48% of the variance, separates the *Ythdf2*$^{\text{CKO}}$ from *Ythdf2*$^{\text{WT}}$ samples. However PC2 which explains 20% of the variance, shows that there is a large amount of sample-to-sample variation within the conditions. The WT12 replicates are as distinct from the other *Ythdf2*$^{\text{WT}}$ samples as two of the *Ythdf2*$^{\text{CKO}}$ samples. The filtered bulk RNA-seq counts were normalised with edgeR and marker genes identified for each sample by performing pairwise contrasts between all six samples using the edgeR quasi-likelihood tests. Marker genes were those that were significantly differentially expressed with a greater than 2-fold increase in mean expression (FDR < 0.1) in every pairwise contrast made for a given sample. A total of 966 marker genes were identified, their expression in CPM is displayed in the heat map in **Fig. 4.9c**. Surprisingly, the KO7 bulk RNA-seq replicates cluster with the WT samples based on their marker gene expression. There is no major separation between the six annotated clusters of single-cells when t-SNE dimensionality reduction analysis is performed using the marker genes present in the filtered scRNA-seq data (Fig. 4.9d). The use of marker genes from the bulk RNA-seq does not appear to improve the single-cell sample assignments.
Unfortunately, as is even apparent from the analysis of the bulk RNA-seq data, the pre-LSCs samples are far too heterogeneous to perform any meaningful investigation into the role of YTHDF2 in regulating transcriptional variability in murine pre-LSCs. Nils Eling (Marioni Lab) performed exploratory analyses of transcriptional variability in the final set of pre-LSC clusters with the BASiCS regression model. The BASiCS MCMC was run 40,000 iterations with 20,000 burn-in iterations and a thinning value of 20. However from the results, it was apparent that there were differences in mean gene expression and transcriptional heterogeneity between the clusters from the same condition (data not shown).

4.3 Discussion

Stochasticity is an inherent feature of biological processes involving molecular interactions (Ecker et al., 2017). When molecules in a system are present in large quantities, biological noise is of little consequence for the stability of the system (Swain et al., 2002). However, molecules are generally found at relatively low abundances in a biological cell. Poly(A)-tails

facilitate nuclear export (Fuke and Ohno, 2008) but the main purpose of mRNA poly(A)-tails is to stabilise the rate of translation initiation (Wakiyama et al., 2000) by protecting transcripts from premature degradation (Lim et al., 2014). Due to its involvement in protecting transcripts from decay, most mRNA degradation pathways involve a de-adenylation step (Beilharz et al., 2009; Mukherjee et al., 2002; Yamashita et al., 2005). Notably it has been shown that one such decay pathway, the miRMD pathway, is involved in transcriptional and protein noise control (Gambardella et al., 2017; Schmiedel et al., 2015). m^6A is the most abundant non-terminal RPTM in vertebrate mRNAs (Desrosiers et al., 1974) and has been shown to regulate transcript stability (Du et al., 2016; Geula et al., 2015; Wang et al., 2013, 2014). Similar to the miRMD pathway, m^6AMD also involves CCR4-NOT deadenylation in the P-bodies (Wang et al., 2013). Furthermore, while the YTHDF2-mediated degradation of m^6A-modified maternal mRNAs is essential for the mammalian MZT (Ivanova et al., 2017), it is dispensable in zebrafish (Zhao et al., 2017). Shared effectors and overlapping targets suggest that there might be interplay between miRNAs and m^6AMD. Given that miRNAs are established regulators of transcriptional noise (Gambardella et al., 2017; Schmiedel et al., 2015), it is possible that m^6AMD also modulates it.

It is unsurprising how few cells from 2-cell stage embryos passed the QC given the developmental block and defects observed at this stage by (Ivanova et al., 2017). Furthermore, most of these cells failed QC due to high levels of mitochondrial gene expression which is indicative of cell death or damage (Ilicic et al., 2016). The significant role of maternal YTHDF2 in regulating zygote transcriptional homoeostasis is apparent from the clear separation between the *Ythdf2*^{CTRL} and *Ythdf2*^{mCKO} zygotes in **Fig. 4.3a**. The impaired m^6AMD of transcripts is enough to differentiate the 2 populations of zygotes. Similar to the results of Ivanova et al. (2017) in metaphase II-arrested oocytes, we identified a large number of genes whose mean expression is altered in the absence of YTHDF2. The vast majority of these are upregulated. Many of these genes were shown to be m^6A-modified based on publicly available m^6A datasets. Furthermore, m^6A peaks were more enriched around the stop codons of upregulated genes than downregulated genes. Interestingly the 3′-UTRs of upregulated genes were more structured than those of the downregulated genes. Some of the *in silico* predicted 3′-UTR structural motifs contained the m^6A consensus sequence as has been described previously for the "m^6A structural switches" characterised by Liu et al. (2015).

GSEA of zygote expressed transcripts revealed the concomitant upregulation of genes associated with NMD independent of the EJC in the absence of maternal YTHDF2. NMD is a eukaryotic QC mechanism for removing transcripts with premature stop codons (Kurosaki and Maquat, 2016). Li et al. (2019) have recently shown that m^6A regulates NMD in human glioblastoma. In the absence of m^6AMD, this may represent a compensatory mechanism

for the removal of accumulating maternal transcripts that would otherwise decay in the cytoplasm in a less regulated manner. 3′-UTR EJC-independent NMD is modulated by the 3D distance between cytoplasmic PABPC1 and the stop codon and can be inhibited if PABPC1 is sufficiently close in 3D space to a stop codon (Silva et al., 2008). Therefore in the absence of YTHDF2-mediated deadenylation, the poly(A)-tails of m^6A-modified transcripts will remain long for a greater period of time and the inhibition of EJC-independent NMD would be relaxed for these transcripts.

Analysis of the transcriptional variability in *Ythdf2*^{CTRL} and *Ythdf2*^{mCKO} zygotes revealed that overall cell-to-cell heterogeneity increases in the absence of maternal YTHDF2. Of the non-differentially expressed genes whose variability is altered with *Ythdf2*^{mCKO}, the vast majority became more variable. Most of these genes were shown to be m^6A-modified based on publicly available m^6A datasets. Interestingly, an analysis of the differential expression and variability using the BASiCS regression model revealed that the abundances for a number of transcripts increased heterogeneously in the absence of maternal YTHDF2. The results of these analyses suggest that in addition to its involvement in clearing maternal transcripts as part of the MZT (Ivanova et al., 2017; Zhao et al., 2017), YTHDF2 may also buffer transcriptional noise in target genes. However, without matching m^6A data from mouse zygotes, this additional regulatory activity remains speculative for now.

The developmental block with *Ythdf2*^{mCKO} at the 2-cell embryo stage (Ivanova et al., 2017) prevented further investigation into the role of YTHDF2 in buffering transcriptional noise during mouse embryogenesis. We had hoped that by studying the impact of YTHDF2 on transcriptional variability in another system, mouse pre-LSCs, for which there was m^6A-seq data available we would determine whether or not this was a zygote-specific or more general post-transcriptional regulatory phenomenon. It appears that the mouse pre-LSCs established by Paris et al. (2019) were not a suitable system for dissecting this additional function of YTHDF2. It was not possible to explore any relationship between m^6AMD and transcriptional noise control using the pre-LSC scRNA-seq dataset. As is even apparent from analysis of the bulk RNA-seq expression, the pre-LSCs samples are far too heterogeneous to perform any meaningful investigation into the role of YTHDF2 in regulating transcriptional variability in murine pre-LSCs. Paris *et al.* derived the six pre-LSC samples we used from six different animals, three *Ythdf2*^{CKO} and three *Ythdf2*^{WT}. Populations of foetal HSPCs were individually co-transduced with MSCV-Meis1a-puro and MSCV-Hoxa9-neo retroviruses and passaged separately in order to generate pre-LSCs. It is probable that this had led to the sample-specific accumulation of differences and has led to the divergence of the biological replicates that is apparent from the bulk and single-cell RNA-seq data. Furthermore, although Paris *et al.* detected 1,282 genes that are differentially expressed in pre-LSCs ($p <$

0.05) with *Ythdf2*^{CKO} (see Fig. 4a, Paris et al. (2019)) very few of these genes have large fold-change increases. A fold-change cut-off is not specified in the text. This suggests that either YTHDF2 transcriptional regulation is relatively inconsequential in mouse pre-LSCs or that their statistical analyses also suffered as a result of sample heterogeneity. Biological replicates are essential for any experiment but they must be generated in a sensible manner that does not compromise the study by introducing artefactual differences between them.
YTHDF2-mediated transcriptional degradation seems to be most significant in differentiation processes (Geula et al., 2015; Ivanova et al., 2017; Lee et al., 2019; Lv et al., 2018; Zhao et al., 2017). Therefore a more standard, better characterised, system that is capable of undergoing differentiation (such as mESCs) would have been a more appropriate model in which to follow-up the analysis of transcriptional noise in the mouse zygote scRNA-seq data.

5

Discussion

5.1 Conclusion

My PhD research has focussed on the regulation of mammalian gene expression. Eukaryotic transcription is the episodic, three stage process by which RNAPs synthesise RNA from a DNA template. Eukaryotic gene expression is a noisy process that is subject to multiple layers of regulation. Key features of this are the 3D chromatin organisation of eukaryotic genomes and the post-transcriptional control of RNA activities. Eukaryotic nuclear DNA is tightly packaged as chromatin that is further folded into higher order structures. The 3D folding of eukaryotic genome lends itself to the formation of interactions between otherwise distant regions of the genome. However, highly condensed, compacted chromatin can obscure DNA regulatory elements from the trans-acting factors that regulate cellular activities. Epigenetic regulation of chromatin accessibility can mitigate this and facilitate episodic transcriptional initiation. The formation of TADs, highly self-interacting regions of chromatin, helps delineate the boundaries of transcription. These are largely stable across animal tissues and conserved between closely related species.

In **Chapter 2**, I investigated the impact of the integration of the genome of a short dsDNA virus, HPV16, on host chromatin architecture and transcriptional regulation in collaboration with the lab of Nick Coleman. We successfully adapted the Promoter Capture Hi-C method developed in the lab of Peter Fraser to enrich for HPV 16 chromatin. We identified a range of interactions between host chromatin and viral integrants when we applied this SCRiBL technique to five cell lines derived from the W12 model for cervical carcinogenesis. These cell lines had been derived previously under non-competitive conditions in order to capture the changes that occur in early cervical carcinogenesis before cancer selection. While it is already established that HR-HPV can interact with host chromatin in the malignant HeLa cell

line, this revealed that viral integrants interact with host chromatin in pre-malignancy, soon after integration. My analysis of viral integration sites revealed that they largely coincide with regions of accessible, transcriptionally permissive chromatin as determined from DNAse I hypersensitive regions and active histone marks. Viral integration occurs in the middle of TADs without disrupting them but can alter interactions within them. The majority of viral integrants in the W12 clones tested were found within host genes. In each instance, this led to the upregulation of the host gene when compared with the average expression in the other clones. Furthermore, transcription from the viral early promoter can continue into neighbouring host genes and produce chimaeric fusion transcripts. HPV16 integrants affect the expression of neighbouring and more distal genes, up to at least 2.5 Mb away, on the chromosome of integration. This is likely due to the combined effect of novel chromatin interaction and transcription from the viral early promoter.

Stochasticity is inherent to biological processes involving interactions between discrete, low abundance molecules. The advent of reliable protocols for performing scRNA-seq has revealed that transcriptional noise, the unstructured cell-to-cell variation in transcript abundances, is a widespread and biologically important feature in many populations of cells. Genes are subject to intrinsic sources, that are more promoter- or gene-specific, and extrinsic sources of noise that are more cell and context-specific. Transcriptional noise is modulated by features affecting the frequency and rate of transcription such as epigenetic modifications and CREs.

Organismal-ageing is associated with, if not even caused by, the progressive decline in biological function as a result of the accrual of molecular damage. This affects many of the epigenetic and chromatin regulatory pathways regulating transcription and transcriptional noise. Ageing is also associated with a decline in male fertility. Increased paternal age is also associated with a greater number of *de novo* mutations in the offspring. Some have attributed both of these phenomena to the clonal expansion of 'selfish' spermatogonial lineages. In **Chapter 3**, I investigated the consequences of ageing and testicular injury & regeneration on the transcriptomes of sorted populations of mouse undifferentiated spermatogonia using bulk and single-cell RNA-seq. While subtle changes in mean gene expression are detectable, it was apparent that cell-to-cell transcriptional variability in undifferentiated spermatogonia declines with ageing. This may reflect the phenomenon of selfish spermatogonial selection. Multiple recent studies have detected increased cell-to-cell transcriptional heterogeneity in aged mammalian somatic tissues. This has led some to suggest that increased transcriptional noise should be added to the nine *Hallmarks of Ageing* defined by López-Otín et al. (2013). However, in light of the results presented in this chapter it may be more appropriate to consider a broader definition of *altered* transcriptional noise. Gametes and their progenitors

are transcriptionally very distinct from somatic cells and what may be true for the soma is not necessarily the case in germ cells.
It has recently been determined that the post-transcriptional regulation of cytoplasmic transcript stability and decay also have a role in modulating and buffering transcriptional noise. Unstable transcripts are more susceptible to noise than stable ones. Cytoplasmic mRNA stability is anti-correlated with expression noise in mESCs. A major determinant of RNA stability is the 3′ poly(A) tail. It inhibits both of the two main mRNA exoribonucleolytic decay pathways and its removal is essential for their progression. Notably, the miRNA-Mediated mRNA Decay pathway has been shown to modulate heterogeneity during the noisy transcriptional shifts that facilitate cell differentiation. This is particularly the case for lowly expressed genes and is enhanced when a transcript is targeted by multiple miRNAs. In addition to targeting transcripts for CCR4-NOT mediated deadenylation, certain miRNAs enhance the post-transcriptional m^6A methylation of their targets. 3′ m^6A methylation, through its reader YTHDF2, also targets transcripts for CCR4-NOT deadenylation. Indeed, miRNAs and m^6A coordinate the decay of transcripts during the zebrafish MZT. Given that miRMD and m^6AMD share effectors and targets it also possible that m^6A and its reader YTHDF2 also modulate transcriptional noise.
In **Chapter 4**, I explored the YTHDF2-regulated transcriptional dynamics in mouse zygotes and pre-leukaemic stem cells using plate- and droplet-based single-cell RNA-seq data. Analysis of single-cell expression in *Ythdf2*$^{\text{mCKO}}$ and *Ythdf2*$^{\text{CTRL}}$ zygotes revealed that depletion of maternally-supplied YTHDF2 leads to the widespread upregulation of genes. The majority of these transcripts seem to m^6A-modified at or around their stop codons based on publicly available m^6A datasets. Notably, their 3′-UTRs are more structured than those of the down-regulated transcripts and many of their *in silco* predicted secondary structures contain the m^6A METTL3/14 consensus motif. Most interestingly, a subset of these transcripts are heterogeneously upregulated in the absence of YTHDF2. This suggests that YTHDF2-mediated deadenylation modulates their noise expression under normal circumstances. Unfortunately, I was unable to determine whether this is a zygote-specific or more general mammalian regulatory phenomenon due to issues with the follow-up scRNA-seq experiment in mouse pre-LSCs. The pre-LSC replicates characterised by Paris *et al.*, 2019 were too divergent from each other for there to be any point in exploring the role of YTHDF2 in buffering noise. Any differences in cell-to-cell heterogeneity could just as easily result from *structured* transcriptional variability due to poor sample replication as from changes in *unstructured* transcriptional noise due to altered YTHDF2 expression. Unless and until the results from the zygotic scRNA-seq data are replicated in system for which matching m^6A data is available, this additional YTHDF2 regulatory activity remains speculative for now. Of note, it has been

shown that overall m^6A methylation in human peripheral blood mononuclear cells declines with age (Min et al., 2018). This suggests that the dysregulation of m^6A methylation (and potentially other RPTMs) might contribute to the alterations in transcriptional noise and decline in biological function associated with ageing.

5.2 Future Direction

Research into post-transcriptional regulation of mRNA stability and decay has undergone a renaissance with the emergence of RNA post-transcriptional modifications. Over 500 studies alone have profiled the m^6A-methylome across many diverse tissues and organisms under various conditions and treatments (Zaccara et al., 2019). It is well established that the m^6A reader YTHDF2 impacts cytoplasmic transcript stability but questions still remain about how m^6A regulates transcript half-lives. The removal of METTL3 has a far greater impact on turnover than the abrogation of YTHDF2 activity (Ke et al., 2017; Wang et al., 2013); this hints at the existence of additional m^6A readers that promote decay. YTHDF2 shows decreased binding specificity for m^6A when YTHDF3 expression is reduced (Shi et al., 2017). The context-specific m^6A methylation and recognition of modified sites remain unresolved. But it seems likely that RNA secondary structure is involved in both processes. METTL16 has been shown to selectivity methylate *Mat2a* in pre-mRNA hairpin structures (Mendel et al., 2018). Additionally, m^6A is most enriched in and around the highly structured transcript $3'$-UTRs (Dominissini et al., 2012; Meyer et al., 2012). The existence of cross-talk between m^6A and other RPTMs may impact upon this. The presence of m^6A antagonises A-to-I RNA editing (Xiang et al., 2018) and it would be unsurprising if other modifications like poly(A) tail terminal-uridylation influence or are influenced by m^6A methylation.
Furthermore, the issue of how dynamic m^6A methylation is remains unresolved. While many m^6A sites seem to be shared between cell-types (Schwartz et al., 2014) and it is almost certain that most m^6A is deposited co-transcriptionally by METTL3/14 (Ke et al., 2017); there still remains a strong possibility that a subset of adenosines are dynamically methylated. The main antibody-based enrichment methods for mapping m^6A, RIP-seq (Dominissini et al., 2012; Meyer et al., 2012) and CLIP seq (Ke et al., 2015; Linder et al., 2015), are unable to determine the stoichiometry of m^6A modifications. While it is true that the mRNA m^6A sites in cytoplasmic and nucleoplasmic fractions are extremely similar Ke et al. (2017); this does not confirm that the proportions of methylated sites or the specific transcript isoforms are identical in both. The translocation of normally cytoplasmic YTHDF2 into the nuclei of heat shocked mouse embryonic fibroblasts gives scope for some element of m^6A methylation or regulatory dynamism (Zhou et al., 2015). If there is any dynamic regulation of m^6A sites

it must occur in the nucleus. The m^{6}A writers METTL3 & METTL14 and ALKBH5, the physiological m^{6}A demethylase, all localise within the nucleus (Liu et al., 2013; Zheng et al., 2013).

With the recent emergence of methods for performing native RNA-seq, many of these questions should be answered in due course. The basecalling for ONT direct RNA-seq data continues to improve with refinements in the underlying models (Wick et al., 2019) and will soon approach the standard of Illumina NGS data. Nanopore direct RNA-seq generates strand-specific, long transcript sequences without any of the biases introduced by PCR amplification or RT that normally affect cDNA-based NGS experiments (Kozarewa et al., 2009). Indeed, the ONT direct RNA-seq protocol is sufficiently straight-forward that even this computational biologist, who had not touched a pipette in 5 years, was able to follow it with ease. Furthermore, I am involved in an protect that is modifying the protocol in order to detect RNA terminal modifications and quantify tail lengths. While the ONT software environment is still in its relative infancy, it is already possible to detect m^{6}A with some confidence by performing signal-level analysis of Nanopore 'squiggles' (Garalde et al., 2018; Liu et al., 2019). The ability to simultaneously interrogate poly(A)-tail lengths, the m^{6}A methylome and other RPTMs at single-nucleotide resolution holds great promise for disentangling the complex post-transcriptional network regulating RNA fates.

References

Adams, I. R. and McLaren, A. (2002). Sexually dimorphic development of mouse primordial germ cells: Switching from oogenesis to spermatogenesis. *Development*.

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M. A., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R., and Venter, J. C. (1991). Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science*.

Adey, A., Burton, J. N., Kitzman, J. O., Hiatt, J. B., Lewis, A. P., Martin, B. K., Qiu, R., Lee, C., and Shendure, J. (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*, 500(7461):207–211.

Agris, P. F. (1996). The importance of being modified: roles of modified nucleosides and Mg2+ in RNA structure and function.

Aiello, N. M. and Stanger, B. Z. (2016). Echoes of the embryo: using the developmental biology toolkit to study cancer. *Disease Models & Mechanisms*, 9(2):105–114.

Akagi, K., Li, J., Broutian, T. R., Padilla-Nash, H., Xiao, W., Jiang, B., Rocco, J. W., Teknos, T. N., Kumar, B., Wangsa, D., He, D., Ried, T., Symer, D. E., and Gillison, M. L. (2014). Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Research*, 24(2):185–199.

Alarcón, C. R., Goodarzi, H., Lee, H., Liu, X., Tavazoie, S., and Tavazoie, S. F. (2015). HNRNPA2B1 Is a Mediator of m6A-Dependent Nuclear RNA Processing Events. *Cell*.

Allen, B. L. and Taatjes, D. J. (2015). The Mediator complex: A central integrator of transcription.

Allen, F. W. (1941). The Biochemistry of the Nucleic Acids, Purines, and Pyrimidines. *Annual Review of Biochemistry*.

Allfrey, V. G., FAULKNER, R., and MIRSKY, A. E. (1964). ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE. *Proceedings of the National Academy of Sciences of the United States of*.

Altschul, S. F., Gish, W., Miller, W. T., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

Amemiya, H. M., Kundaje, A., and Boyle, A. P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*, 9(1):9354.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169.

Andrews, S. (2007). Seqmonk. a tool to visualise and analyse high throughput mapped sequence data.

Angelidis, I., Simon, L. M., Fernandez, I. E., Strunz, M., Mayr, C. H., Greiffo, F. R., Tsitsiridis, G., Ansari, M., Graf, E., Strom, T. M., Nagendran, M., Desai, T., Eickelberg, O., Mann, M., Theis, F. J., and Schiller, H. B. (2019). An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nature communications*.

Antolović, V., Miermont, A., Corrigan, A. M., and Chubb, J. R. (2017). Generation of Single-Cell Transcript Variability by Repression. *Current Biology*, 27(12):1811–1817.e3.

Avarbock, M. R., Brinster, C. J., and Brinster, R. L. (1996). Reconstitution of spermatogenesis from frozen spermatogonial stem cells. *Nat Genet.*, 2.

Baltimore, D. (1964). In vitro synthesis of viral rna by the poliovirus rna polymerase. *Proceedings of the National Academy of Sciences of the United States of America*, 51(3):450.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*.

Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications.

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6(1):4–9.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*.

Bartel, D. P. (2018). Metazoan MicroRNAs. *Cell*, 173(1):20–51.

Baudrimont, A., Jaquet, V., Wallerich, S., Voegeli, S., and Becskei, A. (2019). Contribution of RNA Degradation to Intrinsic and Extrinsic Noise in Gene Expression. *Cell Reports*.

Bedell, M. A., Hudson, J. B., Golub, T. R., Turyk, M. E., Hosken, M., Wilbanks, G. D., and Laimins, L. A. (1991). Amplification of Human Papillomavirus Genomes In Vitro Is Dependent on Epithelial Differentiation. *Journal of Virology*, 65(5):2254–2260.

Beilharz, T. H., Humphreys, D. T., Clancy, J. L., Thermann, R., Martin, D. I., Hentze, M. W., and Preiss, T. (2009). MicroRNA-mediated messenger RNA deadenylation contributes to translational repression in mammalian cells. *PLoS ONE*, 4(8).

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods*, 58(3):268–276.

Benevolenskaya, E. V. (2007). Histone H3K4 demethylases are essential in development and differentiation.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W. K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C. L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W., Ng, P., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Hoon Kim, T., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N., Yu, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., De Bakker, P. I., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K.,

Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and De Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*.

Blythe, A. J., Fox, A. H., and Bond, C. S. (2016). The ins and outs of lncRNA structure: How, why and what comes next?

Bodelon, C., Untereiner, M. E., Machiela, M. J., Vinokurova, S., and Wentzensen, N. (2016). Genomic characterization of viral integration sites in HPV-related cancers. *International Journal of Cancer*, 139(9):2001–2011.

Bokar, J., Shambaugh, M., Polayes, D., Matera, A., and Rottman, F. (1997). Purification and cdna cloning of the adomet-binding subunit of the human mrna (n6-adenosine)-methyltransferase. *Rna*, 3(11):1233–1247.

Bokar, J. A. (2005). The biosynthesis and functional roles of methylated nucleosides in eukaryotic mRNA.

Bolland, D. J., King, M. R., Reik, W., Corcoran, A. E., and Krueger, C. (2013). Robust 3d dna fish using directly labeled probes. *JoVE (Journal of Visualized Experiments)*, (78):e50587.

Bosch, F. X., Manos, M. M., Munoz, N., Sherman, M., Jansen, A. M., Peto, J., Schiffman, M. H., Moreno, V., Kurman, R., and Shah, K. V. (1995). Prevalence of Human Papillomavirus in Cervical Cancer: a Worldwide Perspective. *Journal of the National Cancer Institute*, 50(10):725–727.

Bourc'his, D. and Bestor, T. H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L.

Branco, M. R. and Pombo, A. (2006). Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biology*.

Brawerman, G. (1981). The role of the poly (a) sequence in mammalian messenger rn. *Critical Reviews in Biochemistry*, 10(1):1–38.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424.

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., and Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*.

Brenner, S., Jacob, F., and Meselson, M. (1961). An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*.

Bresson, S. and Tollervey, D. (2018). Tailing Off: PABP and CNOT Generate Cycles of mRNA Deadenylation.

Brown, S. W. (1966). Heterochromatin. *Science*, 151(3709):417–425.

Bucci, L. R. and Meistrich, M. L. (1987). Effects of busulfan on murine spermatogenesis: cytotoxicity, sterility, sperm abnormalities, and dominant lethal mutations. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 176(2):259–268.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218.

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015a). ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, 2015:21.29.1–21.29.9.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015b). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*.

Bulger, M. and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers.

Burk, R. D., Chen, Z., Saller, C., Tarvin, K., Carvalho, A. L., Scapulatempo-Neto, C., Silveira, H. C., Fregnani, J. H., Creighton, C. J., Anderson, M. L., Castro, P., Wang, S. S., Yau, C., Benz, C., Gordon Robertson, A., Mungall, K., Lim, L., Bowlby, R., Sadeghi, S., Brooks, D., Sipahimalani, P., Mar, R., Ally, A., Clarke, A., Mungall, A. J., Tam, A., Lee, D., Chuah, E., Schein, J. E., Tse, K., Kasaian, K., Ma, Y., Marra, M. A., Mayo, M., Balasundaram, M., Thiessen, N., Dhalla, N., Carlsen, R., Moore, R. A., Holt, R. A., Jones, S. J., Wong, T., Pantazi, A., Parfenov, M., Kucherlapati, R., Hadjipanayis, A., Seidman, J., Kucherlapati, M., Ren, X., Xu, A. W., Yang, L., Park, P. J., Lee, S., Rabeno, B., Huelsenbeck-Dill, L., Borowsky, M., Cadungog, M., Iacocca, M., Petrelli, N., Swanson, P., Ojesina, A. I., Ojesina, A. I., Ojesina, A. I., Le, X., Sandusky, G., Adebamowo, S. N., Akeredolu, T., Adebamowo, C., Reynolds, S. M., Shmulevich, I., Shelton, C., Crain, D., Mallery, D., Curley, E., Gardner, J., Penny, R., Morris, S., Shelton, T., Liu, J., Lolla, L., Chudamani, S., Wu, Y., Birrer, M., McLellan, M. D., Bailey, M. H., Miller, C. A., Wyczalkowski, M. A., Fulton, R. S., Fronick, C. C., Lu, C., Mardis, E. R., Appelbaum, E. L., Schmidt, H. K., Fulton, L. A., Cordes, M. G., Li, T., Ding, L., Wilson, R. K., Rader, J. S., Behmaram, B., Uyar, D., Bradley, W., Wrangle, J., Pastore, A., Levine, D. A., Dao, F., Gao, J., Schultz, N., Sander, C., Ladanyi, M., Einstein, M., Teeter, R., Benz, S., Wentzensen, N., Felau, I., Zenklusen, J. C., Bodelon, C., Demchok, J. A., Yang, L., Sheth, M., Ferguson, M. L., Tarnuzzer, R., Yang, H., Schiffman, M., Zhang, J., Wang, Z., Davidsen, T., Olaniyan, O., Hutter, C. M., Sofia, H. J., Gordenin, D. A., Chan, K., Roberts, S. A., Klimczak, L. J., Van Waes, C., Chen, Z., Saleh, A. D., Cheng, H., Parfitt, J., Bartlett, J., Albert, M., Arnaout, A., Sekhon, H., Gilbert, S., Peto, M., Myers, J., Harr, J., Eckman, J., Bergsten, J., Tucker, K., Anne Zach, L., Karlan, B. Y., Lester, J., Orsulic, S., Sun, Q., Naresh, R., Pihl, T., Wan, Y., Zaren, H., Sapp, J., Miller, J., Drwiega, P., Murray, B. A., Zhang, H., Cherniack, A. D., Sougnez, C., Sekhar Pedamallu, C., Lichtenstein, L., Meyerson, M., Noble, M. S., Heiman, D. I., Voet, D., Getz, G., Saksena, G., Kim, J., Shih, J., Cho, J., Lawrence, M. S., Gehlenborg, N., Lin, P., Beroukhim, R., Frazer, S., Gabriel, S. B., Schumacher, S. E., Leraas, K. M., Lichtenberg, T. M., Zmuda, E., Bowen, J., Frick, J., Gastier-Foster, J. M., Wise, L., Gerken, M., Ramirez, N. C., Danilova, L., Cope, L., Baylin, S. B., Salvesen, H. B., Vellano, C. P., Ju, Z., Diao, L., Zhao, H., Chong, Z., Ryan, M. C.,

Martinez-Ledesma, E., Verhaak, R. G., Averett Byers, L., Yuan, Y., Chen, K., Ling, S., Mills, G. B., Lu, Y., Akbani, R., Seth, S., Liang, H., Wang, J., Han, L., Weinstein, J. N., Bristow, C. A., Zhang, W., Mahadeshwar, H. S., Sun, H., Tang, J., Zhang, J., Song, X., Protopopov, A., Mills Shaw, K. R., Chin, L., Olabode, O., DiSaia, P., Radenbaugh, A., Haussler, D., Zhu, J., Stuart, J., Chalise, P., Koestler, D., Fridley, B. L., Godwin, A. K., Madan, R., Ciriello, G., Martinez, C., Higgins, K., Bocklage, T., Todd Auman, J., Perou, C. M., Tan, D., Parker, J. S., Hoadley, K. A., Wilkerson, M. D., Mieczkowski, P. A., Skelly, T., Veluvolu, U., Neil Hayes, D., Kimryn Rathmell, W., Hoyle, A. P., Simons, J. V., Wu, J., Mose, L. E., Soloway, M. G., Balu, S., Meng, S., Jefferys, S. R., Bodenheimer, T., Shi, Y., Roach, J., Thorne, L. B., Boice, L., Huang, M., Jones, C. D., Zuna, R., Walker, J., Gunderson, C., Snowbarger, C., Brown, D., Moxley, K., Moore, K., Andrade, K., Landrum, L., Mannel, R., McMeekin, S., Johnson, S., Nelson, T., Elishaev, E., Dhir, R., Edwards, R., Bhargava, R., Tiezzi, D. G., Andrade, J. M., Noushmehr, H., Gilberto Carlotti, C., da Cunha Tirapelli, D. P., Weisenberger, D. J., Van Den Berg, D. J., Maglinte, D. T., Bootwalla, M. S., Lai, P. H., Triche, T., Swisher, E. M., Agnew, K. J., Simon Shelley, C., Laird, P. W., Schwarz, J., Grigsby, P., and Mutch, D. (2017). Integrated genomic and molecular characterization of cervical cancer. *Nature*, 543(7645):378–384.

Cairns, B. R. (2009). The logic of chromatin architecture and remodelling at promoters.

Carlile, T. M., Rojas-Duran, M. F., Zinshteyn, B., Shin, H., Bartoli, K. M., and Gilbert, W. V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature*.

Carmell, M. A., Girard, A., van de Kant, H. J., Bourc'his, D., Bestor, T. H., de Rooij, D. G., and Hannon, G. J. (2007). MIWI2 Is Essential for Spermatogenesis and Repression of Transposons in the Mouse Male Germline. *Developmental Cell*, 12(4):503–514.

Carrieri, C., Comazzetto, S., Grover, A., Morgan, M., Buness, A., Nerlov, C., and O'Carroll, D. (2017). A transit-amplifying population underpins the efficient regenerative capacity of the testis. *The Journal of Experimental Medicine*, 214(6):1631–1641.

Caspersson, T. (1947). The relations between nucleic acid and protein synthesis. In *Symposia of the Society for Experimental Biology*, number 1, pages 127–151.

Caspersson, T. and Schultz, J. (1939). Pentose nucleotides in the cytoplasm of growing tissues.

Chan, H. M. and La Thangue, N. B. (2001). p300/CBP proteins: HATs for transcriptional bridges and scaffolds. *Journal of Cell Science*.

Chang, H., Lim, J., Ha, M., and Kim, V. N. (2014). TAIL-seq: Genome-wide determination of poly(A) tail length and 3' end modifications. *Molecular Cell*, 53(6):1044–1052.

Chen, C. Y. A. and Shyu, A. B. (2011). Mechanisms of deadenylation-dependent decay.

Chen, C. Y. A., Zheng, D., Xia, Z., and Shyu, A. B. (2009). Ago-TNRC6 triggers microRNA-mediated decay by promoting two deadenylation steps. *Nature Structural and Molecular Biology*, 16(11):1160–1166.

Chen, G., Lustig, A., and Weng, N.-p. (2013). T cell aging: a review of the transcriptional changes determined from genome-wide analysis. *Frontiers in Immunology*, 4:121.

Chen, H.-I. H., Jin, Y., Huang, Y., and Chen, Y. (2016). Detection of high variability in gene expression from single-cell rna-seq profiling. *BMC genomics*, 17(7):508.

Chen, J., Sun, Y., Xu, X., Wang, D., He, J., Zhou, H., Lu, Y., Zeng, J., Du, F., Gong, A., and Xu, M. (2017). YTH domain family 2 orchestrates epithelial-mesenchymal transition/proliferation dichotomy in pancreatic cancer cells. *Cell Cycle*, 16(23):2259–2271.

Chen, M., Wei, L., Law, C. T., Tsang, F. H. C., Shen, J., Cheng, C. L. H., Tsang, L. H., Ho, D. W. H., Chiu, D. K. C., Lee, J. M. F., Wong, C. C. L., Ng, I. O. L., and Wong, C. M. (2018). RNA N6-methyladenosine methyltransferase-like 3 promotes liver cancer progression through YTHDF2-dependent posttranscriptional silencing of SOCS2. *Hepatology*, 67(6):2254–2270.

Chen, T., Hao, Y. J., Zhang, Y., Li, M. M., Wang, M., Han, W., Wu, Y., Lv, Y., Hao, J., Wang, L., Li, A., Yang, Y., Jin, K. X., Zhao, X., Li, Y., Ping, X. L., Lai, W. Y., Wu, L. G., Jiang, G., Wang, H. L., Sang, L., Wang, X. J., Yang, Y. G., and Zhou, Q. (2015). M6A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. *Cell Stem Cell*, 16(3):289–301.

Chesteron, C. J., Coupar, B. E., and Butterworth, P. H. (1974). Transcription of fractionated mammalian chromatin by mammalian ribonucleic acid polymerase. Demonstration of temperature dependent rifampicin resistant initiation sites in euchromatin deoxyribonucleic acid. *Biochemical Journal*.

Christiansen, I. K., Sandve, G. K., Schmitz, M., Dürst, M., and Hovig, E. (2015). Transcriptionally active regions are the preferred targets for chromosomal HPV integration in cervical carcinogenesis. *PLoS ONE*, 10(3):1–11.

Ciccarone, F., Tagliatesta, S., Caiafa, P., and Zampieri, M. (2018). DNA methylation dynamics in aging: how far are we from understanding the mechanisms?

Clerici, M., Faini, M., Muckenfuss, L. M., Aebersold, R., and Jinek, M. (2018). Structural basis of AAUAAA polyadenylation signal recognition by the human CPSF complex. *Nature Structural and Molecular Biology*.

Clermont, Y. (1966). Renewal of spermatogonia in man. *American Journal of Anatomy*.

Cohn, W. E. (1960). Pseudouridine, a carbon-carbon linked ribonucleoside in ribonucleic acids: isolation, structure, and chemical characteristics. *The Journal of biological chemistry*.

Cooper, T. F. (2007). Recombination speeds adaptation by reducing competition between beneficial mutations in populations of Escherichia coli. *PLoS Biology*, 5(9):1899–1905.

Coots, R. A., Liu, X. M., Mao, Y., Dong, L., Zhou, J., Wan, J., Zhang, X., and Qian, S. B. (2017). m6A Facilitates eIF4F-Independent mRNA Translation. *Molecular Cell*, 68(3):504–514.e7.

Cramer, P., Armache, K.-J., Baumli, S., Benkert, S., Brueckner, F., Buchen, C., Damsma, G., Dengl, S., Geiger, S., Jasiak, A., Jawhari, A., Jennebach, S., Kamenski, T., Kettenberger, H., Kuhn, C.-D., Lehmann, E., Leike, K., Sydow, J., and Vannini, A. (2008). Structure of Eukaryotic RNA Polymerases. *Annual Review of Biophysics*.

Crane, E., Bian, Q., McCord, R. P., Lajoie, B. R., Wheeler, B. S., Ralston, E. J., Uzawa, S., Dekker, J., and Meyer, B. J. (2015). Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, 523(7559):240–244.

Cremer, M., Hase, J. V., Volm, T., Brero, A., Kreth, G., Walter, J., Fischer, C., Solovei, I., Cremer, C., and Cremer, T. (2001). Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells. *Chromosome Research*.

Cremer, T. and Cremer, C. (2006). Rise, fall and resurrection of chromosome territories: a historical perspective part ii. fall and resurrection of chromosome territories during the 1950s to 1980s. part iii. chromosome territories and the functional nuclear architecture: experiments and m. *European journal of histochemistry*, pages 223–272.

Cremer, T., Cremer, C., Schneider, T., Baumann, H., Hens, L., and Kirsch-Volders, M. (1982). Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments. *Human Genetics*.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., and Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*.

Crick, F., Barnett, L., Brenner, S., and Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature*.

Crick, F. H. (1958). On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8.

Crick, F. H. (1968). The origin of the genetic code. *Journal of molecular biology*, 38(3):367–379.

Croft, J. A., Bridger, J. M., Boyle, S., Perry, P., Teague, P., and Bickmore, W. A. (1999). Differences in the localization and morphology of chromosomes in the human nucleus. *Journal of Cell Biology*.

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., Bennett, R., Bhai, J., Billis, K., Boddu, S., Cummins, C., Davidson, C., Dodiya, K. J., Gall, A., Girón, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., Kay, M., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Marugán, J. C., Maurel, T., McMahon, A. C., Moore, B., Morales, J., Mudge, J. M., Nuhn, M., Ogeh, D., Parker, A., Parton, A., Patricio, M., Abdul Salam, A. I., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Sparrow, H., Stapleton, E., Szuba, M., Taylor, K., Threadgold, G., Thormann, A., Vullo, A., Walts, B., Winterbottom, A., Zadissa, A., Chakiachvili, M., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Yates, A. D., Zerbino, D. R., and Flicek, P. (2019). Ensembl 2019. *Nucleic Acids Research*, 47(D1):D745–D751.

Custódio, N. and Carmo-Fonseca, M. (2016). Co-transcriptional splicing and the CTD code.

Dahm, R. (2005). Friedrich Miescher and the discovery of DNA.

Dall, K. L., Scarpini, C. G., Roberts, I., Winder, D. M., Stanley, M. A., Muralidhar, B., Herdman, M. T., Pett, M. R., and Coleman, N. (2008). Characterization of naturally occurring HPV16 integration sites isolated from cervical keratinocytes under noncompetitive conditions. *Cancer Research*, 68(20):8249–8259.

Darnell, J. E., Philipson, L., Wall, R., and Adesnik, M. (1971). Polyadenylic acid sequences: Role in conversion of nuclear RNA into messenger RNA. *Science*.

Davila Lopez, M. and Samuelsson, T. (2007). Early evolution of histone mRNA 3' end processing. *RNA*.

Davis, M. P., van Dongen, S., Abreu-Goodger, C., Bartonicek, N., and Enright, A. J. (2013). Kraken: A set of tools for quality control and analysis of high-throughput sequence data. *Methods*, 63(1):41–49.

De Cecco, M., Criscione, S. W., Peterson, A. L., Neretti, N., Sedivy, J. M., and Kreiling, J. A. (2013). Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. *Aging*.

de Magalhães, J. P., Curado, J., and Church, G. M. (2009). Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*.

De Rooij, D. G. and Griswold, M. D. (2012). Questions about spermatogonia posed and answered since 2000. *Journal of Andrology*, 33(6):1085–1095.

de Sanjosé, S., Serrano, B., Tous, S., Alejo, M., Lloveras, B., Quirós, B., Clavero, O., Vidal, A., Ferrándiz-Pulido, C., Pavón, M. Á., Holzinger, D., Halec, G., Tommasino, M., Quint, W., Pawlita, M., Muñoz, N., Bosch, F. X., Alemany, L., and Kulkarni, A. (2018). Burden of Human Papillomavirus (HPV)-Related Cancers Attributable to HPVs 6/11/16/18/31/33/45/52 and 58. *JNCI Cancer Spectrum*, 2(4):1–11.

de Wit, E. and de Laat, W. (2012). A decade of 3C technologies: Insights into nuclear organization. *Genes and Development*.

Deaton, A. M. and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes and Development*.

Denny, L. A., Franceschi, S., de Sanjosé, S., Heard, I., Moscicki, A. B., and Palefsky, J. (2012). Human papillomavirus, human immunodeficiency virus and immunosuppression. *Vaccine*, 30(SUPPL.5):F168–F174.

Desrosiers, R., Friderici, K., and Rottman, F. (1974). Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. *Proceedings of the National Academy of Sciences of the United States of America*.

Desterro, J., Bak-Gordon, P., and Carmo-Fonseca, M. (2019). Targeting mrna processing as an anticancer strategy. *Nature Reviews Drug Discovery*, pages 1–18.

Dey, A., Seshasayee, D., Noubade, R., French, D. M., Liu, J., Chaurushiya, M. S., Kirkpatrick, D. S., Pham, V. C., Lill, J. R., Bakalarski, C. E., Wu, J., Phu, L., Katavolos, P., LaFave, L. M., Abdel-Wahab, O., Modrusan, Z., Seshagiri, S., Dong, K., Lin, Z., Balazs, M., Suriben, R., Newton, K., Hymowitz, S., Garcia-Manero, G., Martin, F., Levine, R. L., and Dixit, V. M. (2012). Loss of the tumor suppressor BAP1 causes myeloid transformation. *Science*.

Dieci, G., Fiorino, G., Castelnuovo, M., Teichmann, M., and Pagano, A. (2007). The expanding RNA polymerase III transcriptome.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

Dohle, G. R. (2010). Male infertility in cancer patients: Review of the literature. *International journal of urology*, 17(4):327–331.

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M., Sorek, R., and Rechavi, G. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, 485(7397):201–206.

Dominissini, D., Nachtergaele, S., Moshitch-Moshkovitz, S., Peer, E., Kol, N., Ben-Haim, M. S., Dai, Q., Di Segni, A., Salmon-Divon, M., Clark, W. C., Zheng, G., Pan, T., Solomon, O., Eyal, E., Hershkovitz, V., Han, D., Doré, L. C., Amariglio, N., Rechavi, G., and He, C. (2016). The dynamic N1 -methyladenosine methylome in eukaryotic messenger RNA. *Nature*.

Dominski, Z. and Marzluff, W. F. (1999). Formation of the 3' end of histone mRNA.

Dong, D., Shao, X., Deng, N., and Zhang, Z. (2011). Gene expression variations are predictive for stochastic noise. *Nucleic Acids Research*.

Dorigo, B., Schalch, T., Bystricky, K., and Richmond, T. J. (2003). Chromatin fiber folding: Requirement for the histone H4 N-terminal tail. *Journal of Molecular Biology*.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., and Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*.

Drolet, M., Benard, E., Perez, N., and Brisson, M. (2019). Population-level impact and herd effects following the introduction of human papillomavirus vaccination programmes: updated systematic review and meta-analysis. *The Lancet*.

Dryden, N. H., Broome, L. R., Dudbridge, F., Johnson, N., Orr, N., Schoenfelder, S., Nagano, T., Andrews, S., Wingett, S., Kozarewa, I., Assiotis, I., Fenwick, K., Maguire, S. L., Campbell, J., Natrajan, R., Lambros, M., Perrakis, E., Ashworth, A., Fraser, P., and Fletcher, O. (2014). Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Research*, 24(11):1854–1868.

Du, H., Zhao, Y., He, J., Zhang, Y., Xi, H., Liu, M., Ma, J., and Wu, L. (2016). YTHDF2 destabilizes m 6 A-containing RNA through direct recruitment of the CCR4-NOT deadenylase complex. *Nature Communications*, 7:1–11.

Dueck, H., Eberwine, J., and Kim, J. (2016). Variation is function: Are single cell differences functionally important?: Testing the hypothesis that single cell variation is required for aggregate function. *BioEssays*, 38(2):172–180.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kellis, M., Kheradpour, P., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Bernstein, B. E., Green, E. D., Gunter, C., Snyder, M., Pazin, M. J., Lowdon, R. F., Dillon, L. A., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Good, P. J., Feingold, E. A., Crawford, G. E., Dekker, J., Elnitski, L., Farnham, P. J., Giddings, M. C., Gingeras, T. R., Guigó, R., Hubbard, T. J., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Stamatoyannopoulos, J. A., Tenenbaum, S. A., Weng, Z., White, K. P., Wold, B., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Eaton, M. L., Dobin, A., Tanzer, A., Lagarde, J., Lin, W., Xue, C., Williams, B. A., Zaleski, C., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Robyr, D., Ruan, X., Sammeth, M., Sandhu, K. S., Schaeffer, L., See, L. H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Hayashizaki, Y., Reymond, A., Antonarakis, S. E., Hannon, G. J., Ruan, Y., Carninci, P., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Grasfeder, L. L., Giresi, P. G., Battenhouse, A., Sheffield, N. C., Showers, K. A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Kim, S. K., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J., Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe, D., Iyer, V. R., Zheng, M., Wang, P., Gertz, J., Vielmetter, J., Partridge, E. C., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., Muratet, M. A., Newberry, K. M., McCue, K., Nesmith, A. S., Fisher-Aylor, K. I., Pusey, B., DeSalvo, G., Parker, S. L., Balasubramanian, S., Davis, N. S., Meadows, S. K.,

Eggleston, T., Newberry, J. S., Levy, S. E., Absher, D. M., Wong, W. H., Blow, M. J., Visel, A., Pennachio, L. A., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Davidson, C., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J. M., Griffiths, E., Harte, R., Hendrix, D. A., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Leng, J., Lin, M. F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J. M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tapanari, E., Tress, M. L., Van Baren, M. J., Washietl, S., Wilming, L., Zadissa, A., Zhang, Z., Brent, M., Haussler, D., Valencia, A., Addleman, N., Alexander, R. P., Auerbach, R. K., Balasubramanian, S., Bettinger, K., Bhardwaj, N., Boyle, A. P., Cao, A. R., Cayting, P., Charos, A., Cheng, Y., Eastman, C., Euskirchen, G., Fleming, J. D., Grubert, F., Habegger, L., Hariharan, M., Harmanci, A., Iyengar, S., Jin, V. X., Karczewski, K. J., Kasowski, M., Lacroute, P., Lam, H., Lamarre-Vincent, N., Lian, J., Lindahl-Allen, M., Min, R., Miotto, B., Monahan, H., Moqtaderi, Z., Mu, X. J., O'Geen, H., Ouyang, Z., Patacsil, D., Raha, D., Ramirez, L., Reed, B., Shi, M., Slifer, T., Witt, H., Wu, L., Xu, X., Yan, K. K., Yang, X., Struhl, K., Weissman, S. M., Penalva, L. O., Karmakar, S., Bhanvadia, R. R., Choudhury, A., Domanus, M., Ma, L., Moran, J., Victorsen, A., Auer, T., Centanin, L., Eichenlaub, M., Gruhl, F., Heermann, S., Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R., Schertel, L., Schneider, S., Sinn, R., Wittbrodt, B., Wittbrodt, J., Jain, G., Balasundaram, G., Bates, D. L., Byron, R., Canfield, T. K., Diegel, M. J., Dunn, D., Ebersol, A. K., Frum, T., Garg, K., Gist, E., Hansen, R. S., Boatman, L., Haugen, E., Humbert, R., Johnson, A. K., Johnson, E. M., Kutyavin, T. V., Lee, K., Lotakis, D., Maurano, M. T., Neph, S. J., Neri, F. V., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Rynes, E., Sanchez, M. E., Sandstrom, R. S., Shafer, A. O., Stergachis, A. B., Thomas, S., Vernot, B., Vierstra, J., Vong, S., Wang, H., Weaver, M. A., Yan, Y., Zhang, M., Akey, J. M., Bender, M., Dorschner, M. O., Groudine, M., MacCoss, M. J., Navas, P., Stamatoyannopoulos, G., Beal, K., Brazma, A., Flicek, P., Johnson, N., Lukk, M., Luscombe, N. M., Sobral, D., Vaquerizas, J. M., Batzoglou, S., Sidow, A., Hussami, N., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M. W., Schaub, M. A., Miller, W., Bickel, P. J., Banfai, B., Boley, N. P., Huang, H., Li, J. J., Noble, W. S., Bilmes, J. A., Buske, O. J., Sahu, A. D., Kharchenko, P. V., Park, P. J., Baker, D., Taylor, J., and Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*.

Dunn, D. (1960). The isolation of 5-methylcytidine from rna. *Biochimica et biophysica acta*, 38:176–178.

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., and Aiden, E. L. (2016a). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems*, 3(1):99–101.

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., and Aiden, E. L. (2016b). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems*, 3(1):95–98.

Durst, M., Croce, C. M., Gissmann, L., Schwarz, E., and Huebner, K. (1987). Papillomavirus sequences integrate near cellular oncogenes in some cervical carcinomas. *Proceedings of the National Academy of Sciences*, 84(4):1070–1074.

Dürst, M., Glitz, D., Schneider, A., and zur Hausen, H. (1992). Human papillomavirus type 16 (HPV 16) gene expression and DNA replication in cervical neoplasia: Analysis by in situ hybridization. *Virology*, 189(1):132–140.

Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R., and Hood, L. (1980). Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. *Cell*.

Ecker, S., Chen, L., Pancaldi, V., Bagger, F. O., Fernández, J. M., Carrillo de Santa Pau, E., Juan, D., Mann, A. L., Watt, S., Casale, F. P., Sidiropoulos, N., Rapin, N., Merkel, A., Stunnenberg, H. G., Stegle, O., Frontini, M., Downes, K., Kuijpers, T. W., Kuijpers, T. W., Rico, D., Valencia, A., Beck, S., Soranzo, N., Paul, D. S., Albers, C. A., Amstislavskiy, V., Ashford, S., Bomba, L., Bujold, D., Burden, F., Busche, S., Caron, M., Chen, S. H., Cheung, W. A., Clarke, L., Colgiu, I., Datta, A., Delaneau, O., Elding, H., Farrow, S., Garrido-Martín, D., Ge, B., Guigo, R., Iotchkova, V., Kundu, K., Kwan, T., Lambourne, J. J., Lowy, E., Mead, D., Pourfarzad, F., Redensek, A., Rehnstrom, K., Rendon, A., Richardson, D., Risch, T., Rowlston, S., Shao, X., Simon, M. M., Sultan, M., Walter, K., Wilder, S. P., Yan, Y., Antonarakis, S. E., Bourque, G., Dermitzakis, E. T., Flicek, P., Lehrach, H., Martens, J. H., Yaspo, M. L., and Ouwehand, W. H. (2017). Genome-wide analysis of differential transcriptional and epigenetic variability across human immune cell types. *Genome Biology*, 18(1).

Eckmann, C. R., Rammelt, C., and Wahle, E. (2011). Control of poly(A) tail length.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461.

Edmonds, M., Vaughan, M. H., and Nakazato, H. (1971). Polyadenylic acid sequences in the heterogeneous nuclear RNA and rapidly-labeled polyribosomal RNA of HeLa cells: possible evidence for a precursor relationship. *Proceedings of the National Academy of Sciences of the United States of America*.

Edupuganti, R. R., Geiger, S., Lindeboom, R. G., Shi, H., Hsu, P. J., Lu, Z., Wang, S. Y., Baltissen, M. P., Jansen, P. W., Rossa, M., Müller, M., Stunnenberg, H. G., He, C., Carell, T., and Vermeulen, M. (2017). N6-methyladenosine (m6A) recruits and repels proteins to regulate mRNA homeostasis. *Nature Structural and Molecular Biology*.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Single Polymerase Molecules. *Science (New York, N.Y.)*.

Eling, N., Richard, A. C., Richardson, S., Marioni, J. C., and Vallejos, C. A. (2018). Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data. *Cell Systems*, 7(3):284–294.e12.

Elowitz, M. B., Levine, A. J., Siggia, E. D., Swain, P. S., Guptasarma, P., Spudich, J. L., McAdams, H. H., Heitzler, P., Ko, M. S., Fiering, S., Lutz, R., Deuschle, U., Maloney, P. C., Rotman, B., Paulsson, J., Ehrenberg, M., Boyd, D., Becskei, A., Serrano, L., Elowitz, M. B., Leibler, S., Thattai, M., van Oudenaarden, A., Alon, U., Capaldo, F. N., Barbour, S. D., Casadaban, M. J., Parkinson, J. S., Houts, S. E., Meyer, B. J., Maurer, R., and Ptashne, M. (2002). Stochastic gene expression in a single cell. *Science (New York, N.Y.)*.

Encode Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Enge, M., Arda, H. E., Mignardi, M., Beausang, J., Bottino, R., Kim, S. K., and Quake, S. R. (2017). Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell*.

Ernst, C., Odom, D. T., and Kutter, C. (2017). The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. *Nature Communications*, 8(1):1–9.

Eser, U., Chandler-Brown, D., Ay, F., Straight, A. F., Duan, Z., Noble, W. S., and Skotheim, J. M. (2017). Form and function of topologically associating genomic domains in budding yeast. *Proceedings of the National Academy of Sciences of the United States of America*.

Esteller, M. (2002). CpG island hypermethylation and tumor suppressor genes: A booming present, a brighter future.

Faure, A. J., Schmiedel, J. M., and Lehner, B. (2017). Systematic Analysis of the Determinants of Gene Expression Noise in Embryonic Stem Cells. *Cell Systems*.

Favre, M., Breitburd, F., Croissant, O., and Orth, G. (1977). Chromatin-like structures obtained after alkaline disruption of bovine and human papillomaviruses. *Journal of Virology*, 21(3):1205–1209.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer*.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G., and Ysebaert, M. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene. *Nature*.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans. *Nature*.

Fraga, M. F. and Esteller, M. (2007). Epigenetics and aging: the targets and the marks.

Franks, L. M. and Payne, J. (1970). The influence of age on reproductive capacity in C57BL mice. *Journal of reproduction and fertility*.

Frye, M., Jaffrey, S. R., Pan, T., Rechavi, G., and Suzuki, T. (2016). RNA modifications: What have we learned and where are we headed?

Fukaya, T., Lim, B., and Levine, M. (2016). Enhancer Control of Transcriptional Bursting. *Cell*.

Fuke, H. and Ohno, M. (2008). Role of poly (A) tail as an identity element for mRNA nuclear export. *Nucleic Acids Research*, 36(3):1037–1049.

Gambardella, G., Carissimo, A., Chen, A., Cutillo, L., Nowakowski, T. J., Di Bernardo, D., and Blelloch, R. (2017). The impact of microRNAs on transcriptional heterogeneity and gene co-expression across single embryonic stem cells. *Nature Communications*, 8(May 2016):1–11.

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., Jordan, M., Ciccone, J., Serra, S., Keenan, J., Martin, S., McNeill, L., Wallace, E. J., Jayasinghe, L., Wright, C., Blasco, J., Young, S., Brocklebank, D., Juul, S., Clarke, J., Heron, A. J., and Turner, D. J. (2018). Highly parallel direct RN A sequencing on an array of nanopores. *Nature Methods*, 15(3):201–206.

Geisberg, J. V., Moqtaderi, Z., Fan, X., Ozsolak, F., and Struhl, K. (2014). Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell*, 156(4):812–824.

Gems, D. and Partridge, L. (2013). Genetics of Longevity in Model Organisms: Debates and Paradigm Shifts. *Annual Review of Physiology*, 75(1):621–644.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition.

Geula, S., Moshitch-Moshkovitz, S., Dominissini, D., Mansour, A. A. F., Kol, N., Salmon-Divon, M., Hershkovitz, V., Peer, E., Mor, N., Manor, Y. S., Ben-Haim, M. S., Eyal, E., Yunger, S., Pinto, Y., Jaitin, D. A., Viukov, S., Rais, Y., Krupalnik, V., Chomsky, E., Zerbib, M., Maza, I., Rechavi, Y., Massarwa, R., Hanna, S., Amit, I., Levanon, E. Y., Amariglio, N., Stern-Ginossar, N., Novershtern, N., Rechavi, G., and Hanna, J. H. (2015). m6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science*.

Goodenough, U. and Heitman, J. (2014). Origins of eukaryotic sexual reproduction. *Cold Spring Harbor Perspectives in Biology*, 6(3).

Goolam, M., Scialdone, A., Graham, S. J., MacAulay, I. C., Jedrusik, A., Hupalowska, A., Voet, T., Marioni, J. C., and Zernicka-Goetz, M. (2016). Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell*.

Gorbunova, V., Seluanov, A., Mao, Z., and Hine, C. (2007). Changes in DNA repair during aging.

Goriely, A., McGrath, J. J., Hultman, C. M., Wilkie, A. O., and Malaspina, D. (2013). "Selfish spermatogonial selection": A novel mechanism for the association between advanced paternal age and neurodevelopmental disorders. *American Journal of Psychiatry*, 170(6):599–608.

Gott, J. M. and Emeson, R. B. (2000). Functions and Mechanisms of RNA Editing. *Annual Review of Genetics*.

Gray, E., Pett, M. R., Ward, D., Winder, D. M., Stanley, M. A., Roberts, I., Scarpini, C. G., and Coleman, N. (2010). In vitro progression of human papillomavirus 16 episome-associated cervical neoplasia displays fundamental similarities to integrant-associated carcinogenesis. *Cancer Research*, 70(10):4081–4091.

Griswold, M. D. (2016). Spermatogenesis: The commitment to Meiosis. *Physiological Reviews*, 96(1):1–17.

Groves, I. J. and Coleman, N. (2015). Pathogenesis of human papillomavirus-associated mucosal disease. *Journal of Pathology*.

Groves, I. J. and Coleman, N. (2018). Human papillomavirus genome integration in squamous carcinogenesis: what have next-generation sequencing studies taught us? *Journal of Pathology*, 245(1):9–18.

Groves, I. J., Knight, E. L. A., Ang, Q. Y., Scarpini, C. G., and Coleman, N. (2016). HPV16 oncogene expression levels during early cervical carcinogenesis are determined by the balance of epigenetic chromatin modifications at the integrated virus genome. *Oncogene*.

Grozhik, A. V. and Jaffrey, S. R. (2018). Distinguishing RNA modifications from noise in epitranscriptome maps.

Grozhik, A. V., Linder, B., Olarerin-George, A. O., and Jaffrey, S. R. (2017). Mapping m6A at individual-nucleotide resolution using crosslinking and immunoprecipitation (MiCLIP). In *Methods in Molecular Biology*. Springer.

Grozhik, A. V., Olarerin-George, A. O., Sindelar, M., Li, X., Gross, S. S., and Jaffrey, S. R. (2019). Antibody cross-reactivity accounts for widespread appearance of m1a in 5?utrs. *bioRxiv*, page 648345.

Grün, D., Kester, L., and Van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nature Methods*.

Gulati, G. S., Sikandar, S. S., Wesche, D. J., Manjunath, A., Bharadwaj, A., Berger, M. J., Ilagan, F., Kuo, A. H., Hsieh, R. W., Cai, S., Zabala, M., Scheeren, F. A., Lobo, N. A., Qian, D., Yu, F. B., Dirbas, F. M., Clarke, M. F., and Newman, A. M. (2019). Single-cell transcriptional diversity is a hallmark of developmental potential. *bioRxiv*, page 649848.

Gunderson, S. I., Polycarpou-Schwarz, M., and Mattaj, I. W. (1998). U1 snRNP inhibits pre-mRNA polyadenylation through a direct interaction between U1 70K and poly(A) polymerase. *Molecular Cell*.

Guo, J., Garrett, M., Micklem, G., and Brogna, S. (2011). Poly(A) Signals Located near the 5' End of Genes Are Silenced by a General Mechanism That Prevents Premature 3'-End Processing. *Molecular and Cellular Biology*.

Haas, B. J., Dobin, A., Stransky, N., Li, B., Yang, X., Tickle, T., Bankapur, A., Ganote, C., Doak, T. G., Pochet, N., Sun, J., Wu, C. J., Gingeras, T. R., and Regev, A. (2017). STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv*, page 120295.

Haberle, V. and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation.

Haddad, R., Maurice, F., Viphakone, N., Voisinet-Hakil, F., Fribourg, S., and Minvielle-Sébastia, L. (2012). An essential role for Clp1 in assembly of polyadenylation complex CF IA and Pol II transcription termination. *Nucleic Acids Research*, 40(3):1226–1239.

Han, S. and Brunet, A. (2012). Histone methylation makes its mark on longevity.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74.

Hansen, K. D., Wu, Z., Irizarry, R. A., and Leek, J. T. (2011). Sequencing technology does not eliminate biological variability.

Harper, C. V., Finkenstädt, B., Woodcock, D. J., Friedrichsen, S., Semprini, S., Ashall, L., Spiller, D. G., Mullins, J. J., Rand, D. A., Davis, J. R. E., and White, M. R. H. (2011). Dynamic analysis of stochastic transcription cycles. *PLoS Biology*.

Harris, C. L., Hanna, S. M., Mizuno, M., Holt, D. S., Marchbank, K. J., and Morgan, B. P. (2003). Characterization of the mouse analogues of CD59 using novel monoclonal antibodies: Tissue distribution and functional comparison. *Immunology*.

Hassan, M. A. and Killick, S. R. (2003). Effect of male age on fertility: Evidence for the decline in male fertility with increasing age. *Fertility and Sterility*, 79(SUPPL. 3):1520–1527.

Hawley-Nelson, P., Vousden, K. H., Hubbert, N. L., Lowy, D. R., and Schiller, J. T. (1989). HPV16 E6 and E7 proteins cooperate to immortalize human foreskin keratinocytes. *The EMBO journal*, 8(12):3905–10.

He, Z., Jiang, J., Kokkinaki, M., Golestaneh, N., Hofmann, M.-C., and Dym, M. (2008). Gdnf Upregulates c-Fos Transcription via the Ras/Erk1/2 Pathway to Promote Mouse Spermatogonial Stem Cell Proliferation. *Stem Cells*.

Heitz, E. (1928). Das Heterochromatin der Moose. *Jahrbücher für wissenschaftliche Botanik*.

Heyne, S., Costa, F., Rose, D., and Backofen, R. (2012). Graphclust: Alignment-free structural clustering of local RNA secondary structures. *Bioinformatics*, 28(12).

Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K., and Sharp, P. A. (2017). A Phase Separation Model for Transcriptional Control. *Cell*, 169(1):13–23.

Ho, J. W., Stefani, M., Dos Remedios, C. G., and Charleston, M. A. (2008). Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics*.

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., and Zamir, A. (1965). Structure of a ribonucleic acid. *Science*.

Holliday, R. (1964). A mechanism for gene conversion in fungi. *Genetical Research*.

Holmes, A., Lameiras, S., Jeannot, E., Marie, Y., Castera, L., Sastre-Garau, X., and Nicolas, A. (2016). Mechanistic signatures of HPV insertions in cervical Carcinomas. *npj Genomic Medicine*, 1(January).

Hong, S. and Laimins, L. A. (2013). Regulation of the life cycle of HPVs by differentiation and the DNA damage response. *Future Microbiology*, 8(12):1547–1557.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*.

Hsu, Y. C., Li, L., and Fuchs, E. (2014). Transit-amplifying cells orchestrate stem cell activity and tissue regeneration. *Cell*, 157(4):935–949.

Hu, Z., Zhu, D., Wang, W., Li, W., Jia, W., Zeng, X., Ding, W., Yu, L., Wang, X., Wang, L., Shen, H., Zhang, C., Liu, H., Liu, X., Zhao, Y., Fang, X., Li, S., Chen, W., Tang, T., Fu, A., Wang, Z., Chen, G., Gao, Q., Li, S., Xi, L., Wang, C., Liao, S., Ma, X., Wu, P., Li, K., Wang, S., Zhou, J., Wang, J., Xu, X., Wang, H., and Ma, D. (2015). Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nature Genetics*, 47(2):158–163.

Huang, H., Weng, H., Sun, W., Qin, X., Shi, H., Wu, H., Zhao, B. S., Mesquita, A., Liu, C., Yuan, C. L., Hu, Y. C., Hüttelmaier, S., Skibbe, J. R., Su, R., Deng, X., Dong, L., Sun, M., Li, C., Nachtergaele, S., Wang, Y., Hu, C., Ferchen, K., Greis, K. D., Jiang, X., Wei, M., Qu, L., Guan, J. L., He, C., Yang, J., and Chen, J. (2018). Recognition of RNA N 6 -methyladenosine by IGF2BP proteins enhances mRNA stability and translation. *Nature Cell Biology*.

Huang, H., Weng, H., Zhou, K., Wu, T., Zhao, B. S., Sun, M., Chen, Z., Deng, X., Xiao, G., Auer, F., Klemm, L., Wu, H., Zuo, Z., Qin, X., Dong, Y., Zhou, Y., Qin, H., Tao, S., Du, J., Liu, J., Lu, Z., Yin, H., Mesquita, A., Yuan, C. L., Hu, Y. C., Sun, W., Su, R., Dong, L., Shen, C., Li, C., Qing, Y., Jiang, X., Wu, X., Sun, M., Guan, J. L., Qu, L., Wei, M., Müschen, M., Huang, G., He, C., Yang, J., and Chen, J. (2019). Histone H3 trimethylation at lysine 36 guides m6A RNA modification co-transcriptionally.

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., and Teichmann, S. A. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*.

Ivanova, I., Much, C., Di Giacomo, M., Azzi, C., Morgan, M., Moreira, P. N., Monahan, J., Carrieri, C., Enright, A. J., and O'Carroll, D. (2017). The RNA m6A Reader YTHDF2 Is Essential for the Post-transcriptional Regulation of the Maternal Transcriptome and Oocyte Competence. *Molecular Cell*, 67(6):1059–1067.e4.

Jacob, F. and Monod, J. (1961a). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3(3):318–356.

Jacob, F. and Monod, J. (1961b). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–356.

Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., and Akeson, M. (2015). Improved data analysis for the minion nanopore sequencer. *Nature methods*, 12(4):351.

Javierre, B. M., Sewitz, S., Cairns, J., Wingett, S. W., Várnai, C., Thiecke, M. J., Freire-Pritchett, P., Spivakov, M., Fraser, P., Burren, O. S., Cutler, A. J., Todd, J. A., Wallace, C., Wilder, S. P., Kreuzhuber, R., Kostadima, M., Zerbino, D. R., Stegle, O., Kreuzhuber, R., Burden, F., Farrow, S., Rehnström, K., Downes, K., Grassi, L., Kostadima, M., Ouwehand, W. H., Frontini, M., Kreuzhuber, R., Burden, F., Farrow, S., Rehnström, K., Downes, K., Kostadima, M., Ouwehand, W. H., Frontini, M., Hill, S. M., Wang, F., Stunnenberg, H. G., Ouwehand, W. H., Frontini, M., Ouwehand, W. H., Martens, J. H., Kim, B., Sharifi, N., Janssen-Megens, E. M., Yaspo, M. L., Linser, M., Kovacsovics, A., Clarke, L., Richardson, D., Datta, A., and Flicek, P. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, 167(5):1369–1384.e19.

Jia, G., Fu, Y., Zhao, X., Dai, Q., Zheng, G., Yang, Y., Yi, C., Lindahl, T., Pan, T., Yang, Y. G., and He, C. (2011). N6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nature Chemical Biology*.

Jones, P. L., Veenstra, G. J. C., Wade, P. A., Vermaak, D., Kass, S. U., Landsberger, N., Strouboulis, J., and Wolffe, A. P. (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nature Genetics*.

Jung, M. and Pfeifer, G. P. (2015). Aging and DNA methylation. *BMC Biology*.

Kafri, T., Ariel, M., Brandeis, M., Shemer, R., Urven, L., McCarrey, J., Cedar, H., and Razin, A. (1992). Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. *Genes and Development*.

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., Van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., Taatjes, D. J., Dekker, J., and Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*.

Kajitani, N., Satsuka, A., Kawate, A., and Sakai, H. (2012). Productive lifecycle of human papillomaviruses that depends upon squamous epithelial differentiation. *Frontiers in Microbiology*, 3(APR):1–12.

Kasowitz, S. D., Ma, J., Anderson, S. J., Leu, N. A., Xu, Y., Gregory, B. D., Schultz, R. M., and Wang, P. J. (2018). Nuclear m6A reader YTHDC1 regulates alternative polyadenylation and splicing during mouse oocyte development. *PLoS Genetics*, 14(5):1–28.

Ke, S., Alemu, E. A., Mertens, C., Gantman, E. C., Fak, J. J., Mele, A., Haripal, B., Zucker-Scharff, I., Moore, M. J., Park, C. Y., Vågbø, C. B., Kuśnierczyk, A., Klungland, A., Darnell, J. E., and Darnell, R. B. (2015). A majority of m6A residues are in the last exons, allowing the potential for 3? UTR regulation. *Genes and Development*.

Ke, S., Pandya-Jones, A., Saito, Y., Fak, J. J., Vågbø, C. B., Geula, S., Hanna, J. H., Black, D. L., Darnell, J. E., and Darnell, R. B. (2017). m6A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover. *Genes and Development*, 31(10):990–1006.

Kirschner, K., Chandra, T., Kiselev, V., Flores-Santa Cruz, D., Macaulay, I. C., Park, H. J., Li, J., Kent, D. G., Kumar, R., Pask, D. C., Hamilton, T. L., Hemberg, M., Reik, W., and Green, A. R. (2017). Proliferation Drives Aging-Related Functional Decline in a Subpopulation of the Hematopoietic Stem Cell Compartment. *Cell Reports*.

Knudson, A. G. (1971). Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823.

Koch, C. M., Andrews, R. M., Flicek, P., Dillon, S. C., Karaöz, U., Clelland, G. K., Wilcox, S., Beare, D. M., Fowler, J. C., Couttet, P., James, K. D., Lefebvre, G. C., Bruce, A. W., Dovey, O. M., Ellis, P. D., Dhami, P., Langford, C. F., Weng, Z., Birney, E., Carter, N. P., Vetrie, D., and Dunham, I. (2007). The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Research*.

Koh, P. W., Sinha, R., Barkal, A. A., Morganti, R. M., Chen, A., Weissman, I. L., Ang, L. T., Kundaje, A., and Loh, K. M. (2016). An atlas of transcriptional, chromatin accessibility, and surface marker changes in human mesoderm development. *Scientific Data*.

Kojima, S., Sher-Chen, E. L., and Green, C. B. (2012). Circadian control of mRNA polyadenylation dynamics regulates rhythmic protein expression. *Genes and Development*.

Kornberg, R. D. (1977). Structure of chromatin. *Annual review of biochemistry*, 46(1):931–954.

Kornberg, R. D. (1999). Eukaryotic transcriptional control. *Trends in biochemical sciences*, 24(12):M46–M49.

Kosak, S. T., Skok, J. A., Medina, K. L., Riblet, R., Le Beau, M. M., Fisher, A. G., and Singh, H. (2002). Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. *Science*.

Kowalczyk, M. S., Tirosh, I., Heckl, D., Rao, T. N., Dixit, A., Haas, B. J., Schneider, R. K., Wagers, A. J., Ebert, B. L., and Regev, A. (2015). Single-cell rna-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome research*, 25(12):1860–1872.

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., and Turner, D. J. (2009). Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (g+ c)-biased genomes. *Nature methods*, 6(4):291.

Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., and Cech, T. R. (1982). Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell*.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645.

Kühn, U., Gündel, M., Knoth, A., Kerwitz, Y., Rüdel, S., and Wahle, E. (2009). Poly (a) tail length is controlled by the nuclear poly (a)-binding protein regulating the interaction between poly (a) polymerase and the cleavage and polyadenylation specificity factor. *Journal of Biological Chemistry*, 284(34):22803–22814.

Kühnert, B. and Nieschlag, E. (2004). Reproductive functions of the ageing male. *Human Reproduction Update*, 10(4):327–339.

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., Medvedeva, Y. A., Magana-Mora, A., Bajic, V. B., Papatsenko, D. A., Kolpakov, F. A., and Makeev, V. J. (2018). HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research*.

Kumar, R. M., Cahan, P., Shalek, A. K., Satija, R., Keyser, A. D., Li, H., Zhang, J., Pardee, K., Gennert, D., Trombetta, J. J., Ferrante, T. C., Regev, A., Daley, G. Q., and Collins, J. J. (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature*.

Kurosaki, T. and Maquat, L. E. (2016). Nonsense-mediated mRNA decay in humans at a glance. *Journal of Cell Science*.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science*.

Lajoie, B. R., Dekker, J., and Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods*, 72(C):65–75.

Lamfrom, H., McLaughlin, C. S., and Sarabhai, A. (1966). Direction of reading the genetic message in reticulocytes.

Lammerding, J. (2011). Mechanics of the nucleus. *Comprehensive Physiology*.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Hong, M. L., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel,

N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., De La Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., and Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*.

Landry, J. J. M., Pyl, P. T., Rausch, T., Zichner, T., Tekkedil, M. M., Stütz, A. M., Jauch, A., Aiyar, R. S., Pau, G., Delhomme, N., Gagneur, J., Korbel, J. O., Huber, W., and Steinmetz, L. M. (2013). The Genomic and Transcriptomic Landscape of a HeLa Cell Line. *G3: Genes|Genomes|Genetics*, 3(8):1213–1224.

Langemeier, J., Schrom, E. M., Rabner, A., Radtke, M., Zychlinski, D., Saborowski, A., Bohn, G., Mandel-Gutfreund, Y., Bodem, J., Klein, C., and Bohne, J. (2012). A complex immunodeficiency is based on U1 snRNP-mediated poly(A) site suppression. *EMBO Journal*.

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

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*.

Larsson, A. J., Johnsson, P., Hagemann-Jensen, M., Hartmanis, L., Faridani, O. R., Reinius, B., Segerstolpe, Å., Rivera, C. M., Ren, B., and Sandberg, R. (2019). Genomic encoding of transcriptional burst kinetics.

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*.

Lee, H., Bao, S., Qian, Y., Geula, S., Leslie, J., Zhang, C., Hanna, J. H., and Ding, L. (2019). Stage-specific requirement for mettl3-dependent m 6 a mrna methylation during haematopoietic stem cell differentiation. *Nature cell biology*, 21(6):700.

Lee, Y., Kim, M., Han, J., Yeom, K. H., Lee, S., Baek, S. H., and Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *EMBO Journal*.

Lehner, B. (2010). Conflict between noise and plasticity in yeast. *PLoS Genetics*, 6(11).

Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: Emerging characteristics and insights into transcriptional regulation.

Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*.

Levesque, M. J. and Raj, A. (2013). Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nature Methods*.

Li, A., Chen, Y. S., Ping, X. L., Yang, X., Xiao, W., Yang, Y., Sun, H. Y., Zhu, Q., Baidya, P., Wang, X., Bhattarai, D. P., Zhao, Y. L., Sun, B. F., and Yang, Y. G. (2017). Cytoplasmic m 6 A reader YTHDF3 promotes mRNA translation. *Cell Research*, 27(3):444–447.

Li, E. and Zhang, Y. (2014). DNA methylation in mammals. *Cold Spring Harbor Perspectives in Biology*.

Li, F., Yi, Y., Miao, Y., Long, W., Long, T., Chen, S., Cheng, W., Zou, C., Zheng, Y., Wu, X., et al. (2019). N6-methyladenosine modulates nonsense-mediated mrna decay in human glioblastoma. *Cancer research*, pages canres–2868.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*.

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

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

Li, Y., Zhang, Z., Chen, J., Liu, W., Lai, W., Liu, B., Li, X., Liu, L., Xu, S., Dong, Q., Wang, M., Duan, X., Tan, J., Zheng, Y., Zhang, P., Fan, G., Wong, J., Xu, G. L., Wang, Z., Wang, H., Gao, S., and Zhu, B. (2018). Stella safeguards the oocyte methylome by preventing de novo methylation mediated by DNMT1.

Lieberman-Aiden, E., Berkum, N. L. V., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 33292(October):289–294.

Lim, J., Ha, M., Chang, H., Kwon, S. C., Simanshu, D. K., Patel, D. J., and Kim, V. N. (2014). Uridylation by TUT4 and TUT7 marks mRNA for degradation. *Cell*, 159(6):1365–1376.

Lim, J., Maher, G. J., Turner, G. D., Dudka-Ruszkowska, W., Taylor, S., Meyts, E. R. D., Goriely, A., and Wilkie, A. O. (2012). Selfish spermatogonial selection: Evidence from an immunohistochemical screen in testes of elderly men. *PLoS ONE*, 7(8).

Lim, L. and Canellakis, E. S. (1970). Adenine-rich polymer associated with rabbit reticulocyte messenger RNA. *Nature*.

Linder, B., Grozhik, A. V., Olarerin-George, A. O., Meydan, C., Mason, C. E., and Jaffrey, S. R. (2015). Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nature Methods*.

Lister, R. and Ecker, J. R. (2009). Finding the fifth base: Genome-wide sequencing of cytosine methylation.

Liu, C., Cheng, Y. J., Wang, J. W., and Weigel, D. (2017a). Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. *Nature Plants*.

Liu, H., Begik, O., Lucas, M. C., Ramirez, J. M., Mason, C. E., Wiener, D., Schwartz, S., Mattick, J. S., Smith, M. A., and Novoa, E. M. (2019). Accurate detection of m6A RNA modifications in native RNA sequences. *Nature Communications*.

Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., Jia, G., Yu, M., Lu, Z., Deng, X., Dai, Q., Chen, W., and He, C. (2013). A METTL3?METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nature Chemical Biology*, 10(2):93–95.

Liu, J., Yue, Y., Liu, J., Cui, X., Cao, J., Luo, G., Zhang, Z., Cheng, T., Gao, M., Shu, X., Ma, H., Wang, F., Wang, X., Shen, B., Wang, Y., Feng, X., and He, C. (2018). VIRMA mediates preferential m6A mRNA methylation in 3?UTR and near stop codon and associates with alternative polyadenylation. *Cell Discovery*.

Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M., and Pan, T. (2015). N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature*, 518(7540):560–564.

Liu, N. and Pan, T. (2015). RNA epigenetics.

Liu, N., Zhou, K. I., Parisien, M., Dai, Q., Diatchenko, L., and Pan, T. (2017b). N6-methyladenosine alters RNA structure to regulate binding of a low-complexity protein. *Nucleic Acids Research*, 45(10):6051–6063.

Liu, Y., Lu, Z., Xu, R., and Ke, Y. (2016). Comprehensive mapping of the human papillomavirus (HPV) DNA integration sites in cervical carcinomas by HPV capture technology. *Oncotarget*, 7(5).

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. *Cell*, 153(6):1194–217.

Louloupi, A., Ntini, E., Conrad, T., and Ørom, U. A. V. (2018). Transient N-6-Methyladenosine Transcriptome Sequencing Reveals a Regulatory Role of m6A in Splicing Efficiency. *Cell Reports*.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21.

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 å resolution. *Nature*, 389(6648):251.

Luger, K. and Richmond, T. J. (1998). The histone tails of the nucleosome. *Current Opinion in Genetics and Development*.

Lun, A. T. and Marioni, J. C. (2017). Overcoming confounding plate effects in differential expression analyses of single-cell rna-seq data. *Biostatistics*, 18(3):451–464.

Lun, A. T., Riesenfeld, S., Andrews, T., Dao, T. P., Gomes, T., and Marioni, J. C. (2019). EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biology*.

Lun, A. T. L., McCarthy, D. J., and Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. *F1000Res.*, 5:2122.

Lv, J., Zhang, Y., Gao, S., Zhang, C., Chen, Y., Li, W., Yang, Y.-G., Zhou, Q., and Liu, F. (2018). Endothelial-specific m 6 a modulates mouse hematopoietic stem and progenitor cell development via notch signaling. *Cell research*, 28(2):249.

Macarthur, B. D. and Lemischka, I. R. (2013). Statistical mechanics of pluripotency.

Machanick, P. and Bailey, T. L. (2011). MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics*, 27(12):1696–1697.

MacLeod, A. O. and McCarty, M. (1944). Studies of the chemical nature of the substance inducing transformation of pneumococcal types. induction of transformation by a deoxyribonucleic acid fraction isolated from pneumococcus type iii. *Journal of Experimental Medicine*, 79:137–158.

Maekawa, S., Imamachi, N., Irie, T., Tani, H., Matsumoto, K., Mizutani, R., Imamura, K., Kakeda, M., Yada, T., Sugano, S., Suzuki, Y., and Akimitsu, N. (2015). Analysis of RNA decay factor mediated RNA stability contributions on RNA abundance. *BMC Genomics*.

Magnúsdóttir, E. and Azim Surani, M. (2014). How to make a primordial germ cell. *Development (Cambridge)*.

Maksakova, I. A., Romanish, M. T., Gagnier, L., Dunn, C. A., Van De Lagemaat, L. N., and Mager, D. L. (2006). Retroviral elements and their hosts: Insertional mutagenesis in the mouse germ line.

Martinez-Jimenez, C. P., Eling, N., Chen, H. C., Vallejos, C. A., Kolodziejczyk, A. A., Connor, F., Stojic, L., Rayner, T. F., Stubbington, M. J., Teichmann, S. A., De La Roche, M., Marioni, J. C., and Odom, D. T. (2017). Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science*, 355(6332):1433–1436.

Matharu, N. and Ahituv, N. (2015). Minor Loops in Major Folds: Enhancer?Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease.

Mauer, J., Luo, X., Blanjoie, A., Jiao, X., Grozhik, A. V., Patil, D. P., Linder, B., Pickering, B. F., Vasseur, J. J., Chen, Q., Gross, S. S., Elemento, O., Debart, F., Kiledjian, M., and Jaffrey, S. R. (2017). Reversible methylation of m 6 A m in the 5? cap controls mRNA stability. *Nature*.

Mayr, C. (2017). Regulation by 3?-Untranslated Regions. *Annual Review of Genetics*.

McBride, A. A., Sakakibara, N., Stepp, W. H., and Jang, M. K. (2012). Hitchhiking on host chromatin: How papillomaviruses persist. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1819(7):820–825.

McCarthy, D. J., Campbell, K. R., Lun, A. T. L., and Wills, Q. F. (2017). Scater: pre-processing, quality control, normalisation and visualisation of single-cell rna-seq data in r. *Bioinformatics*, 14 Jan.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297.

McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S. D., Wickens, M., and Bentley, D. L. (1997). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature*.

McDonald, M. J., Rice, D. P., and Desai, M. M. (2016). Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature*, 531(7593):233–236.

McGettigan, P. A. (2013). Transcriptomics in the RNA-seq era.

Meijer, H. A., Bushell, M., Hill, K., Gant, T. W., Willis, A. E., Jones, P., and de Moor, C. H. (2007). A novel method for poly(A) fractionation reveals a large population of mRNAs with a short poly(A) tail in mammalian cells. *Nucleic Acids Research*.

Meister, P., Towbin, B. D., Pike, B. L., Ponti, A., and Gasser, S. M. (2010). The spatial dynamics of tissue-specific promoters during C. elegans development. *Genes and Development*.

Mendel, M., Chen, K. M., Homolka, D., Gos, P., Pandey, R. R., McCarthy, A. A., and Pillai, R. S. (2018). Methylation of Structured RNA by the m 6 A Writer METTL16 Is Essential for Mouse Embryonic Development. *Molecular Cell*.

Mercer, T. R., Gerhardt, D. J., Dinger, M. E., Crawford, J., Trapnell, C., Jeddeloh, J. A., Mattick, J. S., and Rinn, J. L. (2012). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nature Biotechnology*, 30(1):99–104.

Meyer, K. D. and Jaffrey, S. R. (2014). The dynamic epitranscriptome: N6-methyladenosine and gene expression control.

Meyer, K. D., Patil, D. P., Zhou, J., Zinoviev, A., Skabkin, M. A., Elemento, O., Pestova, T. V., Qian, S. B., and Jaffrey, S. R. (2015). 5? UTR m6A Promotes Cap-Independent Translation. *Cell*.

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3? UTRs and near stop codons. *Cell*, 149(7):1635–1646.

Miescher-Rüsch, F. (1871). *Ueber die chemische Zusammensetzung der Eiterzellen*.

Mifsud, B., Martincorena, I., Darbo, E., Sugar, R., Francis, T., and Wca, L. (2017). GOTHiC, simple probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. *PLoS ONE*, pages 1–15.

Min, K. W., Zealy, R. W., Davila, S., Fomin, M., Cummings, J. C., Makowsky, D., Mcdowell, C. H., Thigpen, H., Hafner, M., Kwon, S. H., Georgescu, C., Wren, J. D., and Yoon, J. H. (2018). Profiling of m6A RNA modifications identified an age-associated regulation of AGO2 mRNA stability. *Aging Cell*.

Mirny, L. A. (2011). The fractal globule as a model of chromatin architecture in the cell. *Chromosome Research*.

Mizutani, S. and Temin, H. M. (1970). An RNA-Dependent DNA Polymerase in Virions of Rous Sarcoma Virus. *Cold Spring Harbor Symposia on Quantitative Biology*.

Mojtahedi, M., Skupin, A., Zhou, J., Castaño, I. G., Leong-Quong, R. Y., Chang, H., Trachana, K., Giuliani, A., and Huang, S. (2016). Cell Fate Decision as High-Dimensional Critical State Transition. *PLoS Biology*.

Monk, M., Boubelik, M., and Lehnert, S. (1987). Temporal and regional changes in dna methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. *Development*, 99(3):371–382.

Morgan, M., Much, C., DiGiacomo, M., Azzi, C., Ivanova, I., Vitsios, D. M., Pistolic, J., Collier, P., Moreira, P. N., Benes, V., Enright, A. J., and O'Carroll, D. (2017). MRNA 3′ uridylation and poly(A) tail length sculpt the mammalian maternal transcriptome. *Nature*, 548(7667):347–351.

Moskowitz, D. M., Zhang, D. W., Hu, B., Le Saux, S., Yanes, R. E., Ye, Z., Buenrostro, J. D., Weyand, C. M., Greenleaf, W. J., and Goronzy, J. J. (2017). Epigenomics of human CD8 T cell differentiation and aging. *Science Immunology*, 2(8):1–14.

Mudge, J. M. and Harrow, J. (2015). Creating reference gene annotation for the mouse C57BL6/J genome assembly.

Muhlrad, D., Decker, C. J., and Parker, R. (1994). Deadenylation of the unstable mRNA encoded by the yeast MFA2 gene leads to decapping followed by 5' ? 3' digestion of the transcript. *Genes and Development*.

Mukherjee, C., Patil, D. P., Kennedy, B. A., Bakthavachalu, B., Bundschuh, R., and Schoenberg, D. R. (2012). Identification of Cytoplasmic Capping Targets Reveals a Role for Cap Homeostasis in Translation and mRNA Stability. *Cell Reports*, 2(3):674–684.

Mukherjee, D., Gao, M., O'Connor, J. P., Raijmakers, R., Pruijn, G., Lutz, C. S., and Wilusz, J. (2002). The mammalian exosome mediates the efficient degradation of mRNAs that contain AU-rich elements. *EMBO Journal*, 21(1-2):165–174.

Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*.

Murthy, K. G. and Manley, J. L. (1995). The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation. *Genes and Development*.

Nagano, T., Várnai, C., Schoenfelder, S., Javierre, B. M., Wingett, S. W., and Fraser, P. (2015). Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biology*, 16(1):1–13.

Nagase, N. and Brinton, M. C. (2017). The gender division of labor and second births: Labor market institutions and fertility in Japan. *Demographic Research*.

Nakagawa, T., Sharma, M., Nabeshima, Y. I., Braun, R. E., and Yoshida, S. (2010). Functional hierarchy and reversibility within the murine spermatogenic stem cell compartment. *Science*, 328(5974):62–67.

Nan, X., Ng, H. H., Johnson, C. A., Laherty, C. D., Turner, B. M., Eisenman, R. N., and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*.

Napoli, C., Lemieux, C., and Jorgensen, R. (1990). Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. *Plant Cell*.

Nelson, N. J. (2001). Microarrays Have Arrived: Gene Expression Tool Matures. *JNCI Journal of the National Cancer Institute*.

Nicolas, D., Zoller, B., Suter, D. M., and Naef, F. (2018). Modulation of transcriptional burst frequency by histone acetylation. *Proceedings of the National Academy of Sciences*, 115(27):7153–7158.

Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P., and Barron, A. E. (2011). Landscape of next-generation sequencing technologies.

Noordermeer, D., Branco, M. R., Splinter, E., Klous, P., Van Ijcken, W., Swagemakers, S., Koutsourakis, M., Van Der Spek, P., Pombo, A., and De Laat, W. (2008). Transcription and chromatin organization of a housekeeping gene cluster containing an integrated β-globin locus control region. *PLoS Genetics*.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., and Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*.

Norbury, C. J. (2013). Cytoplasmic RNA: A case of the tail wagging the dog. *Nature Reviews Molecular Cell Biology*.

Oakberg, E. (1971). Spermatogonial stem-cell renewal in the mouse. *The Anatomical Record*, 169(3):515–531.

Oakberg, E. F. (1956). A description of spermiogenesis in the mouse and its use in analysis of the cycle of the seminiferous epithelium and germ cell renewal. *American Journal of Anatomy*, 99(3):391–413.

Oatley, J. M. and Brinster, R. L. (2008). Regulation of Spermatogonial Stem Cell Self-Renewal in Mammals. *Annual Review of Cell and Developmental Biology*.

Oatley, J. M., Kaucher, A. V., Avarbock, M. R., and Brinster, R. L. (2010). Regulation of Mouse Spermatogonial Stem Cell Differentiation by STAT3 Signaling1. *Biology of Reproduction*.

O'Donnell, L., Nicholls, P. K., O'Bryan, M. K., McLachlan, R. I., and Stanton, P. G. (2011). Spermiation: The process of sperm release. *Spermatogenesis*.

Ojesina, A. I., Lichtenstein, L., Freeman, S. S., Pedamallu, C. S., Imaz-Rosshandler, I., Pugh, T. J., Cherniack, A. D., Ambrogio, L., Cibulskis, K., Bertelsen, B., Romero-Cordoba, S., Treviño, V., Vazquez-Santillan, K., Guadarrama, A. S., Wright, A. A., Rosenberg, M. W., Duke, F., Kaplan, B., Wang, R., Nickerson, E., Walline, H. M., Lawrence, M. S., Stewart, C., Carter, S. L., McKenna, A., Rodriguez-Sanchez, I. P., Espinosa-Castilla, M., Woie, K., Bjorge, L., Wik, E., Halle, M. K., Hoivik, E. A., Krakstad, C., Gabiño, N. B., Gómez-Macías, G. S., Valdez-Chapa, L. D., Garza-Rodríguez, M. L., Maytorena, G., Vazquez, J., Rodea, C., Cravioto, A., Cortes, M. L., Greulich, H., Crum, C. P., Neuberg, D. S., Hidalgo-Miranda, A., Escareno, C. R., Akslen, L. A., Carey, T. E., Vintermyr, O. K., Gabriel, S. B., Barrera-Saldaña, H. A., Melendez-Zajgla, J., Getz, G., Salvesen, H. B., and Meyerson, M. (2014). Landscape of genomic alterations in cervical carcinomas. *Nature*, 506(7488):371–375.

Ong, C. T. and Corces, V. G. (2014). CTCF: An architectural protein bridging genome topology and function.

ONS (2017). Births by parents' characteristics in England and Wales: 2016. *Office for National Statistics*, pages 1–12.

Orgel, L. E. (1968). Evolution of the genetic apparatus. *Journal of molecular biology*, 38(3):381–393.

Orphanides, G., Lagrange, T., and Reinberg, D. (1996). The general transcription factors of RNA polymerase II.

Pan, T. (2018). Modifications and functional genomics of human transfer RNA. *Cell Research*, 28(4):395–404.

Pandolfini, L., Barbieri, I., Bannister, A. J., Hendrick, A., Andrews, B., Webster, N., Murat, P., Mach, P., Brandi, R., Robson, S. C., Migliori, V., Alendar, A., D'Onofrio, M., Balasubramanian, S., and Kouzarides, T. (2019). METTL1 Promotes let-7 MicroRNA Processing via m7G Methylation. *Molecular Cell*.

Paris, C., Pentland, I., Groves, I., Roberts, D. C., Powis, S. J., Coleman, N., Roberts, S., and Parish, J. L. (2015). CCCTC-Binding Factor Recruitment to the Early Region of the Human Papillomavirus 18 Genome Regulates Viral Oncogene Expression. *Journal of Virology*, 89(9):4770–4785.

Paris, J., Morgan, M., Campos, J., Spencer, G. J., Shmakova, A., Ivanova, I., Mapperley, C., Lawson, H., Wotherspoon, D. A., Sepulveda, C., Vukovic, M., Allen, L., Sarapuu, A., Tavosanis, A., Guitart, A. V., Villacreces, A., Much, C., Choe, J., Azar, A., van de Lagemaat, L. N., Vernimmen, D., Nehme, A., Mazurier, F., Somervaille, T. C., Gregory, R. I., O'Carroll, D., and Kranc, K. R. (2019). Targeting the RNA m6A Reader YTHDF2 Selectively Compromises Cancer Stem Cells in Acute Myeloid Leukemia. *Cell Stem Cell*, pages 1–12.

Patil, D. P., Chen, C. K., Pickering, B. F., Chow, A., Jackson, C., Guttman, M., and Jaffrey, S. R. (2016). M6A RNA methylation promotes XIST-mediated transcriptional repression. *Nature*, 537(7620):369–373.

Paul, C. and Robaire, B. (2013). Ageing of the male germ line.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., David, E., Cohen, N., Lauridsen, F. K. B., Haas, S., Schlitzer, A., Mildner, A., Ginhoux, F., Jung, S., Trumpp, A., Porse, B. T., Tanay, A., and Amit, I. (2015). Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell*.

Paule, M. R. and White, R. J. (2000). Transcription by RNA polymerases I and III. *Nucleic Acids Research*.

Peaston, A. E., Evsikov, A. V., Graber, J. H., de Vries, W. N., Holbrook, A. E., Solter, D., and Knowles, B. B. (2004). Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Developmental Cell*.

Pennacchio, L. A., Loots, G. G., Nobrega, M. A., and Ovcharenko, I. (2007). Predicting tissue-specific enhancers in the human genome. *Genome Research*.

Pentland, I., Campos-León, K., Cotic, M., Davies, K. J., Wood, C. D., Groves, I. J., Burley, M., Coleman, N., Stockton, J. D., Noyvert, B., Beggs, A. D., West, M. J., Roberts, S., and Parish, J. L. (2018). Disruption of CTCF-YY1-dependent looping of the human papillomavirus genome activates differentiation-induced viral oncogene transcription. *PLoS biology*, 16(10):e2005752.

Perry, R. P., Kelley, D. E., Friderici, K., and Rottman, F. (1975). The methylated constituents of L cell messenger RNA: Evidence for an unusual cluster at the 5? terminus. *Cell*, 4(4):387–394.

Pett, M. and Coleman, N. (2007). Integration of high-risk human papillomavirus: A key event in cervical carcinogenesis? *Journal of Pathology*, 212(4):356–367.

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, 9(1):171–181.

Plaschka, C., Larivière, L., Wenzeck, L., Seizl, M., Hemann, M., Tegunov, D., Petrotchenko, E. V., Borchers, C. H., Baumeister, W., Herzog, F., Villa, E., and Cramer, P. (2015). Architecture of the RNA polymerase II-Mediator core initiation complex. *Nature*.

Popescu, N. C., Dipaolo, J. A., and Amsbaugh, S. C. (1987). Integration sites of human papillomavirus 18 dna sequences on hela cell chromosomes. *Cytogenetic and Genome Research*.

Popp, C., Dean, W., Feng, S., Cokus, S. J., Andrews, S., Pellegrini, M., Jacobsen, S. E., and Reik, W. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature*.

Posadas, D. M. and Carthew, R. W. (2014). MicroRNAs and their roles in developmental canalization. *Current Opinion in Genetics and Development*, 27:1–6.

Poss, Z. C., Ebmeier, C. C., and Taatjes, D. J. (2013). The Mediator complex and transcription regulation.

Proudfoot, N. J. (2011). Ending the message: Poly(A) signals then and now.

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

Radman-Livaja, M. and Rando, O. J. (2010). Nucleosome positioning: How is it established, and why does it matter?

Ragoczy, T., Bender, M. A., Telling, A., Byron, R., and Groudine, M. (2006). The locus control region is required for association of the murine β-globin locus with engaged transcription factories during erythroid maturation. *Genes and Development*.

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., and Manke, T. (2014). deeptools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research*, 42(W1):W187–W191.

Ransom, B., Goldman, S. A., Meldolesi, J., Zhou, L., Murai, K. K., Harris, K. M., Mccarthy, K. D., Li, N., Doyle, R. T., Haydon, P. G., Zielke, H. R., Ni, Y., Sunjara, V., Hua, X., Parpura, V., Jiang, L., Nedergaard, M., Xu, J., Xu, Q., Kang, J., Mulligan, S. J., Macvicar, B. A., Harder, D. R., Alkayed, N. J., Lange, A. R., Gebremedhin, D., and Roman, R. J. (2008). Proliferating Cells Express mRNAs with Shortened 3? Untranslated Regions and Fewer MicroRNA Target Sites. *Science*, 320(June):1643–1647.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., and Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680.

Raser, J. M. and O'Shea, E. K. (2005). Noise in Gene Expression : Origins, Consequences and Control. *Science*, 309:2010–2014.

Ravarani, C. N., Chalancon, G., Breker, M., De Groot, N. S., and Babu, M. M. (2016). Affinity and competition for TBP are molecular determinants of gene expression noise. *Nature Communications*.

Richter, J. D. (2000). Influence of polyadenylation-induced translation on metazoan development and neuronal synaptic function. *Cold Spring Harbor Monograph Archive*, 39:785–805.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*.

Roden, R. B. and Stern, P. L. (2018). Opportunities and challenges for human papillomavirus vaccination in cancer. *Nature Reviews Cancer*, 18(4):240–254.

Roeder, R. G. and Rutter, W. J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature*.

Rogers, J., Early, P., Carter, C., Calame, K., Bond, M., Hood, L., and Wall, R. (1980). Two mRNAs with different 3? ends encode membrane-bound and secreted forms of immunoglobulin μ chain. *Cell*.

Rosa-Mercado, N. A., Withers, J. B., and Steitz, J. A. (2017). Settling the m6A debate: Methylation of mature mRNA is not dynamic but accelerates turnover. *Genes and Development*.

Rosenfeld, J. A., Wang, Z., Schones, D. E., Zhao, K., DeSalle, R., and Zhang, M. Q. (2009). Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics*.

Rossi, P., Sette, C., Dolci, S., and Geremia, R. (2000). Role of c-kit in mammalian spermatogenesis.

Roundtree, I. A., Luo, G. Z., Zhang, Z., Wang, X., Zhou, T., Cui, Y., Sha, J., Huang, X., Guerrero, L., Xie, P., He, E., Shen, B., and He, C. (2017). YTHDC1 mediates nuclear export of N6-methyladenosine methylated mRNAs. *eLife*, 6:1–28.

Safra, M., Sas-Chen, A., Nir, R., Winkler, R., Nachshon, A., Bar-Yaacov, D., Erlacher, M., Rossmanith, W., Stern-Ginossar, N., and Schwartz, S. (2017). The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution. *Nature*.

Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development*.

Salas, M., Smith, M., Stanley Jr, W., Wahba, A., and Ochoa, S. (1965). Direction of reading of the genetic message. *The Journal of biological chemistry*, 240(10):3988.

Saletore, Y., Meyer, K., Korlach, J., Vilfan, I. D., Jaffrey, S., and Mason, C. E. (2012). The birth of the Epitranscriptome: deciphering the function of RNA modifications. *Genome biology*.

Sanger, F., Nicklen, S., and Coulson, a. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7.

Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*.

Scarpini, C. G., Groves, I. J., Pett, M. R., Ward, D., and Coleman, N. (2014). Virus transcript levels and cell growth rates after naturally occurring HPV16 integration events in basal cervical keratinocytes. *Journal of Pathology*, 233(3):281–293.

Schiffman, M., Wentzensen, N., Wacholder, S., Kinney, W., Gage, J. C., and Castle, P. E. (2011). Human papillomavirus testing in the prevention of cervical cancer. *Journal of the National Cancer Institute*, 103(5):368–383.

Schmidt, J. A., Avarbock, M. R., Tobias, J. W., and Brinster, R. L. (2009). Identification of Glial Cell Line-Derived Neurotrophic Factor-Regulated Genes Important for Spermatogonial Stem Cell Self-Renewal in the Rat1. *Biology of Reproduction*.

Schmiedel, J. M., Klemm, S. L., Zheng, Y., Sahay, A., Blüthgen, N., Marks, D. S., and Van Oudenaarden, A. (2015). MicroRNA control of protein expression noise. *Science*, 348(6230):128–131.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B. M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S. W., Dimitrova, E., Dimond, A., Edelman, L. B., Elderkin, S., Tabbada, K., Darbo, E., Andrews, S., Herman, B., Higgs, A., LeProust, E., Osborne, C. S., Mitchell, J. A., Luscombe, N. M., and Fraser, P. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research*, 25(4):582–597.

Schwartz, S., Mumbach, M. R., Jovanovic, M., Wang, T., Maciag, K., Bushkin, G. G., Mertins, P., Ter-Ovanesyan, D., Habib, N., Cacchiarelli, D., Sanjana, N. E., Freinkman, E., Pacold, M. E., Satija, R., Mikkelsen, T. S., Hacohen, N., Zhang, F., Carr, S. A., Lander, E. S., and Regev, A. (2014). Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5' sites. *Cell Reports*, 8(1):284–296.

Sergushichev, A. A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*.

Sharma, M. and Braun, R. E. (2018). Cyclical expression of GDNF is required for spermatogonial stem cell homeostasis. *Development (Cambridge)*, 145(5).

Sherman, M. S., Lorenz, K., Lanier, M. H., and Cohen, B. A. (2015). Cell-to-Cell Variability in the Propensity to Transcribe Explains Correlated Fluctuations in Gene Expression. *Cell Systems*, 1(5):315–325.

Shi, H., Wang, X., Lu, Z., Zhao, B. S., Ma, H., Hsu, P. J., Liu, C., and He, C. (2017). YTHDF3 facilitates translation and decay of N 6-methyladenosine-modified RNA. *Cell Research*.

Shi, H., Wei, J., and He, C. (2019). Where, When, and How: Context-Dependent Functions of RNA Methylation Writers, Readers, and Erasers.

Siciliano, V., Garzilli, I., Fracassi, C., Criscuolo, S., Ventre, S., and Di Bernardo, D. (2013). MiRNAs confer phenotypic robustness to gene networks by suppressing biological noise. *Nature Communications*.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(539).

Silva, A. L., Ribeiro, P., Inácio, Â., Liebhaber, S. A., and Romão, L. (2008). Proximity of the poly (a)-binding protein to a premature termination codon inhibits mammalian nonsense-mediated mrna decay. *Rna*, 14(3):563–576.

Sim, G. K., Kafatos, F. C., Jones, C. W., Koehler, M. D., Efstratiadis, A., and Maniatis, T. (1979). Use of a cDNA library for studies on evolution and developmental expression of the chorion multigene families. *Cell*.

Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*.

Singh, I., Lee, S. H., Sperling, A. S., Samur, M. K., Tai, Y. T., Fulciniti, M., Munshi, N. C., Mayr, C., and Leslie, C. S. (2018). Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nature Communications*.

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*.

Smith, E. M., Lajoie, B. R., Jain, G., and Dekker, J. (2016). Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus. *American Journal of Human Genetics*, 98(1):185–201.

Solovei, I., Cavallo, A., Schermelleh, L., Jaunin, F., Scasselati, C., Cmarko, D., Cremer, C., Fakan, S., and Cremer, T. (2002). Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). *Experimental Cell Research*.

Soutourina, J. (2018). Transcription regulation by the Mediator complex.

Stanley, M. A., Browne, H. M., Appleby, M., and Minson, A. C. (1989). Properties of a non?tumorigenic human cervical keratinocyte cell line. *International Journal of Cancer*, 43(4):672–676.

Steege, D. A. (2000). Emerging features of mRNA decay in bacteria.

Stegeman, R. and Weake, V. M. (2017). Transcriptional Signatures of Aging. *Journal of Molecular Biology*, 429(16):2427–2437.

Steger, D. J., Lefterova, M. I., Ying, L., Stonestrom, A. J., Schupp, M., Zhuo, D., Vakoc, A. L., Kim, J.-E., Chen, J., Lazar, M. A., Blobel, G. A., and Vakoc, C. R. (2008). DOT1L/KMT4 Recruitment and H3K79 Methylation Are Ubiquitously Coupled with Gene Transcription in Mammalian Cells. *Molecular and Cellular Biology*.

Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H., and Bartel, D. P. (2014). Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature*.

Swain, P. S. (2004). Efficient attenuation of stochasticity in gene expression through post-transcriptional control. *Journal of Molecular Biology*.

Swain, P. S., Elowitz, M. B., and Siggia, E. D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800.

Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J., and Stahl, F. W. (1983). The double-strand-break repair model for recombination.

Thomas, M. C. and Chiang, C. M. (2006). The general transcription machinery and general cofactors.

Thorland, E. C., Myers, S. L., Persing, D. H., Sarkar, G., McGovern, R. M., Gostout, B. S., and Smith, D. I. (2000). Human papillomavirus type 16 integrations in cervical tumors frequently occur in common fragile sites. *Cancer research*, 60(21):5916–5921.

Torres-Padilla, M. E. and Chambers, I. (2014). Transcription factor heterogeneity in pluripotent stem cells: A stochastic advantage.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386.

Trotta, E. (2014). On the normalization of the minimum free energy of RNAs by sequence length. *PLoS ONE*, 9(11).

Tsukamoto, H., Clise-Dwyer, K., Huston, G. E., Duso, D. K., Buck, A. L., Johnson, L. L., Haynes, L., and Swain, S. L. (2009). Age-associated increase in lifespan of naive CD4 T cells contributes to T-cell homeostasis but facilitates development of functional defects. *Proceedings of the National Academy of Sciences*, 106(43):18333–18338.

Ucar, D., Márquez, E. J., Chung, C. H., Marches, R., Rossi, R. J., Uyar, A., Wu, T. C., George, J., Stitzel, M. L., Karolina Palucka, A., Kuchel, G. A., and Banchereau, J. (2017). The chromatin accessibility signature of human immune aging stems from CD8+ T cells. *Journal of Experimental Medicine*.

Uesaka, T., Jain, S., Yonemura, S., Uchiyama, Y., Milbrandt, J., and Enomoto, H. (2007). Conditional ablation of GFRα1 in postmigratory eneric neurons triggers unconventional neuronal death in the colon and causes Hirschsprungs's disease phenotype. *Development*, 134(11):2171–2181.

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., Sanli, K., Von Feilitzen, K., Oksvold, P., Lundberg, E., Hober, S., Nilsson, P., Mattsson, J., Schwenk, J. M., Brunnström, H., Glimelius, B., Sjöblom, T., Edqvist, P. H., Djureinovic, D., Micke, P., Lindskog, C., Mardinoglu, A., and Ponten, F. (2017). A pathology atlas of the human cancer transcriptome. *Science*, 357(6352).

Valadares Barroso, G., Puzovic, N., and Dutheil, J. Y. (2018). The Evolution of Gene-Specific Transcriptional Noise Is Driven by Selection at the Pathway Level. *Genetics*, 208(January):173–189.

Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLoS Computational Biology*, 11(6).

Vallejos, C. A., Richardson, S., and Marioni, J. C. (2016). Beyond comparisons of means: Understanding changes in gene expression at the single-cell level. *Genome Biology*.

Vasiliauskaité, L., Berrens, R. V., Ivanova, I., Carrieri, C., Reik, W., Enright, A. J., and O'Carroll, D. (2018). Defective germline reprogramming rewires the spermatogonial transcriptome. *Nature Structural and Molecular Biology*.

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Hieter, P., Vogelstein, B., and Kinzler, K. W. (1997). Characterization of the yeast transcriptome. *Cell*.

Vitsios, D. M. and Enright, A. J. (2015). Chimira: Analysis of small RNA sequencing data and microRNA modifications. *Bioinformatics*.

Voit, R. and Grummt, I. (2011). The RNA polymerase i transcription machinery. *Protein Reviews*.

Wagner, W., Bork, S., Horn, P., Krunic, D., Walenda, T., Diehlmann, A., Benes, V., Blake, J., Huber, F. X., Eckstein, V., Boukamp, P., and Ho, A. D. (2009). Aging and replicative senescence have related effects on human stem and progenitor cells. *PLoS ONE*.

Wakiyama, M., Imataka, H., and Sonenberg, N. (2000). Interaction of elF4G with poly(A)-binding protein stimulates translation and is critical for Xenopus oocyte maturation. *Current Biology*, 10(18):1147–1150.

Walboomers, J. M. M., Jacobs, M. V., Manos, M. M., Bosch, F. X., Kummer, J. A., Shah, K. V., Snijder, P. J. F., Peto, J., Meijer, C. J. L. M., and Munzo, N. (1999). Human Papillomavirus Is a Necessary Cause. *Journal of pathology*, 19(189):12–19.

Wallace, N. A., Khanal, S., Robinson, K. L., Wendel, S. O., Messer, J. J., and Galloway, D. A. (2017). High-Risk Alphapapillomavirus Oncogenes Impair the Homologous Recombination Pathway. *Journal of Virology*, 91(20):1–22.

Walsh, C. P., Chaillet, J. R., and Bestor, T. H. (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation [4].

Wang, P., Doxtader, K. A., and Nam, Y. (2016a). Structural Basis for Cooperative Function of Mettl3 and Mettl14 Methyltransferases. *Molecular Cell*.

Wang, X., Feng, J., Xue, Y., Guan, Z., Zhang, D., Liu, Z., Gong, Z., Wang, Q., Huang, J., Tang, C., Zou, T., and Yin, P. (2016b). Structural basis of N6-adenosine methylation by the METTL3-METTL14 complex. *Nature*.

Wang, X., Lu, Z., Gomez, A., Hon, G. C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., Ren, B., Pan, T., and He, C. (2013). N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*, 505(7481):117–120.

Wang, X., Zhao, B. S., Roundtree, I. A., Lu, Z., Han, D., Ma, H., Weng, X., Chen, K., Shi, H., and He, C. (2015). N6-methyladenosine modulates messenger RNA translation efficiency. *Cell*, 161(6):1388–1399.

Wang, Y., Li, Y., Toth, J. I., Petroski, M. D., Zhang, Z., and Zhao, J. C. (2014). N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nature Cell Biology*, 16(2):191–198.

Warda, A. S., Kretschmer, J., Hackert, P., Lenz, C., Urlaub, H., Höbartner, C., Sloan, K. E., and Bohnsack, M. T. (2017). Human mettl16 is a n6-methyladenosine (m6a) methyltransferase that targets pre-mrnas and various non-coding rnas. *EMBO reports*, 18(11):2004–2014.

Watson, J. and Crick, F. (1953). A structure for deoxyribose nucleic acid. *Nature*, 171:737–738.

Watson, J. D. (1965). Molecular biology of the gene. *Molecular biology of the gene.*, (1st edn).

Weake, V. M. and Workman, J. L. (2010). Inducible gene expression: Diverse regulatory mechanisms.

Webster, M. W., Chen, Y. H., Stowell, J. A., Alhusaini, N., Sweet, T., Graveley, B. R., Coller, J., and Passmore, L. A. (2018). mRNA Deadenylation Is Coupled to Translation Rates by the Differential Activities of Ccr4-Not Nucleases. *Molecular Cell*.

Wei, C. M., Gershowitz, A., and Moss, B. (1975). Methylated nucleotides block 5? terminus of HeLa cell messenger RNA. *Cell*.

Wen, J., Lv, R., Ma, H., Shen, H., He, C., Wang, J., Jiao, F., Liu, H., Yang, P., Tan, L., Lan, F., Shi, Y. G., He, C., Shi, Y., and Diao, J. (2018). Zc3h13 Regulates Nuclear RNA m6A Methylation and Mouse Embryonic Stem Cell Self-Renewal. *Molecular Cell*.

White-Cooper, H. and Bausek, N. (2010). Evolution and spermatogenesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1546):1465–1480.

Wick, R. R., Judd, L. M., and Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*.

Willis, A., Jung, E. J., Wakefield, T., and Chen, X. (2004). Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. *Oncogene*, 23(13):2330–2338.

Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., and Andrews, S. (2015). HiCUP: pipeline for mapping and processing Hi-C data [version 1; referees: 2 approved, 1 approved with reservations]. *F1000Research*, 1310(May):1–12.

Woese, C. R. (1967). The genetic code: the molecular basis for genetic expression. Technical report.

Wolf, J. and Passmore, L. A. (2014). MRNA deadenylation by Pan2-Pan3. *Biochemical Society Transactions*.

Wood, J. G. and Helfand, S. L. (2013). Chromatin structure and transposable elements in organismal aging.

Wood, J. G., Jones, B. C., Jiang, N., Chang, C., Hosier, S., Wickremesinghe, P., Garcia, M., Hartnett, D. A., Burhenn, L., Neretti, N., and Helfand, S. L. (2016). Chromatin-modifying genetic interventions suppress age-associated transposable element activation and extend life span in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*.

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Zuzarte, P. C., Gilpatrick, T., Razaghi, R., Quick, J., Sadowski, N., Holmes, N., de Jesus, J. G., Jones, K. L., Snutch, T. P., Loman, N., Paten, B., Loose, M., Simpson, J. T., Olsen, H. E., Brooks, A. N., Akeson, M., and Timp, W. (2018). Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv*.

Wu, S., Li, K., Li, Y., Zhao, T., Li, T., Yang, Y. F., and Qian, W. (2017). Independent regulation of gene expression level and noise by histone modifications. *PLoS Computational Biology*.

Xiang, J. F., Yang, Q., Liu, C. X., Wu, M., Chen, L. L., and Yang, L. (2018). N6-Methyladenosines Modulate A-to-I RNA Editing. *Molecular Cell*.

Xiao, W., Adhikari, S., Dahal, U., Chen, Y.-S., Hao, Y.-J., Sun, B.-F., Sun, H.-Y., Li, A., Ping, X.-L., Lai, W.-Y., Wang, X., Ma, H.-L., Huang, C.-M., Yang, Y., Huang, N., Jiang, G.-B., Wang, H.-L., Zhou, Q., Wang, X.-J., Zhao, Y.-L., and Yang, Y.-G. (2016). Nuclear m 6 A Reader YTHDC1 Regulates mRNA Splicing. *Molecular Cell*, 61(4):507–519.

Xu, B., Chotewutmontri, S., Wolf, S., Klos, U., Schmitz, M., Dürst, M., and Schwarz, E. (2013). Multiplex Identification of Human Papillomavirus 16 DNA Integration Sites in Cervical Carcinomas. *PLoS ONE*, 8(6).

Xu, C., Wang, X., Liu, K., Roundtree, I. A., Tempel, W., Li, Y., Lu, Z., He, C., and Min, J. (2014). Structural basis for selective binding of m6A RNA by the YTHDC1 YTH domain. *Nature Chemical Biology*, 10(11):927–929.

Xu, K., Lin, J., Zandi, R., Roth, J. A., and Ji, L. (2016). MicroRNA-mediated target mRNA cleavage and 3?-uridylation in human cells. *Scientific Reports*, 6(July):1–14.

Xuan, J. J., Sun, W. J., Lin, P. H., Zhou, K. R., Liu, S., Zheng, L. L., Qu, L. H., and Yang, J. H. (2018). RMBase v2.0: Deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Research*.

Yaffe, E. and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, 43(11):1059–1065.

Yamashita, A., Chang, T. C., Yamashita, Y., Zhu, W., Zhong, Z., Chen, C. Y. A., and Shyu, A. B. (2005). Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover. *Nature Structural and Molecular Biology*, 12(12):1054–1063.

Yang, Z., Li, J., Feng, G., Gao, S., Wang, Y., Zhang, S., Liu, Y., Ye, L., Li, Y., and Zhang, X. (2017). MicroRNA-145 Modulates N6-methyladenosine levels by targeting the 3'-untranslated mRNA Region of the N6-Methyladenosine Binding YTH domain family 2 protein. *Journal of Biological Chemistry*, 292(9):3614–3623.

Yekta, S., Shih, I. H., and Bartel, D. P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. *Science*, 304(5670):594–596.

Yi, H., Park, J., Ha, M., Lim, J., Chang, H., and Kim, V. N. (2018). PABP Cooperates with the CCR4-NOT Complex to Promote mRNA Deadenylation and Block Precocious Decay. *Molecular Cell*, 70(6):1081–1088.e5.

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*.

Youssef, K. K., Lapouge, G., Bouvrée, K., Rorive, S., Brohée, S., Appelstein, O., Larsimont, J. C., Sukumaran, V., Van De Sande, B., Pucci, D., Dekoninck, S., Berthe, J. V., Aerts, S., Salmon, I., Del Marmol, V., and Blanpain, C. (2012). Adult interfollicular tumour-initiating cells are reprogrammed into an embryonic hair follicle progenitor-like fate during basal cell carcinoma initiation. *Nature Cell Biology*, 14(12):1282–1294.

Yu, G. and He, Q. Y. (2016). ReactomePA: An R/Bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*.

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology*.

Yu, G., Wang, L. G., and He, Q. Y. (2015). ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, 31(14):2382–2383.

Yudkovsky, N., Logie, C., Hahn, S., and Peterson, C. L. (1999). Recruitment of the SWI/SNF chromatin remodeling complex by transcriptional activators. *Genes and Development*.

Zaccara, S., Ries, R. J., and Jaffrey, S. R. (2019). Reading, writing and erasing mRNA methylation. *Nature reviews. Molecular cell biology*.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., and Shirley, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*.

Zhao, B. S., Wang, X., Beadell, A. V., Lu, Z., Shi, H., Kuuspalu, A., Ho, R. K., and He, C. (2017). M6A-dependent maternal mRNA clearance facilitates zebrafish maternal-to-zygotic transition. *Nature*, 542(7642):475–478.

Zhao, J., Hyman, L., and Moore, C. (1999). Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiology and molecular biology reviews : MMBR*.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K. S., Singh, U., Pant, V., Tiwari, V., Kurukuti, S., and Ohlsson, R. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*.

Zheng, C., Lu, X., Hansen, J. C., and Hayes, J. J. (2005). Salt-dependent intra- and internucleosomal interactions of the H3 tail domain in a model oligonucleosomal array. *Journal of Biological Chemistry*.

Zheng, G., Dahl, J. A., Niu, Y., Fedorcsak, P., Huang, C. M., Li, C. J., Vågbø, C. B., Shi, Y., Wang, W. L., Song, S. H., Lu, Z., Bosmans, R. P., Dai, Q., Hao, Y. J., Yang, X., Zhao, W. M., Tong, W. M., Wang, X. J., Bogdan, F., Furu, K., Fu, Y., Jia, G., Zhao, X., Liu, J., Krokan, H. E., Klungland, A., Yang, Y. G., and He, C. (2013). ALKBH5 Is a Mammalian RNA Demethylase that Impacts RNA Metabolism and Mouse Fertility. *Molecular Cell*, 49(1):18–29.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*.

Zheng, Zhi Ming, Baker, C. C. (2008). Papillomavirus Genome Structure, Expression, and Post-Transcriptional Regulation. *Frontiers in Bioscience*, 11:2286–2302.

Zhou, J., Wan, J., Gao, X., Zhang, X., Jaffrey, S. R., and Qian, S. B. (2015). Dynamic m6A mRNA methylation directs translational control of heat shock response. *Nature*, 526(7574):591–594.

Zhu, T., Roundtree, I. A., Wang, P., Wang, X., Wang, L., Sun, C., Tian, Y., Li, J., He, C., and Xu, Y. (2014). Crystal structure of the YTH domain of YTHDF2 reveals mechanism for recognition of N6-methyladenosine. *Cell Research*, 24(12):1493–1496.

Zopf, C. J., Quinn, K., Zeidman, J., and Maheshri, N. (2013). Cell-Cycle Dependence of Transcription Dominates Noise in Gene Expression. *PLoS Computational Biology*.

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148.

zur Hausen, H. (2002). Papillomaviruses and cancer: from basic studies to clinical application. *Nature reviews. Cancer*, 2(5):342–50.