Deposit & Copying of Dissertation Declaration

UNIVERSITY OF CAMBRIDGE

Board of Graduate Studies

Please note that you will also need to bind a copy of this Declaration into your final, hardbound copy of thesis - this has to be the very first page of the hardbound thesis.

1 Surname (Family Name)	Forenames(s)	Title
STUBBS	THOMAS MICHAEL	MR

2 Title of Dissertation as approved by the Degree Committee
DNA METHYLATION : A MODEL SYSTEM FOR THE STUDY OF AGEING

In accordance with the University Regulations in *Statutes and Ordinances* for the PhD, MSc and MLitt Degrees, I agree to deposit one print copy of my dissertation entitled above and one print copy of the summary with the Secretary of the Board of Graduate Studies who shall deposit the dissertation and summary in the University Library under the following terms and conditions:

1. Dissertation Author Declaration

I am the author of this dissertation and hereby give the University the right to make my dissertation available in print form as described in 2. below.

My dissertation is my original work and a product of my own research endeavours and includes nothing which is the outcome of work done in collaboration with others except as declared in the Preface and specified in the text. I hereby assert my moral right to be identified as the author of the dissertation.

The deposit and dissemination of my dissertation by the University does not constitute a breach of any other agreement, publishing or otherwise, including any confidentiality or publication restriction provisions in sponsorship or collaboration agreements governing my research or work at the University or elsewhere.

2. Access to Dissertation

I understand that one print copy of my dissertation will be deposited in the University Library for archival and preservation purposes, and that, unless upon my application restricted access to my dissertation for a specified period of time has been granted by the Board of Graduate Studies prior to this deposit, the dissertation will be made available by the University Library for consultation by readers in accordance with University Library Regulations and copies of my dissertation may be provided to readers in accordance with applicable legislation.

3 Signature	Date
	10/05/18

Corresponding Regulation

Before being admitted to a degree, a student shall deposit with the Secretary of the Board one copy of his or her hard-bound dissertation and one copy of the summary (bearing student's name and thesis title), both the dissertation and the summary in a form approved by the Board. The Secretary shall deposit the copy of the dissertation together with the copy of the summary in the University Library where, subject to restricted access to the dissertation for a specified period of time having been granted by the Board of Graduate Studies, they shall be made available for consultation by readers in accordance with University Library Regulations and copies of the dissertation provided to readers in accordance with applicable legislation.

DNA methylation: a model system for the study of ageing

Thomas Michael Stubbs

Darwin College

This dissertation is submitted for the degree of Doctor of Philosophy

September 2017

Graduate School of Life Sciences

UNIVERSITY OF CAMBRIDGE

DNA methylation: A model system for the study of ageing

Thomas Michael Stubbs

DNA methylation is an important epigenetic mark spanning all of life's kingdoms. In humans, DNA methylation has been associated with a wide range of age-related pathologies, including type II diabetes and cancer. More recently, in humans, changes in DNA methylation at specific positions in the genome have been found to be predictive of chronological age. Interestingly, DNA methylation age is also predictive of health status and time-to-death. A better understanding of what these DNA methylation changes represent and whether they might be causative in the ageing process will be important to ascertain. However, at present there is no animal model system with which this process can be studied at a mechanistic level.

Furthermore, it is becoming increasingly apparent that many disease states that increase in prevalence with age are not caused by all cells within the individual, but are often the result of changes to a subset of cells. This underscores the importance of studying these processes at the single cell level. The recent advances in single cell sequencing approaches now mean that we can study multiple layers of biology within the same single cell, such as the epigenome and the transcriptome (scM&T-Seq). Unfortunately, we are still only able to probe these important aspects of single cell biology in a static sense. This is a major limitation in the study of ageing because ageing and age-related disease processes are inherently dynamic. As such, it is incumbent upon us to develop approaches to assay single cell biology in a dynamic manner.

In this thesis, I describe an epigenetic age predictor in the mouse. This predictor is tissue-independent and can accurately predict age (with an error of 3.33 weeks) and can record deviations in biological age upon interventions including ovariectomy and high fat diet both of which are known to reduce lifespan. Next, I describe the analysis of a homogeneous population of muscle satellite cells (MuSCs) that I have interrogated at the single cell level, using single cell combined transcriptome and methylome sequencing (scM&T-seq). I found that with age there was increased global transcriptional variability and increased feature-specific methylome variability. These findings explain the loss of functionality of these cells with age. Lastly, I describe two imaging approaches to study DNA methylation dynamically in single cells. Using these methods, I demonstrate that it is possible to accurately determine methylation status across a wide spectrum of global methylation levels and that by using such approaches novel information about dynamic methylation processes can be obtained. These methods represent the first to study DNA methylation dynamically.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface 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. This thesis does not exceed the word limit of 60,000 words required by the University of Cambridge School of Biological Sciences.

Thomas M. Stubbs
September 2017

Table of Acknowledgments of Assistance

Data/materials provided by someone else:

- TA-Hi MuSC single cells from young and old individuals were isolated and sorted by Dr. Brendan Evano
- IF staining was conducted by Dr. Fatima Santos

Data/materials produced jointly:

- Mouse samples were collected and processed jointly with Dr Ferdinand von Meyenn and Dr Anne-Katrien Stark.
- UMI-hunter was defined in collaboration with Dr. Felix Krueger.
- Elastic-net regression models were derived in collaboration with Dr. Marc Jan Bonder.
- Primary DDM analysis was defined in collaboration with Dr. Simon Andrews and Dr. Andrew Harrison.
- Cell lines used in Chapter 5 were generated in collaboration with Dr. Melanie Eckersley-Maslin.
- FRAP imaging and primary analysis were conducted in collaboration with Kris Grenz and Dr Fatima Santos.
- DDM imaging was conducted in collaboration with Kris Grenz.

- Secondary DDM analysis was defined in collaboration with Dr. Andrew Harrison.
- Single cell libraries were generated in collaboration with Dr. Stephen Clark.
- Single cell analysis was conducted in collaboration with Dr. Irene Hernando Herraez.

Data obtained from technical service provider:

- Next generation sequencing were done by Wellcome Trust Sanger Institute and Babraham Institute NGS facility.
- Mice were bred for ageing purposes by the BI ageing clock team.
- Images were taken using microscopes that are maintained by the Imaging Facility.

Summary

DNA methylation is an important epigenetic mark spanning all of life's kingdoms. In humans, DNA methylation has been associated with a wide range of age-related pathologies, including type II diabetes and cancer. More recently, in humans, changes in DNA methylation at specific positions in the genome have been found to be predictive of chronological age. Interestingly, DNA methylation age is also predictive of health status and time-to-death. A better understanding of what these DNA methylation changes represent and whether they might be causative in the ageing process will be important to ascertain. However, at present there is no animal model system with which this process can be studied at a mechanistic level.

Furthermore, it is becoming increasingly apparent that many disease states that increase in prevalence with age are not caused by all cells within the individual, but are often the result of changes to a subset of cells. This underscores the importance of studying these processes at the single cell level. The recent advances in single cell sequencing approaches now mean that we can study multiple layers of biology within the same single cell, such as the epigenome and the transcriptome (scM&T-Seq). Unfortunately, we are still only able to probe these important aspects of single cell biology in a static sense. This is a major limitation in the study of ageing because ageing and age-related disease processes are inherently dynamic. As such, it is incumbent upon us to develop approaches to assay single cell biology in a dynamic manner.

In this thesis, I describe an epigenetic age predictor in the mouse. This predictor is tissue-independent and can accurately predict age (with an error of 3.33 weeks) and can record deviations in biological age upon interventions including ovariectomy and high fat diet both of which are known to reduce lifespan. Next, I describe the analysis of a homogeneous population of muscle satellite cells (MuSCs) that I have interrogated at the single cell level, using single cell combined transcriptome and methylome sequencing (scM&T-seq). I found that with age there was increased global transcriptional variability and increased feature-specific methylome variability. These findings explain the loss of functionality of these cells with age. Lastly, I describe two imaging approaches to study DNA methylation dynamically in single cells. Using these methods, I demonstrate that it is possible to accurately determine methylation status across a wide spectrum of global methylation levels and that by using such approaches novel information about dynamic methylation processes can be obtained. These methods represent the first to study DNA methylation dynamically.

Acknowledgements

No PhD is conducted in isolation, and the work conducted herein is no exception. I could not have accomplished anything without the help and support of many individuals. Individuals that up until now have never been thanked formally in black and white.

First and foremost, I am incredibly grateful to my PhD supervisor Wolf Reik for the opportunities and support he has given me. I will never forget the intellectual freedom he has entrusted me with over the past three years.

Next, I would like to express how grateful I am to “the inmates” Wendy Dean and Fatima Santos. Thank you both for your thought-provoking discussions, and your selfless kindness.

To Melanie Eckersley-Maslin, it’s been a roller-coaster and I’m sure it wasn’t easy having me as your first mentee, but I want to thank you for all the help and support you have given me over the time I have spent in the lab.

To my collaborators: Daniel Herranz, Marc Jan Bonder, Brendan Evano and Andrew Harrison, thank you for opening new doors to exciting science and providing me with fresh perspectives on the world of epigenetics.

To the wolf pack past and present, thank you! I cannot express my gratitude enough to each and every one of you. From acquiring an astute knowledge of niche Portuguese music, to arguments about the correct name for “stubbies” or “tiny beers”; you have all made my time in the lab that much more enjoyable. In particular, it would be remis of me not to highlight a number of individuals for special thanks. Ines “the miracle”

Milagre, thank you for all your kind support. I don't think I will ever meet anyone so selfless, fun and driven. You have been and still are an inspiration to me. Ferdinand "foidy" von Meyenn, it has been a pleasure to have had the chance to work alongside you, however fleetingly. Irene Herraez, analysis with you will never be boring. Thank you for showing me how much fun computational work can be. Poppy "PJ" Gould a.k.a. "Beers" Gould, I could not have asked for a better PhD companion to have shared this journey with me. I don't know what I would have done without your friendship in the lab. Lastly, Stephen "single-cell" Clark, the daily/weekly midday method development meetings sponsored by The George will be among my fondest memories of my time in the lab.

I would also like to thank those individuals who edited and proof-read this thesis. Namely, Aartjan te Velthuis, Melanie Eckersley-Maslin, Julia Spindel and Wolf Reik.

I would like to thank the BBSRC for the funding that supported this research.

Moving away from the lab and into Cambridge, I would like to thank those who incessantly reminded me that there's a whole new and exciting world just outside of the confines of the Babraham security barriers.

I would also like to thank my family and friends at home for their support and encouragement, without which I am not sure that I would have ever made it to the stage of writing a thesis. A special thanks to the tripod members: Charles Ball and James Whitecross, I think we can finally get blindo.

This thesis is a reflection of these individuals and what they have inspired, so for those of you bold enough to read on, I hope you enjoy it.

Thank you.

Publications & Patents

0.1 Publications

Clark SJ, Argelaguet R, Kapourani CA, **Stubbs TM**, Lee HJ, Alda-Catalinas C, Krueger F, Sanguinetti G, Kelsey G, Marioni JC, Stegle O, Reik W. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat Commun. 2018 Feb 22; 9(1):781. doi: 10.1038/s41467-018-03149-4.

Martin-Herranz DE, Ribeiro AJM, Krueger F, Thornton JM, Reik W, **Stubbs TM**. cuRRBS: simple and robust evaluation of enzyme combinations for reduced representation approaches. Nucleic Acids Research, gkx814, https://doi.org/10.1093/nar/gkx814.

Stubbs TM, Bonder MJ, Stark AK, Krueger F; BI Ageing Clock Team, von Meyenn F, Stegle O, Reik W. Multi-tissue DNA methylation age predictor in mouse. Genome Biol. 2017 April 11; 18(1):68. doi: 10.1186/s13059-017-1203-5

Hahn O, Grönke S, **Stubbs TM**, Ficz G, Hendrich O, Krueger F, Andrews S, Zhang Q, Wakelam MJ, Beyer A, Reik W, Partridge L. Dietary restriction protects from age-associated DNA methylation and induces epigenetic reprogramming of lipid metabolism. Genome Biol. 2017 Mar 28; 18(1):56. doi: 10.1186/s13059-017-1187-1.

Milagre I, **Stubbs TM**, King MR, Spindel J, Santos F, Krueger F, Bachman M, Segonds-Pichon A, Balasubramanian S, Andrews SR, Dean W, Reik W. Gender Differences in Global but Not Targeted Demethylation in iPSC Reprogramming. Cell Rep. 2017 Jan 31; 18(5):1079-1089. doi: 10.1016/j.celrep.2017.01.008.

Eckersley-Maslin MA, Svensson V, Krueger C, **Stubbs TM**, Giehr P, Krueger F, Miragaia RJ, Kyriakopoulos C, Berrens RV, Milagre I, Walter J, Teichmann SA, Reik W. MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs. Cell Rep. 2016 Sep 27; 17(1):179-92. doi: 10.1016/j.celrep.2016.08.087.

Stubbs TM, Te Velthuis AJ. The RNA-dependent RNA polymerase of the influenza A virus. Future Virol. 2014 Sep;9(9):863-876.

0.2 Patents

Stubbs TM, 2018 Feb 9, Novel Age Calculation Method, PCT/GB2018/050362.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Ageing and the hope of reversing it have fascinated humans around the globe for centuries. More recently however, the study of ageing has evolved from a curiosity to an integral area of research in order to meet the demands of the currently ageing population. But in spite of recent effort to understand the process of ageing, our understanding of it is still limited and many fundamental questions remain. Such questions include: what is ageing? Is ageing a selected or adaptive trait? Is ageing a multicellular ensemble phenomenon or does it exist as a concept at the level of the single cell? And what is the molecular definition of ageing and can we manipulate it in order to reverse it? Definitions abound to try and describe ageing at different complexity levels, from physiological definitions to molecular ones. In this thesis, I define ageing broadly as a reduction in a cell or organism's ability to efficiently interact with its environment for its own propagation.

1.1 Theories of ageing

In the late 19th century, August Weismann proposed that ageing was the imperfect adaptation of an individual to the detrimental injuries that occurred over the course of their lifetime and that this imperfect adaptive process of ageing would in turn "make room for the young" (Weismann, 1889). However, this concept of group selection-based adaptive ageing lost favour with evolutionary biologists due to its teleological narrative, which did not offer a clear mechanism for the selection process. It was therefore replaced by the Mutation Accumulation Theory of ageing. This theory postulates that non-adaptive ageing is due to the reduced power of natural selection at advanced age. This results in an increased burden of deleterious mutations that specifically effect the aged (Medawar, 1952). This theory suggests that late-acting deleterious mutations in the germline are responsible for the ageing phenomenon. It has since lost favour, because of the observation that many genes are expressed in a tissue-specific, and developmental stage-specific manner. As such, any late-life specific effect would require a late-life specific genetic program to be activated in the first instance.

A theory that sought to rectify this discrepancy was of the theory of Antagonistic Pleiotropy. This theory, also proposed under the umbrella of non-adaptive ageing, differed from the Mutation Accumulation Theory of ageing in its assessment of the influence of natural selection upon the formation of the ageing phenotype itself. According to this theory, natural selection acts strongly on genes that are beneficial early in life, even if they will have detrimental effects in mature individuals (Hamilton, 1966, Kirkwood and Melov, 2011 and Williams, 2001). This theory of ageing thus gives natural selection an integral role in the ageing process. Moreover, the theory is not considered an adaptive ageing theory, because the selection pressure is on the phenotypic traits that are beneficial during early life. As such, this theory hypothesises that ageing is not directly caused by a genetic mechanism.

However, there is increasing observational and experimental evidence that calls this Antagonistic Pleiotropy theory of ageing into question. For instance, it has been shown in fruit flies that were selectively bred for a long lifespan that a long lifespan is not correlated with a concomitant decrease in reproductive fitness (Leroi et al., 1994), as would be expected from Antagonstic Pleiotropy. In addition, it has been found that certain ageing-related diseases have conserved genetic mechanisms that need to be activated for the disease to occur (Bowles, 2000) and that such genetic mechanisms don't need to have a beneficial effect early in life (Brack et al., 2007, Goldsmith, 2006, Kenyon, 2010, Khaidakov et al., 2006, Linnane et al., 1990, Mitteldorf, 2004, Pishel et al., 2012 and Thum et al., 2008).

The last of the three major non-adaptive theories of ageing is the Disposable Soma Theory. This theory posits that there is a finite amount of resources that an individual has at their disposal, whether this resource be energy (as originally proposed by Kirkwood, 1977) or time (as an adaptation of the theory proposed by Lorenzini et al., 2011), and that these resources must be distributed among the cells of the body to maximise successful procreation. From this it follows that a far greater proportion of resources ought to be given to the cells of the germline than to those of the soma. In turn this would predict that a reduction in available resources would result in a reduction in lifespan and reproduction, assuming that the distribution ratio does not change. Indeed, a reduction in caloric consumption reduces the reproductive output in caloric restriction studies. However, caloric restriction experiments conducted since the 1930s do not show that a reduction in energy input reduces lifespan, in fact it has been shown to increase lifespan (Fontana et al., 2010), depending on the species, strains and sexes studied (Colman et al., 2014, Sohal and Forster, 2014 and Swindell, 2012). This suggests that this simple assumption of a maintained germline-soma ratio is not reflective of the reality. Perhaps instead, individuals are able to respond to short term "hardship" in order to ensure successful procreation is feasible in the more "affluent" long-term.

One characteristic that these non-adaptive ageing theories all share is the prediction that faster ageing will evolve when there is a higher extrinsic death rate. In other words, when selection cares less about the old due to an increased extrinsic death rate, mutations that are prohibitive to long life would be allowed to survive within the population. But this idea does not seem to be supported by evidence: guppies living in regions of higher predation do not have an increased intrinsic death rate due to aging (Reznick et al., 2004). In an attempt to account for such contradictory observations, adaptations to these classical non-adaptive theories of ageing have been proposed. One such adaptation is the Theory of Robustness (Kriete, 2013). In addition to the proposal of adaptations to these theories, these contradictions have also resulted in the re-popularisation of adaptive ageing theories, as first proposed by Weismann and derivatives thereof (Bowles, 1998, Goldsmith, 2004, Goldsmith, 2008, Martins, 2011, Mitteldorf, 2006, Mitteldorf and Pepper, 2009, Skulachev, 2001, Travis, 2004 and Woodberry et al., 2007). This is in large part due to the realisation that adaptive theories of ageing are compatible with both the evidence for and contradictory evidence against these non-adaptive theories of ageing (Goldsmith, 2006, Mitteldorf, 2004 and Mitteldorf, 2010). In addition, these new adaptive theories of ageing are also compatible with known individually-adverse behaviours present in nature. Such behaviours include: altruism, sexual reproduction and suicidal behaviour in semelparous species.

Modern adaptive theories of ageing assume that evolution is acting at a group or kinship level and incorporate ageing as an altruistic component (Mitteldorf and Wilson, 2000, Taylor, 1992, Wilson et al., 1992 and Yang, 2013). This description of adaptive ageing within a group or kinship selection context has been widely criticised owing to interpretations from analyses of classical evolutionary theories, including: Price's Theorem (Price et al., 1970), Evolutionarily Stable Strategy theory (Smith and Parker, 1976), and Kin Selection theory (Hamilton, 1964). However, these theories are limited owing to their failure to adequately take into account differences between individuals within a population and the impact of population dynamics. As such, analyses based on them are inherently over-simplistic and biased. Such limitations are now becoming more widely appreciated, which makes it increasingly apparent that ageing *per se* may be actionable by evolution when considered as a kinship attribute within a viscous population (Mitteldorf and Wilson, 2000, Mitteldorf, 2006). A viscous population is defined as one with limited movement that in turn results in increased genetic relatedness of individuals within it (Mitteldorf and Wilson, 2000). That being said, for such kinship-based ageing adaptations to develop, there must be a dynamic carrying capacity within the population, otherwise the kinship benefits are negated by the increased pressure of the attribute (Taylor, 1992 and Wilson et al., 1992). Excitingly, in recent years these kinship theories have been modified to incorporate kinship selection at the level of the individual (within viscous populations; Yang, 2013). This study considered all offspring to be non-equivalent, and further highlighted the potential for the evolution of ageing as a kinship attribute, by showing that the evolution of "programmed death" requires a less cooperative environment than does the evolution of gender (Yang, 2013).

1.2 Ageing at the molecular level

At the level of the individual, the above theories of ageing can be broadly split into two categories: those implicating passive decay, potentially as a result of damage (for example somatic mutations in the case of the Disposable Soma Theory); or those that involve an actively procured state, either in a non-programmed manner such as a misfiring genetic program due to a mutation (as in Antagonistic Pleiotropy Theory) or programmed (as in adaptive theories of ageing). However, it must be noted that, at the molecular level, the characteristics of these various processes are by no means mutually exclusive. In fact, it is often challenging to discern what the differences would be between the evolutionary theories at the molecular level. For instance, one could imagine that there existed a "program" that resulted in the accumulation of somatic mutations responsible for the ageing process, and that this phenomenon would not necessarily require an independent "ageing program". For the purposes of understanding whether or not the ageing process could be stalled or reversed, it is perhaps better to understand, firstly, its molecular mechanism and then, secondarily, the evolutionary rationale (if this existed) behind it. This is because the feasibility of stalling or reversing the aged-state will be primarily due to the molecular causes of the aged-state in the first place and not necessarily whether it was evolutionarily generated or not.

Another interesting question that remains largely unanswered is whether ageing is a phenomenon that can be described at the single cell level or whether it is an ensemble property of a multi-cellular organism. In other words, could the phenomenon of ageing be described purely from the view of a single cell? Could a cell in and of itself be described as aged, or can the ageing phenomenon only be considered in an organismal context? Studies conducted on single-celled organisms such as S. cerevisiae have attempted to address this point, but so far it has remained unclear whether the mother cell is truly ageing in the same sense as we define it in a human setting, or whether it is simply acquiring a new state that serves an evolutionary purpose within a population (Frenk et al., 2017).

The evidence available for answering these questions is in most instances circumstantial. This is because a significant portion of the available evidence comes from observations of the natural ageing process (either in a laboratory setting or in the field) and not from hypothesis-driven experiments. Another reason is that ageing is intrinsically complex and time-consuming to study in an experimental setting. In recent years, some evidence collected at both the organismal and cellular level has emerged that strengthens the idea that ageing could be an actively procured state. For instance, from observational studies evidence for the actively procured state has come from organism-level observations such as: the existence of semelparous organisms like *Oncorhynchus spp.*; the huge variability in lifespans of evolutionarily close species, such as in whales; and the existence of vast lifespans in the kingdom *Plantae*, such as that of *Pinus longaeva*, which is reported to live thousands of years (Crespi and Teo, 2002, Foote, 2008 and Schulman et al., 1956). From experimental studies, evidence in support of the actively procured state has come from reprogramming, rejuvenation, senescence and immortalisation experiments that have been conducted both at a cellular level *in vitro*, and *in vivo*. And lastly, in the case of reprogramming experiments, it has recently been shown that differentiated cell types isolated from young and old individuals can be reprogrammed to an induced pluripotent stem cell (iPSC) state and then re-differentiated, at which point they both have the same "youthful" phenotype. In this sense the aged phenotype appears molecularly reversible (Frobel et al., 2014). This work also suggests that ageing as defined in this context is a cell-intrinsic property that can be reset during the iPSC reprogramming process. It has been noted sporadically in the literature that older cells are more refractory to the iPSC reprogramming process and more prone to abnormalities (Frobel et al., 2014, Sardo et al., 2017 and Takahashi and Yamanaka, 2006). However, it is accepted in the field that the reprogramming efficiency of any one cell type is dependent on a range of factors including the donor genotype, as such, it is yet to be determined whether this idea of age-related inefficiency would hold true in studies with larger donor sample sizes (Ebrahimi, 2015).

In addition to observations on *in vitro* cellular reprogramming, scientists have demonstrated the rejuvenating effect of reprogramming *in vivo*. Here, researchers cyclically induced expression of the Yamanaka factors (Oct4, Sox2, Klf4, c-Myc; Mosteiro et al., 2016 and Ocampo et al., 2016) in mice using doxycycline. The expression of these factors resulted in the rejuvenation of multiple tissues, with a correlation between the extent of rejuvenation and the degree of reprogramming in the specific tissue (Mosteiro et al., 2016). The pancreas is an example of a tissue within which reprogramming is highly efficient (Mosteiro et al., 2016). In addition, it has been shown in a progeria mouse model that this cyclical reprogramming protocol can extend lifespan (Ocampo et al., 2016). However, it should be noted that such a lifespan extension in a progeria model might not be surprising, given that fibroblasts from progeria patients lost the hallmarks of the disease for several passages when they were reprogrammed to iPSCs and subsequently re-differentiated (Liu et al., 2011). Furthermore, experiments using cyclical reprogramming in wild type mice have shown no such lifespan extension (M. Serrano personal communication).

Mechanistically, the rejuvenation that occurs during the cyclical protocol is thought to be due to the removal of aged cells not cleared properly by the immune system prior to the start of the reprogramming protocol. These aged cells are thought to be cleared during the removal of the vast numbers of senescent cells induced during the reprogramming process itself. Senescent cells are defined as cells that have irreversibly ceased cellular replication (Hayflick, 1965). This rejuvenation is then further compounded by replacement of the aged cells with phenotypically young descendants of the *in vivo* iPSCs generated during the reprogramming. The findings from these *in vivo* reprogramming experiments are suggestive of an actively procured state, with the iPSCs having had their state reset. However, the removal of senescent cells from tissues would also be consistent with a passive decay process.

The phenomenon of cellular senescence was first observed by Hayflick in human fibroblasts, in which he observed a limit to the number of passages that a cell can undergo before it ceases to divide (Hayflick and Moorhead, 1961). This limit has since become known as the Hayflick limit (Hayflick, 1965). It is caused by maximal telomere erosion, and the induction of a DNA damage response, and it has therefore been termed replicative senescence. Cellular senescence can also be induced in a telomere-independent manner, for instance through oncogenic induction or in response to DNA damage (Di Leonardo et al., 1994, Lin et al., 1998 and Serrano et al., 1997). Two typically defined characteristics of senescent cells are their expression of senescence-associated beta-galactosidase (SA-β-Gal) and cyclin-dependent kinase inhibitor 2A (p16$^{\text{INK4A}}$). p16 expression is commonly seen in benign cancers, but it appears to be lost in malignant samples, suggesting that in this setting senescence is tumour-suppressive (Braig et al., 2005, Haugstetter et al., 2010 and Michaloglou et al., 2005). In addition, these cells have a characteristic secretome termed Senescence Associated Secretory Phenotype (SASP), which contains inflammatory cytokines, growth factors and proteases (Acosta et al., 2008, Coppé et al., 2008 and Kuilman et al., 2008). This SASP also induces senescence in neighbouring cells (Acosta et al., 2013).

SASP is relevant to the ageing field because it contributes to a number of ageing-related diseases such as atherosclerosis, cancer and type 2 diabetes through modulation of the immune system (Tchkonia et al., 2013). Interestingly, although controversially, it has been suggested that this senescent state could be reversible (Beauséjour et al., 2003). This study proposes that the cause of senescence, whether by p53 inactivation or pRB inactivation for instance, will determine how the senescence barrier can be overcome. This study alongside others highlights two things: firstly that ageing, when considered as a senescent state, is a cell-intrinsic phenomenon (although extrinsic factors are hugely influential in the formation of the state); and secondly that the ageing process as considered in this way is potentially reversible. However, it should be noted that although senescence can be overcome, this reversal process is commonly associated with chromatin defects such as ploidy and could potentially be one of the ways in which cancers are able to evolve (Beauséjour et al., 2003).

Progeroid Syndromes are a family of rare genetic diseases characterised by the presence of ageing phenotypes, from where they derive their name. The most commonly studied Progeroid Syndromes within the field of ageing are Werner Syndrome (“adult progeria”) and Hutchinson-Gilford Progeria Syndrome (HGPS; “progeria”), because they are believed to represent aspects of the "natural" ageing process. Werner Syndrome is an autosomal recessive disease characterised in the majority of cases by a mutation in the gene WRN. WRN encodes a protein with strong homology to recQ helicases and is thought to be required for unwinding DNA during DNA repair and replication (Gray et al., 1997). HGPS, conversely, is a rare genetic disease that is most commonly the result of a *de novo* C-to-T mutation in the LMNA gene at position 1824 in the coding sequence (found in exon 11), which results in the incorporation of a cryptic donor splice site into the transcript and a fifty-amino acid truncation of the Lamin A protein (De Sandre-Giovannoli et al., 2003). This truncation results in incorrect processing of the prelamin A protein within the nucleus, which in turn affects the nuclear lamina (De Sandre-Giovannoli et al., 2003).

Studies of cells derived from people with HGPS, have provided us with insights into the nature of ageing from a senescence perspective. HGPS patient-derived fibroblasts exhibit premature cellular senescence alongside multiple other nuclear and chromatin abnormalities, such as blebbing of the nuclear envelope and reduced telomere length (Allsopp et al., 1992, Decker et al., 2009, Goldman et al., 2004, Huang et al., 2008, Liu et al., 2005, Liu et al., 2006 and Shumaker et al., 2006). Experiments reprogramming these to iPSCs showed that the lamin A-associated nuclear defects, alongside many of the other aberrant phenotypes, were absent in the HGPS-iPS cell lines but these nuclear defects returned upon re-differentiation (Liu et al., 2011). This study highlights that the progeria phenotype is reversible and that the hallmarks of senescence are reversible too. It also reveals that the nuclear structure and chromatin are important features in the induction of senescence and the propensity for its reversal.

This understanding of senescence has encouraged researchers to test its relevance to ageing *in vivo* and whether or not it can be experimentally manipulated in order to improve health or lifespan. Scientists have adopted several different approaches to assess this. One such approach was to utilise a genetically engineered mouse called the INK-ATTAC mouse (Baker et al., 2011). This mouse is genetically engineered to allow apoptosis induction in p16-expressing cells upon injection of AP20187. After biweekly induction of apoptosis in otherwise wild-type mice (from the first year onwards), it was shown that the median but not maximum life expectancy was extended independently of the genetic background or sex of the animal (Baker et al., 2011). Moreover, the researchers indicated that "removal of p16-positive cells delayed tumorigenesis and attenuated age-related deterioration of several organs without apparent side effects, including kidney, heart and fat, where clearance preserved the functionality of glomeruli, cardio-protective KATP channels, and adipocytes, respectively." (Baker et al., 2016). These results highlight that a subpopulation of senescent cells can have a detrimental impact on an organism as a whole. This is consistent with both the passive decay theories of ageing and ideas ascribing to an active process.

In addition to reprogramming and senescence experiments, parabiosis experiments have provided insight into the nature of ageing and ways in which certain age-related symptoms, such as cardiac hypertrophy, can be alleviated in mammals (Carlson et al., 2008, Conboy et al., 2005, Demontis et al., 2014, Elabd et al., 2014, Horrington et al., 1960, Katsimpardi et al., 2014, Loffredo et al., 2013, McCay et al., 1957, Ruckh et al., 2012, Villeda et al., 2011, Villeda et al., 2014, Wagers et al., 2002 and Wright et al., 2001). Parabiosis experiments involve the joining of the circulatory systems of two individuals. In the case of ageing research, parabiosis experiments are conducted by joining old (typically more than 20 months of age) and young (typically 3 months of age) individuals together (heterochronic) and comparing them to control pairings (isochronic - old/old and young/young pairings). Experiments conducted as such have shown that many of the ageing phenotypes, for instance hypertrophy of the heart and thinning of the epidermis, can be reversed in old mice (Carlson et al., 2008, Conboy et al., 2005, Demontis et al., 2014, Elabd et al., 2014, Horrington et al., 1960, Katsimpardi et al., 2014, Loffredo et al., 2013, McCay et al., 1957, Ruckh et al., 2012, Villeda et al., 2011, Villeda et al., 2014, Wagers et al., 2002 and Wright et al., 2001) . However, it should be noted that not all ageing phenotypes in all tissues are rejuvenated. Importantly, this rejuvenation phenomenon has been validated in settings where cell transfer between the two parabiosed individuals has been inhibited, highlighting that it is factors within the blood itself that are responsible for the rejuvenation. One such identified factor, although it remains controversial, is growth differentiation factor 11 (GDF-11;

Loffredo et al., 2013 and Sinha et al., 2014). The findings from these parabiosis and factor-intervention experiments are consistent with aging being the result of an active process. These observations of rejuvenation would be more difficult to explain through a passive decay process.

Interestingly, the young mice used in these parabiosis experiments appear to show an increasingly aged phenotype concurrent with the rejuvenation of the older animals (Katsimpardi et al., 2014). This suggests that, at least in these crude experiments, ageing phenotypes are malleable in both directions. In addition, it raises the tantalising possibility that the process responsible for the one phenomenon may also underlie the other. In other words, perhaps there are finite amounts of "youthful" factors whose concentrations are reduced upon the joining of the old and young individuals, resulting in rejuvenation in the one and ageing in the other.

At present, it is still not clear why certain tissues become rejuvenated during parabiosis experiments whilst others do not, nor how the rejuvenating effects are enacted at the molecular level. This is largely due to the costly, time-consuming and technically challenging nature of the experiments. However, some factors have been identified that could play a role, such as GDF-11 and oxytocin (Elabd et al., 2014, Loffredo et al., 2013 and Sinha et al., 2014). In addition, the rejuvenation of skeletal muscle appears to be partly due to rejuvenation of the stem cell niche environment, which in turn results in improved functionality of resident muscle satellite cells (Conboy et al., 2005).

1.3 Epigenetic observations

1.3.1 Definition of epigenetics

Epigenetics was first defined by Conrad Waddington in the early 1940s as "the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being" (Waddington, 1953). For the purposes of this thesis, epigenetics is defined as: the mechanism by which heritable changes in gene expression can be derived, without alterations to the underlying DNA sequence. Epigenetics plays a key role in cell fate determination and differentiation (Zhu et al., 2013), with aberrant modifications present in a wide range of diseases. For instance, some cancers contain epigenetic but no genetic abnormalities (Mack et al., 2014 and Versteeg, 2014). Additionally, dosage effects such as X-chromosome inactivation (Barakat and Gribnau, 2010) and genetic imprinting (Magenis et al., 1987) are also thought to be controlled by epigenetic changes. In order for a modification to be classified as an epigenetic mark it must be self-propagating, replicative and resulting in a phenotype. This thesis will predominantly discuss one specific type of epigenetic mechanism/modification that is relevant in the context of aging: DNA methylation.

1.3.2 Epigenetics

Many epigenetic changes occur during ageing (López-Otín et al., 2013). These include changes in modifications of histones, non-coding RNAs and modifications to the DNA itself. Core histones are octameric protein complexes that are formed from four protein subunits: H2A, H2B, H3 and H4 (Luger et al., 1997). These subunits are arranged in two H2A-H2B dimers and a H3-H4 tetramer (Luger et al., 1997). These octameric complexes are referred to as nucleosomes and each can bind 147 base pairs of DNA (Richmond and Davey, 2003). The histone subunits can be chemically modified by the addition of acetyl and methyl groups among other modifications (Jenuwein and Allis, 2001). In turn, these modifications recruit chromatin remodelling proteins and thus alter the chromatin state. Two such marks are methylation of lysine 4 of H3 (H3K4me), associated with active chromatin, and trimethylation of lysine 27 of H3 (H3K27me3), associated with repressive chromatin (Barski et al., 2007, Boyer et al., 2006 and Lauberth et al., 2013). The function of nuclear non-coding RNAs is less well understood (Chu et al., 2015), but they also appear to be involved in nuclear organisation (Luo et al., 2015) and transcriptional regulation (Dorn and Matkovich, 2015). Xist, which plays a pivotal role in the inactivation of the X chromosome, is an example that has been the subject of intensive research (Gendrel and Heard, 2011).

1.3.3 Epigenetic changes with age

Research into the role of epigenetics in ageing has interrogated a diverse number of cell types and tissues in a number of important model organisms. These studies have resulted in the determination of some common epigenetic changes with age. For example, it is often seen that histones become sparser and their positioning across the genome becomes less well defined with age (Das and Tyler, 2012). In addition, studies have shown increased levels of spurious transcription, alterations in splicing efficiencies and mRNA processing (Busuttil et al., 2007, Heintz et al., 2017 and Rangaraju et al., 2015). Common changes in DNA methylation with age have also been determined, and these will be described in section 1.3.4. One important caveat of these epigenetic studies is that all measurements were made at a cell population level. This means that any underlying heterogeneity within the cell population could have been masked or averaged out. For ageing studies, this is a significant limitation because it may obscure crucial characteristics of the decline in organismal fitness.

1.3.4 DNA methylation

DNA methylation is found in all kingdoms of life and involves the addition of a methyl group onto a nucleobase. The most commonly methylated base is cytosine, although there is some interspecies variation (Willbanks et al., 2016). For instance, adenine is commonly methylated in many bacterial species with asymmetric DNA methyltransferases capable of methylating both N6' of adenine and C5' of cytosine (Ryazanova et al., 2012). The high structural conservation among DNA methyltransferases suggests that the emergence of DNA methylation was an evolutionarily distant event (Iyer et al., 2011). Some species have since lost the ability to methylate their genome, including some model organisms such as C. elegans (Soojin, 2012). Although the catalytic subdomain, within the methyltransferase domain of DNA methyltransferases, is strongly conserved at the structural level, the way in which the targeting of this activity has evolved varies hugely. For instance plants can target DNA methylation using siRNA-guided mechanisms not seen in other organisms (Zhang and Zhu, 2011). Additional to the targeting, the actual downstream role or function that the methylation itself has (Ryazanova et al., 2012) varies hugely between species and even within different cell types of the same organism.

In mammals, DNA methylation occurs primarily on the C5' position of cytosines that reside in the symmetric, heritable context of CpG dinucleotides. The enzymes responsible for this modification in mice and humans are known as DNA methyltransferases (DNMTs). Of the enzymes encoded in the mammalian genome, four are known to have catalytic activity on DNA and one on tRNAs (Barau et al., 2016, Bestor, 2000 and Dong et al., 2001). Three of these DNMTs are *de novo* DNMTs: DNMT3A, DNMT3B and the most recently annotated, DNMT3C. DNMT3A and DNMT3B have distinct but partially overlapping functions in different cell types (Challen et al., 2014 and Li et al., 2015). In addition, both enzymes have been shown to complex with the catalytically inactive DNMT3L, which increases their activity *in vitro* (Suetake et al., 2004). The principal DNA sequence methylated by both enzymes is 5'-CpG-3', with additional activity at 5'-CpA-3' albeit 10-100-fold lower (Aoki et al., 2001). In addition to this sequence specificity, these enzymes have preferences regarding the nucleobases upstream and downstream of this target site. The preferred sequence 5'-RCGY-3' (where R=purine and Y=pyrimidine) can confer more than 10 times the enzymatic activity of the least preferred sequence, 5'-YCGR-3' (Handa and Jeltsch, 2005). DNMT3A and DNMT3B methylate DNA differently: DNMT3A methylates in a distributive fashion whilst DNMT3B methylates in a progressive fashion (Gowher and Jeltsch, 2001 and Norvil et al., 2016). This is seemingly due to a more positively-charged DNA binding region in DNMT3B ensuring increased processivity (Norvil et al., 2016). DNMT3A can form tetrameric complexes with DNMT3L (DNTM3A-DNMT3L-DNMT3L-DNMT3A) that bind target DNA such that the catalytic sites of the two DNMT3A subunits reside 8-10 bp apart, allowing them to catalyse DNA methylation simultaneously (Jia et al., 2007). Indeed, this 8-10 bp methylation spacing is seen in many maternally-imprinted genes in mice and some highly-methylated regions in humans (Jia et al., 2007). DNMT3C is distinct from the other two *de novo* DNMTs in that it is solely expressed in the male germ line, where it is responsible for methylating the promoters of evolutionarily young retrotransposons (Barau et al., 2016).

DNMT1, the maintenance DNMT, is primarily responsible for the symmetric re - establishment of DNA methylation, predominantly at CpG sites that are hemi-methylated (i.e. only one of the two cytosine bases within this double-stranded DNA dinucleotide is methylated). During DNA replication, maintenance of DNA methylation is ensured by interaction of DNMT1 with the Ubiquitin-like with PHD and RING finger domains 1 (UHRF1) protein, which can bind hemi-methylated CpG dinucleotides and 5'- hydroxymethylcytosine in DNA through a SET and Ring finger Associated (SRA) domain (Bostick et al., 2007 and Sharif et al., 2007). UHRF1 also ubiquitylates histone H3 lysine K23 through its Really Interesting New Gene (RING) domain, which in turn recruits DNMT1 to sites of hemi-methylation (Nishiyama et al., 2013). In addition, this activity is facilitated by the interaction of the N-terminus of DNMT1 with Proliferating Cell Nuclear Antigen (PCNA) via a putative PCNA Recognition Domain (PRD) (Bestor et al., 1988, Bostick et al., 2007 and Sharif et al., 2007). The specificity for these hemi-methylated sites is achieved by the presence of Bromo-Adjacent Homology Domains (BAH Domains) within the Target Recognition Domain (TRD). These BAH domains contain tryptophan and other hydrophobic residues that create a shallow cleft for recognition of the hemimethylated CpG islands.

1.3.5 The cytosine DNA methyltransferase domain

The cytosine DNA methyltransferase domain responsible for the addition of the methyl group on to the 5' position of cytosine nucleobases in the context of a DNA polymer consists of a large and small subdomain, separated by a DNA binding cleft (Figure 1.1) (Klimasauskas et al., 1994 and Song et al., 2011). The tertiary structure of the large (catalytic) subdomain consists of 7 β-sheets and 3 α-helices. Six of the 7 β-sheets are arranged in a parallel orientation. The large subdomain can be further broken down into two separate regions. The region containing β-sheets 1-3 is responsible for S-Adenosyl Methionine (SAM or AdoMet) binding and the region containing β-sheets 4-7 is responsible for the binding of the target cytosine. The small subdomain is known as the TRD.

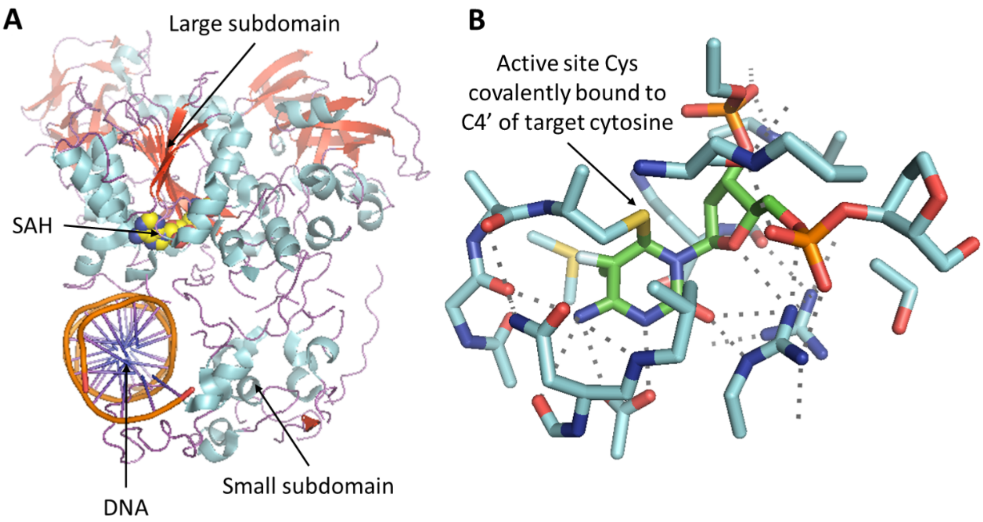


Figure 1.1: Structure of the DNA Methyltransferase Domain: (A) Visualisation of the overall structure of the C5 DNA methyltransferase domain with subdomains and cofactor highlighted. Structure shown is Dnmt1, PDB code 3PT6 (Song et al., 2011). (B) shows the active site of the C5' DNA methyltransferase domain with the target cytosine covalently bound. Structure shown is of HhaI DNA methyltransferase, PDB code 1MHT (Klimasauskas et al., 1994)

This subdomain is responsible for target specificity and as such varies far more between C5' DNA methyltransferases than the large (catalytic) subdomain. Interestingly, insights into the processivity of different methyltransferases suggest that there is some cross-talk between these two subdomains. For instance, DNMT1 appears to methylate hemimethylated DNA distributively and unmethylated DNA processively with tens of methylation events per processed stretch of DNA (Vilkaitis et al., 2005).

1.3.6 Catalysis of methylation by DNA methyltransferases

Cytosine methylation occurs via an SN_2 mechanism involving a covalently linked intermediary step (Figure 1.1). The overall mechanism can be broken down into three steps (Du et al., 2016). Firstly, the cytosine is flipped out of the DNA double helix. Next, the enzyme-substrate intermediate is formed. This intermediate is formed by covalent attachment of the side-chain sulphur atom of a particular cysteine (Cys) residue to the C4' of the target cytosine. However, there may be exceptions to the use of this cysteine, as highlighted by mutational studies of DNMT3A (Gowher and Jeltsch, 2002 and Reither et al., 2003). Lastly, the methyl group present in the SAM cofactor is attached through an SN_2-type reaction to the C5' of the target cytosine. This 5'-methylcytosine (5mC) product can then reform following basic-catalysed breakage of the covalent 5mC-Cys bond. This release of 5mC has been shown to be the rate-limiting step for *MHhaI* (a bacterial methyltransferase; Du et al., 2016 and Yang et al., 2013).

The above is the primary reaction of the enzyme, but side reactions can occur and there are a number of publications that have validated the occurrence of such side-reactions *in vivo*. One side reaction involves the deamination of cytosine to uracil (Figure 1.1). This occurs when SAM and the cofactor product S-Adenosylhomocysteine (SAH) are not bound, which allows water molecules to enter the active site and deaminate the covalently bound cytosine. Interestingly, this side reaction introduces the possibility for mutation should this formation of uracil go unchecked (Shen et al., 1995) similar to the effects described for Activation-Induced cytidine Deaminase (AID) and apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G (APOBEC3G). Another side reaction occurs in the presence of SAM and absence of DNA, resulting in irreversible automethylation of the cysteine responsible for the covalent attachment of the cytosine (Siddique et al., 2011). This automethylation is thought to be an important regulatory mechanism. Indeed, DNMT3A in complex with DNMT3L is far more likely to undergo this automethylation than when DNMT3A is on its own in solution (Siddique et al., 2011).

1.3.7 Demethylation and oxidative products

In contrast to this singular mechanism by which the DNA methyltransferases add methylation to the cytosine 5' position correctly, there is a huge variety of different mechanisms by which this modification can be removed. Some of these processes have been validated experimentally and others still await experimental validation (Hill et al., 2014).

These demethylation mechanisms can be broken down into passive or active removal of the 5mC (Seisenberger et al., 2013). Passive removal is the result of DNA repli-

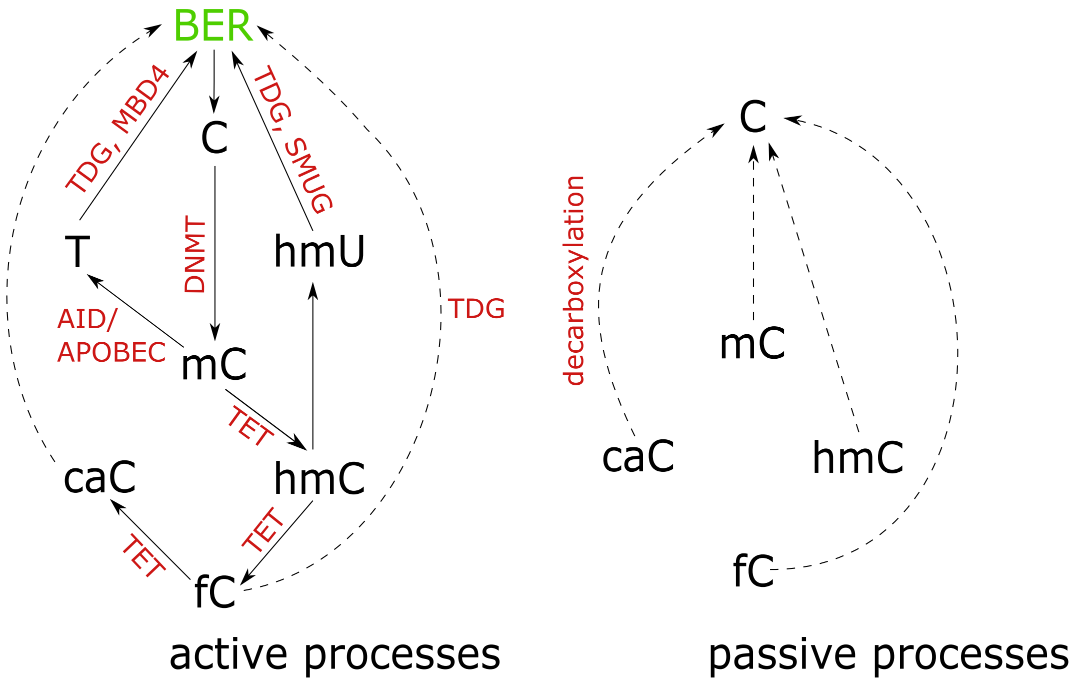


Figure 1.2: Mechanisms of DNA demethylation: DNA demethylation can occur by active or passive mechanisms, shown separately (“active process” and “passive process” respectively). Dotted arrows are used to indicate where there are intermediary steps that aren’t shown.

cation in the absence of sufficient epigenetic maintenance (von Meyenn et al., 2016). Active removal can involve the oxidation of the methyl group with Ten-Eleven Translocase (TET) enzymes, of which there are three (TET1, TET2 and TET3; Iyer et al., 2009, Tahiliani et al., 2009 and Wu and Zhang, 2017; Figure 1.2). These enzymes are methylcytosine dioxygenases that catalyse the oxidation of methylcytosine to hydroxymethylcytosine. In addition, they may also catalyse the formation of formylcytosine and carboxycytosine. However, structural studies have suggested that these enzymes are fine-tuned to ensure stability of the hydroxymethyl mark (Hu et al., 2015), suggesting that this mark maybe a functional epigenetic mark in its own right (Iurlaro et al., 2013). The enzymes themselves require both Fe(II)- and 2-oxoglutarate (2OG) for their oxidation activity, and are competitively inhibited by 2-hydroxyglutarate (2HG) owing to its similarity to 2OG (Yang et al., 2014). Alternatively, active removal could occur through deamination via proteins such as AID or members of the APOBEC superfamily of proteins such as APOBEC3G (Nabel et al., 2012). These oxidation and/or deamination events can then be recognised as mismatches by the Base Excision

Repair (BER) or Nucleotide Excision Repair (NER) pathway proteins and repaired, resulting in the removal of the modified base and replacement where successful by a cytosine (Zhu, 2009). Such proteins include Thymine DNA Glycosylase (TDG), Uracil DNA Glycosylase (UDG) and Methyl Binding Domain 4 (MBD4); all of which are DNA glycosylases. The exact mechanism employed for a given methylation event will invariably involve a combination of these processes and multiple compensatory systems may act within the same pathway. The mechanisms of demethylation described here are indirect in nature owing to the difficulty in cleaving this stable aromatic-proximal C-C bond. However, it has been shown *in vitro* that the DNMT enzymes themselves are capable of catalysing this reverse reaction under correct substrate conditions (Chen et al., 2013b).

1.3.8 DNA methylation in the study of ageing

The nature of the relationship between DNA methylation and ageing is complex. Scientists have sought to explain this relationship by condensing observations down into singular statements of global hypomethylation and focal hyper-methylation with age (Jung and Pfeifer, 2015). However, the literature is contradictory in this regard. For instance, numerous studies claim that 5mC levels drop with age (Heyn et al., 2012), while many other studies report a gain in 5mC with age, including but not limited to liver samples and haematopoietic stem cell (HSC) studies (Hahn et al., 2017 and Sun et al., 2014). One possible reason for this variation is that the individual samples considered are very rarely taken in a longitudinal fashion from the same individuals, and as such they could be driven by variation between individuals irrespective of age. The potential for this is highlighted by the magnitude of the global differences reported typically being minute (<5%), and as such within the range of population variation (Bjornsson et al., 2008, Oey et al., 2015 and Xia et al., 2015). In fact, in one of the few longitudinal studies that could address the question of whether there was a global hypo- or hyper- methylation with age, the result was that in any given individual either could be the case (Bjornsson et al., 2008). This highlights the importance of longitudinal studies in ageing research, but also the care in evaluating the importance of drawing conclusions from large datasets containing small sample sizes.

Importantly, although these global simplifications of the link between DNA methylation and ageing are often misleading, there are certain aspects of this relationship that do appear to be highly conserved across cell types and across species (Jung and Pfeifer, 2015).

One such aspect is that when genomic features are assessed for their enrichment in ageing-associated DNA hypermethylation, these regions are highly enriched for Polycomb Group (PcG) target genes (Johnson et al., 2014, Maegawa et al., 2010 and Teschendorff et al., 2010). Interestingly, this is also seen in studies that have sought to understand DNA hyper-methylation events in ageing-related diseases such as cancer (Hahn et al., 2008, Kalari et al., 2013, Ohm et al., 2007, Rauch et al., 2006, Schlesinger et al., 2007 and Widschwendter et al., 2007). Polycomb Group proteins act to stably repress gene expression networks specific for cell states present in early development. This is achieved through the regulation of chromatin (Bernstein et al., 2006, Blackledge et al., 2014, Boyer et al., 2006, Bracken et al., 2006 and Lee et al., 2006). In mammals, these proteins comprise two complexes: Polycomb repressive complex 1 (PRC1) and polycomb repressive complex 2 (PRC2) (Di Croce and Helin, 2013 and Shao et al., 1999). These complexes were first identified in *D. melanogaster*, but are conserved in all animals, including humans (Lewis, 1978). The components of PRC1 are CBX (Pc homolog), PHC1, 2, and 3 (PH homologs), Ring1a and Ring1b (dRING homologs), BMI1 and six minor others (PSC homologs) (Levine et al., 2002) and PRC2 is composed of enhancer of zeste homolog 1 and 2 (EZH1/2), suppressor of zeste 12 (SUZ12), embryonic ectoderm development (EED) and the histone binding protein RbBP4 (Di Croce and Helin, 2013). Although these are the main components of PRC1 and PRC2 there are additional variants and homologs that could also function as part of these complexes (Brien et al., 2012, Di Croce and Helin, 2013, Hunkapiller et al., 2012 and Pasini et al., 2010). PRC2 is responsible for the tri-methylation of H3K27, via the catalytic action of EZH2, to H3K27me3 (Francis et al., 2004). This mark its self is stably maintained over cell division and is repressive in nature (Grossniklaus and Paro, 2014). However, it can also recruit PRC1, which recognises H3K27me3 through CBX, resulting in monoubiquitination of H2A on K119 via RING1/2, further compacting the chromatin and resulting in stable repression (Eskeland et al., 2010 and Francis et al., 2004).

The mechanistic link between DNA hypermethylation and loss of PRC2 marks is still not fully understood, but clues from the literature suggest that the most likely mechanism is one of competition between the *de novo* DNMTs (DNMT3A and DNMT3B) and PRC2 (Cedar and Bergman, 2012, Gal-Yam et al., 2008 and Jung and Pfeifer, 2015). This erosion of PRC1 and PRC2 protection with age would then result in *de novo* methylation and less plastic repression of gene expression, since DNA methylation is assumed to be a less plastic mark. This mechanism has been hypothesised to be occurring in cancer as well, where it is referred to as Polycomb switching (Gal-Yam et al., 2008). One piece of evidence for this model comes from the suggestion that absence of DNA methylation at CG dense regions of the genome is sufficient for recruitment of PcG (Lynch et al., 2012). This was shown in mouse embryonic stem cells (ESCs) that were double knock-out for both DNMT3A and DNMT3B. Another study showed that regions of the genome that are predominantly hyper-methylated become enriched for H3K27me3 upon inhibition of DNA methylation with 5-aza-2'-deoxycytidine (a competitive inhibitor for DNA methyltransferases; Reddington et al., 2013). One mechanism by which PRC2 is thought to be recruited to unmethylated DNA is through Lysine Demethylase 2B (KDM2B), which demethylates lysine K4 and K36 in histone H3. Although this strong association for hypermethylation has been found, so far it has not been possible to strongly link any one particular genomic feature or mechanism to the DNA hypomethylation that occurs. However, there are hypotheses that include: loss of activity of DNMT1 with age, DNMT3A and DNMT3B (Xiao et al., 2008), reduced expression of DNMT proteins with age, reduction in dietary intake of crucial one-carbon metabolites, reduced basal metabolic rates resulting in reduced levels of methionine and thus SAM alongside many others (Chiang et al., 1996, James et al., 2002, Mason, 2003, Schrack et al., 2014 and Ulrey et al., 2005).

Another conserved aspect of DNA methylation changes that occur with age, which is suggestive of potentially the nature of ageing its self, is that the changes that are seen appear to result in an entropic increase in the methylation at a given site (Hannum et al., 2013 and Slieker et al., 2016). Entropy here defined as the Shannon Information Entropy of the position. Put another way, sites that are lowly methylated will tend to increase in methylation with age and sites that are highly methylated will tend to become hypomethylated with age (i.e. tending towards an entropic maximum of 50%; Hannum et al., 2013 and Slieker et al., 2016). This is of interest because it suggests that perhaps the nature of epigenetic changes that are seen with age follow what would be expected for ageing being a passive decay process.

However, it has recently become apparent that DNA methylation changes that occur with age are predictive of chronological age in a highly accurately manner suggesting that either this passive decay process at a population level is ordered enough to result in a measurable direction (similar to radioactive decay processes) or that this process is perhaps the result of an actively procured state (Bocklandt et al., 2011, Florath et al., 2013, Hannum et al., 2013, Horvath, 2013 and Weidner et al., 2014).

1.3.9 Estimating age from DNA methylation

Many studies conducted in humans have shown that there are age-related changes in DNA methylation or age-related differentially methylated regions (DMRs) and that changes in methylation appear to occur consistently within large cohorts, suggesting that they represent a consistent phenomenon. More recently, largely through utilisation of the powerful DNA methylation array-based technology developed by illumina (27k, 450k and EPIC arrays), scientists have been able to utilise methylation changes at single cytosine positions in the genome to define predictors of age (Bocklandt et al., 2011, Florath et al., 2013, Hannum et al., 2013, Horvath, 2013 and Weidner et al., 2014).

These ‘epigenetic clocks’ have been defined using linear regression-based models. In the case of Weidner and Florath, these models were derived from multivariate linear regression models from sites that were previously selected as being age-correlated either from Pearson correlation or from Spearman correlation respectively (Florath et al., 2013 and Weidner et al., 2014). The data for the Weidner model came from four studies that utilised the 27k MethylArray and all utilised whole blood samples. In the case of Florath, the data utilised in the study came from the ESTHER cohort (Bock et al., 2016) and all samples were again whole blood, though this time run on the more recent 450k MethylArray. The Florath predictor was defined from 17 cytosine sites in the genome and the Weidner predictor utilised only three. In both the median absolute deviation (MAD) was <4yrs (Florath et al., 2013 and Weidner et al., 2014).

In the case of the Hannum predictor and the Horvath predictor, multivariate linear regression was again performed, this time using implementations of an elastic-net linear regression (Hannum et al., 2013 and Horvath, 2013). An elastic-net regression is a combination of both a Lasso-style model in which the model attempts to greatly reduce the number of parameters in order to define the predictor and a ridge-style model in which the predictor utilises all positions in the model. This model was chosen in both instances, because it has proven useful for linear regression-based approaches where the number of parameters (or CG sites in this case) greatly outweigh the number of samples that are available. An elastic-net based model, therefore, utilises both the L1 absolute value penalty (implemented in Lasso-based regression) and L2 quadratic value penalty (implemented in Ridge-based regression). Both studies utilised the "glmnet" implementation of the elastic-net model (Friedman et al., 2010). The extent to which the elastic-net regression model utilises these two penalties is defined by a term *alpha* (α), which is a ratio term and varies between 1 for a fully Lasso-based model and 0 for a fully Ridge-based model. In glmnet the *lambda* terms that define the severity of the L1 and L2 penalties are optimised via gradient descent. The *alpha* term that defines the ratio between the two penalties can either be chosen arbitrarily or optimised during training.

The Hannum predictor was defined using 450k MethylArray data of 656 whole blood samples (Hannum et al., 2013). The predictor that ended up being defined, contained 71 cyotsines and had a test MAD error of 4.9 years (Hannum et al., 2013). In addition, to assess whether certain individuals within their study age faster than others and what could be the cause of this, they defined an individual-specific term: apparent methylomic aging rate (AMAR). This term could then be used to test a number of hypotheses about what may affect the apparent ageing rate, including ethnicity, BMI and gender. For instance, they found that males were significantly more likely to age faster than women, but that the distribution between the two was not different. In addition, it was found in a number of studies that there are longevity-associated SNPs in the human genome, and so, to understand whether there were any SNPs within their human samples that could be affecting the AMAR of a given individual, they performed exome sequencing on 252 of the individuals (Hannum et al., 2013). Although a relatively small sample size for GWAS, this information would enable them to decipher not only whether there were SNPs that were effecting methylation at the designated cytosines, but also whether the cyotsines themselves were being mutated at a relatively low level. As such, it would be possible to assess whether the observed methylation changes are actually a result of changing methylation and not of an underlying level of DNA mutation, which are known to accumulate with age (Girard et al., 2016). This analysis highlighted 303 methylation Quantitative Trait Loci (meQTLs; Hannum et al., 2013). One that was particularly interesting, was a SNP (rs140692) in an intron contained within MBD4 and the age-associated cytosine found just upstream of the coding sequence of this gene. Importantly, the SNP fell outside of the range of the probe sequence its self and so couldn't be causing this effect from a purely technical stand point. Importantly, this study also tested their 71 cytosine model in tissues other than blood using The Cancer Genome Atlas (TCGA) dataset. The importance of this is due to it being known that the blood composition changes with age in a very defined and predictable manner, and it is also known that the different blood cell types within blood have different methylation signatures and so you could imagine how a predictor could be built that only took these features into account, although previous studies on purified cell types have shown that there are age-related changes in blood independently of cell heterogeneity changes (Rakyan et al., 2010). They found that their model was able to predict age in these different tissues but found that there was a clear linear offset in both the gradient and the slope (Hannum et al., 2013). Showing that the epigenetic changes that they were using to predict in the blood are perhaps more widespread. In addition, they show that when cancer samples are interrogated they appeared to have an accelerated epigenetic age independent of the source of the tissue studied. Suggesting that cancer incidence accelerates the epigenetic ageing process. Additional validation was conducted in whole genome bisulfite sequencing (WGBS) samples validating that the predictor could function cross-platform. This predictor is exciting because it suggested for the first time that not only could a chronological predictor of age be defined from changes in DNA methylation with age, but that also inter-individual differences in this prediction are perhaps biologically meaningful.

The Horvath predictor was different from the previous predictors mentioned in that it was derived *ab initio* from multiple different tissues (Horvath, 2013). Horvath utilised data from a large number of different datasets (82), combining more than 7,000 samples from both the 27k MethylArray and the 450k MethylArray, as such he utilised fewer sites in the training of the model than Hannum had used in his model (21,369 cytosines). In addition, Horvath performed a split transformation of age, since he found that this provided a better fit to the training data in addition to having a number of additional properties. For instance, adding one to the function enabled negative ages (or pre-birth) samples to be considered:

If age <= adult.age (set to 20 years of age in humans):

$$F(age) = \log(age + 1) - \log(adult.age + 1) \tag{1.1}$$

Else if age > adult.age:

$$F(age) = \frac{(age - adult.age)}{adult.age + 1} \tag{1.2}$$

The Horvath predictor contained 353 cytosine sites, although he also showed that a similar predictive accuracy could be achieved with a reduced 110 of these sites. The test MAD error of the 353-site predictor was 3.6 years, slightly less than that achieved by Hannum, but more than the Weidner predictor (although this was only a blood-based predictor; Horvath, 2013).

Horvath, was able to show that he could not only predict chronological age in human tissues, but that he could also predict age in liquid samples such as saliva, in cell culture samples, such as fibroblasts (Horvath, 2013). In addition, he was able to show that iPSCs have an epigenetic age of less than zero in the majority of cases (Horvath, 2013). He was also able to show that, although nowhere near as accurate and only from a very small sample size, he could predict chronological age in other great apes (Chimpanzees, Bonobos and Gorillas; Horvath, 2013). This comparison is hindered because the 450k MethylArray is optimised for the human genome and owing to it not being a sequencing technology, it relies upon strong homology at the bases surrounding the cytosine that is under study, which will often not be the case (Hernando-Herraez et al., 2013). In addition, the age transformation in the case of the great ape comparisons was slightly altered with adult.age equating to 15 instead of 20. The Horvath predictor also highlighted that there were a number of tissues that were less well calibrated. These tissues fell into two categories: those that were hormonally regulated such as breast tissue and uterine endothelium and muscle tissues such as skeletal muscle and cardiac samples (Horvath, 2013). However, in contrast to Hannum's predictor, the Horvath study found no clear directionality to the change that cancer would have to the epigenetic age, i.e. depending on the cancer or cancer cell line epigenetic age is either accelerated or decelerated (Hannum et al., 2013 and Horvath, 2013). In addition, contrary to what could be expected from the known lifespan shortening effect of HGPS, in immortalised B cells derived from these patients there was no significant alteration to the epigenetic age in these individuals (Horvath, 2013). This could reflect small sample sizes, since the majority of epigenetic age differences that are seen are small and require large sample sizes in order to have the statistical power to call them as significant, or it may be that this disease does not reflect the process of natural ageing. Another interesting observation is that cells in culture age across passages, for instance mesenchymal stromal cells have both been shown to have an increased epigenetic age with increased passage (Horvath, 2013). More interesting still is the finding that whether cells are proliferating in culture or whether they are senescent they still age at the same rate, suggesting that this predictor is not measuring a mitotic rate, but that it is potentially measuring "the cumulative effect of an epigenetic maintenance system" (Horvath, 2013) that is at least partially independent of cell replication. Additional circumstantial evidence for this is provided by samples from brain tissues that are commonly thought of as being predominantly post-mitotic and samples from more actively turned over tissues such as blood having the same epigenetic ageing rate. One interesting corollary of this is that perhaps this epigenetic maintenance system in different tissues is maintained or managed through different processes, for instance in non-replicative tissues perhaps hypomethylating sites are more closely linked to processes of active demethylation involving hydroxymethylation. This is backed up by the high levels of 5hmC seen in postmitotic tissues relative to replicative tissues (Meng et al., 2014).

Both predictors, are intriguing from the point of view of suggesting that the phenomenon of epigenetic drift or epigenetic ageing appears to be reflected in multiple tissues (Hannum et al., 2013 and Horvath, 2013). Though it is important to note that this does not mean that there are not large age-associated changes happening that are not tissue specific. In addition, both studies find certain cytosine positions within the gene Elongation Of Very Long chain fatty acids protein 2 (ELOVL2) to be strongly correlated with age. ELOVL2 is responsible for the catalysis of the rate limiting first step of the fatty acid elongation cycle (Leonard et al., 2004) and has previously been associated with ageing (Garagnani et al., 2012). In fact, in all the predictors defined by Hannum et al. cytosine sites in ELOVL2 are present (Hannum et al., 2013). Aside from sites over this gene and a number of other genes it is interesting to note that there is very little overlap between the sites chosen by these two models suggestive that the epigenetic changes that are occurring may perhaps be happening at a more global level.

Excitingly, both studies have shown that they are not only able to estimate chronological age but also “biological age” (Hannum et al., 2013 and Horvath, 2013). At present “biological age” is loosely defined as a comparative measure of a disease state, attribute or trait that is altered with increasing age. For instance, someone with obvious signs of muscle weakness above what is expected for their age-group would be considered to have an accelerated or increased “biological age”. In the case of the Horvath predictor this has been more widely addressed in subsequent studies. Such studies have addressed the impact of Werner Syndrome, HIV-1, Down’s Syndrome, Menopause, fatty liver disease, diet, Parkinson’s and many more (Carroll et al., 2017, Horvath et al., 2014, Horvath et al., 2015a, Horvath et al., 2015b, Horvath and Levine, 2015, Horvath and Ritz, 2015, Horvath et al., 2016a, Horvath et al., 2016b, Levine et al., 2015a, Levine et al., 2015b, Lu et al., 2016, Marioni et al., 2015b and Vidal-Bralo et al., 2016). Interestingly it has also been shown that this phenomenon of tissues coherently ageing falls apart in super-centenarians for whom regions of the brain age at different rates with the cerebellum for instance ageing more slowly than the cortex (Lu et al., 2016).

In addition to these observational studies, inadvertent manipulation based studies have also been conducted that address questions as to the plasticity of the clock and how the extrinsic environment effects it. A number of studies have now attempted to address the impact of donor vs recipient age on the incumbent age of transplanted haematopoietic stem cells (HSCs). In essence, they studied the introduction of HSCs into recipients of older or younger age than the donors. These studies have utilised both the Weidner and the Horvath clock (Stölzel et al., 2017 and Weidner et al., 2015). Together they show that upon introduction into the recipient there is a brief reduction of age up until 6 months of age (relative to the donor) after which the age of the blood begins to rise. At about 1 year the age of the blood is back in line with the age of the donor. However, by three years of age the epigenetic age is accelerated relative to the donor and this continues to rise until the end of measurement (Stölzel et al., 2017). This result suggests that the cells are intrinsically ageing and that their new environment can alter the rate of epigenetic age increase but that it is unable to result in large changes to the epigenetic age *per se*.

Interestingly, alongside studying the implication that different parameters have on biological age, researchers have started applying these predictors and others built off the same principles, to the question of mortality and lifespan prediction (Chen et al., 2016, Marioni et al., 2015a and Thinggaard et al., 2016). In other words, can we predict not simply how old someone is but how long they have to live and it appears that we can. In a recent paper, Chen B. *et al*, showed that it was possible to predict time to death using all measures of age acceleration (Chen et al., 2016). In another such paper Marioni R. *et al*, show that it is possible to predict all-cause mortality from these derivative metrics (Marioni et al., 2015a).

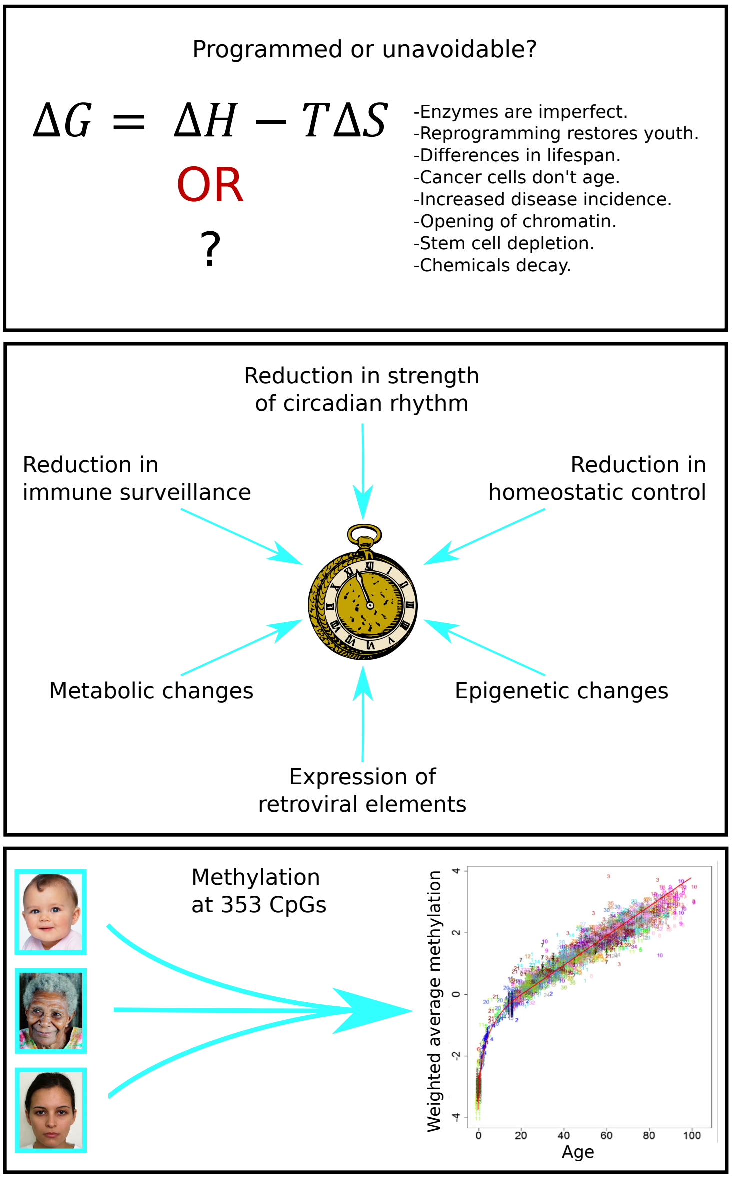

Programmed or unavoidable?
ΔG = ΔH − TΔS
OR
?
-Enzymes are imperfect.
-Reprogramming restores youth.
-Differences in lifespan.
-Cancer cells don't age.
-Increased disease incidence.
-Opening of chromatin.
-Stem cell depletion.
-Chemicals decay.
Reduction in strength of circadian rhythm
Reduction in immune surveillance
Reduction in homeostatic control
Metabolic changes
Epigenetic changes
Expression of retroviral elements
Methylation at 353 CpGs
Weighted average methylation
-4
-2
0
2
4
0
20
40
60
80
100
Age

Figure 1.3: Mechanisms of ageing: (A) The nature of ageing: is it programmed or a stochastic and inevitable decline? (B) Schematic of changes associated with ageing that could be responsible for our ability to predict age from epigenetics. (C) A depiction of the human DNAmAge clock, highlighting that individuals of different ages can have their methylomes interrogated at 353 CpG sites and an accurate prediction of their age made.

Although it seems unlikely that DNA methylation is the mechanism underlying ageing in of itself, since this epigenetic modification is not found in all organisms (Lee et al., 2010). It does provide a useful readout of underlying chromatin and metabolic changes that are potentially occurring with age. As such, a greater understanding of the mechanism behind these epigenetic clocks would be immensely useful as a research tool. In addition, they will be of key relevance to understand ageing in more detail and will also be instrumental for the design of future interventions. However, at present all of these predictors have been developed in humans and so there is a growing need to firstly test the conservation of such an epigenetic system evolutionarily but also to define an epigenetic age predictor in an experimentally tractable system.

As such, in this thesis I have set out to define a multi-tissue epigenetic predictor of age in the mouse. At present the only evidence that such an epigenetic predictor may also function in the mouse is based on one recent study using the Sequenom EpiTYPER to assess the directionality of methylation changes in mouse at ageing-associated cytosine sites identified in human, including at ELOVL2 (Spiers et al., 2016).

1.3.10 The study of heterogeneity

The study of heterogeneity is that of variation within a sample. This is a concept that lies at the very heart of experimental science, where heterogeneity (or differences) are often phenotypes that can lead investigations. At a single cell level, the study of heterogeneity has been possible for many years owing to microscopy providing us with the ability to interrogate individual cells visually.

In more recent times, the ability to study single cells in unprecedented detail is now becoming possible with methodologies that utilise the power of next generation sequencing (NGS) based approaches. These approaches are hugely valuable for being able to assess and study difference between single cells as they carry within them a vast amount of data that can be analysed. At present, it is possible to interrogate genomic (scDNA-Seq; Xu et al., 2012), transcriptomic (scRNA-Seq; Islam et al., 2014 and Picelli et al., 2013), nuclear conformation (scHi-C; Nagano et al., 2013), DNA accessibility (scATAC-Seq and scNOMe-Seq; Buenrostro et al., 2015 and Pott, 2017), DNA methylation (scBS-Seq; Clark et al., 2017 and Smallwood et al., 2014), hydroxymethylation and formylcytosine (scAba-Seq and CLEVER-Seq; Mooijman et al., 2016 and Zhu et al., 2017) heterogeneity using single cell studies; with more techniques likely to expand this list of possibilities in the near future. Many of these approaches have now been expanded into high-throughput methods that can be used to study vast numbers of single cells in any one experiment.

In addition, it is possible to now study both genomic and transcriptomic (scG&T-Seq; Macaulay et al., 2015), DNA methylation and transcriptomic (scM&T-Seq; Angermueller et al., 2016) and DNA methylation, transcriptomic and DNA accessibility (scNMT-Seq; (Clark, S, *et al.*, in review)) heterogeneity at single cell resolution in combination. This has allowed, for the first time, a direct link between these different layers of information to be studied. For instance, the association between promoter methylation and transcriptional output can be directly interrogated without compromise, as in bulk studies (Angermueller et al., 2016).

In addition to being able to assess these multiple layers of regulation and expression both in single and combined settings, it is now becoming possible to assess these differences between cells in relation to their relationships to the other cells present. This is being done using a number of different approaches that more generally are referred to as lineage tracing methods, some of which include the incorporation of barcodes into the cells of interest, while others utilise the inherent behaviour of already present host epigenetic marks, such as hydroxymethylation (Davis et al., 2016, Fischer et al., 2016, Kalhor et al., 2016, Li et al., 2016, McKenna et al., 2016, Mooijman et al., 2016 and Woodworth et al., 2017). These methods will be incredibly useful in the future for the study of heterogeneity, cancer and ageing, since they will allow the scientist to understand the life history of a given cell and how it came to be in its current state.

However, there are currently limited methods available for the study of a given cell across multiple time points. This would be incredibly useful for being able to look at dynamics that are currently only being modelled/inferred from NGS-based single cell approaches. In particular, there are no methods currently available for the quantitation of DNA methylation dynamics over time from the same cell. From the literature, there have been a number of papers that have utilised a methyl binding domain (MBD) fused to a fluorescent protein (Ingouff et al., 2017, Kimura et al., 2010, Ueda et al., 2014, Yamazaki et al., 2007 and Yamagata, 2010). In most cases this protein was the MBD of methyl binding domain protein 1 (MBD1) fused to enhanced green fluorescent protein (eGFP). This construct is hereto referred to as MBD1-eGFP and has been used to assess the global dynamics of DNA methylation at a qualitative level in previous studies (Yamagata, 2010). For instance, it has been used to describe the changes in the nuclear organisation of methylated regions of the genome during early development (Yamazaki et al., 2007), however it has never been used for a quantitative assessment of DNA methylation. In addition, other studies have assessed DNA methylation qualitatively in live cells using constructs capable of differentiating between methylation in contexts not commonly seen in mammalian settings such as CHH context (H = C, T or A; Ingouff et al., 2017). With the quantitative experiments in this study all conducted on fixed cells, negating the ability to study the cells across multiple time points.

As such, in this thesis I have detailed experiments set out to define a novel system within which it is possible to study global methylation dynamics over a period of time from the same cell. The method under development is called Differential Dynamic Microscopy (DDM; Cerbino and Trappe, 2008) and has been previously used for the study of colloidal particles in Brownian motion (Lu et al., 2012) and more recently for the study of ensemble bacterial motion (Lu et al., 2012 and Wilson et al., 2011).

1.3.11 Heterogeneity in ageing

The study of heterogeneity at the single cell level is of great interest in the study of ageing because it will provide us with insight into the very nature of ageing its self. Hopefully allowing us to tease apart whether ageing is the result of a passive decay process, or an actively procured state. This will be made possible by the unprecedented detail that we can now study single cells at allowing us to test hypotheses about how

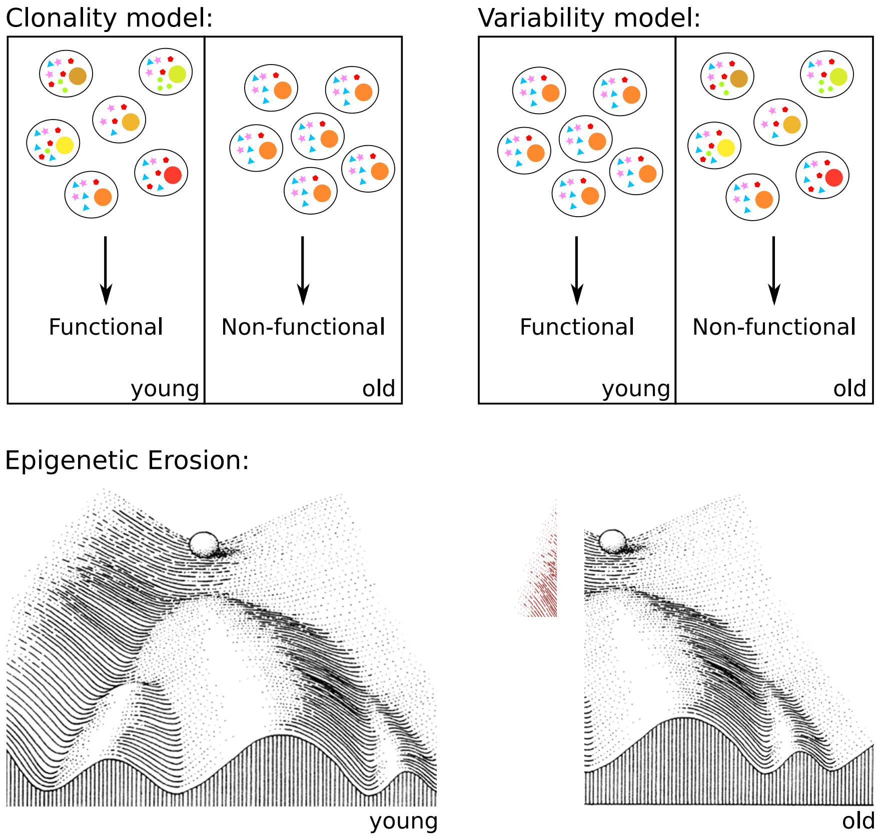


Figure 1.4: Potential models for heterogeneity with age: (A) Two proposed models for how heterogeneity may be altered in different compartments during ageing. (B) A model for how epigenetic drift with age may alter the capacity of an organism to maintain its required cell types and in their required stoichiometries. Both epigenetic erosion and changes in heterogeneity are likely playing a role in the ageing process, though erosion has been far better documented than heterogeneic changes.

these two concepts could be differentiated. In addition, to the advances that studying at this level of detail will provide in this abstract sense, it will also provide us with huge amounts of information as to what defines the aged state of different cell types and tissues. It will enable us to define and understand whether the proportions of cell types within a given tissue are changing with age or whether the very nature of a given cell type is changed with age. Information that will hopefully answer questions such as why age-related diseases increase with age.

At present, there are very few studies that have addressed the nature of ageing at the single cell level, one such study was performed in cardiomyocytes (Bahar et al., 2006). This study was based on previous work that showed that with ageing there was an increase in DNA mutations and genomic rearrangements in cardiomyocytes, which they showed in mouse embryonic fibroblasts (MEFs) could result in increased transcriptional variability (Dolle et al., 2000 and Dollé and Vijg, 2002). As such they decided to characterise whether such variability could be seen for the cardiomyocytes themselves by taking a panel of house-keeping and cardiomyocyte-specific transcripts. They found that with age there was an increase in transcriptional heterogeneity and that this increase in transcriptional heterogeneity could provide a mechanism to explain the functional decline of the tissue with age, although no functional link was made (Bahar et al., 2006 and Baris et al., 2015).

Another single cell transcriptome study, this time focused on ageing of the HSC compartment, also yielded valuable insights into the nature of ageing at the single cell level. It is already known from bulk studies that with age the HSC compartment changes in its propensity to differentiate into all of the defined lineages (Rimmelé et al., 2014 and Rossi et al., 2005). This change in propensity results in a myeloid lineage skew with age (Rossi et al., 2005). In addition, it has been shown with age that the HSC compartment becomes more clonal in nature owing to a certain number of HSCs escaping quiescence and expanding within the niche (Beerman et al., 2010). This phenomenon has been further validated by the observation that in humans there are a large number of cells carrying similar mutations in the peripheral blood suggesting that the majority of them are derived from the same mother cells (Buscarlet et al., 2017). Of particular interest the two most frequently mutated genes in this system are DNMT3A and TET2 (Buscarlet et al., 2017 and Zhang et al., 2016b). Excitingly, this single cell study has identified that there was an age-specific subpopulation of HSCs that were myeloid biased and appeared to be expressing pro- and anti-inflammatory signals at the same time (Kirschner et al., 2017). The fact that these cells appeared to be myeloid biased and expressing markers consistent with proliferation and DNA damage (p53), suggests that they could represent the sub-population from which this clonal phenotype seen in humans is derived.

Another interesting observation pertaining to stem cells is that it appears that they are able to undergo asymmetric cell division at a number of levels, from chromosomal asymmetric division to asymmetric division of organelles such as mitochondria. One curious hypothesis of this behaviour is that evolution is trying to maintain the most faithful daughter cell in the least differentiated state (Katajisto et al., 2015). This would suggest that you would expect alterations to the extent of heterogeneity within a tissue dependent upon its differentiation potential, a hypothesis that is yet to be tested in the setting of DNA methylation.

In this thesis, I describe the work conducted to address the question of whether a homogeneous population of cells will exhibit concerted changes with age that would be reminiscent of an actively procured state, or whether they would exhibit evidence of cells that would more closely resemble that of a passive decay process. This question has been addressed in an *in vivo* context, utilising the combined single cell DNA methylome and transcriptome sequencing method recently developed in the group (Angermueller et al., 2016). This data, has enabled this question to be addressed at both the epigenetic and transcriptomic level. The homogeneous cell population used as the model system in this study was a highly quiescent skeletal *tibialis anterior* (TA) muscle satellite cell subpopulation (TA-Hi MuSCs; Sambasivan et al., 2011). In addition to being a useful model system, these cells are of clinical importance in the study of ageing owing to the burden of frailty in the general population and particularly with age (Janssen et al., 2002).

Chapter 2

Material and Methods

2.1 Materials

Table 2.1: Kits

Product	**Commercial supplier**
Miniprep kit	Qiagen
Gel extraction kit	Qiagen
PCR Purification kit	Qiagen
DNeasy Blood & Tissue Kit	Qiagen
Kapa Library Quantification kit	Kapa Biosystems
High Sensitivity DNA kit	Agilent
Nextera XT kit	Illumina
Imprint® DNA modification kit	Sigma

Table 2.2: Instruments

Product	Commercial supplier
Pipettes	Gilson Pipetman P
NanoPhotometer	NanoDrop® Technologies
Thermocyclers	Biorad
LSR Fortessa Cell Analyser	BD Biosciences
FACS MoFlo Astrios	Beckman Coulter
Tablecentrifuge	Eppendorf
Centrifuge	Eppendorf
Vortex	Genius 3
Bioanalyzer	Agilent
Bravo robotic system	Agilent
HiSeq 2000 instrument	Illumina
HiSeq 2500 instrument	Illumina
LSM780 confocal microscope	Zeiss
Revolution spinning disk confocal microscope	Andor
Eclipse Ti-E microscope	Nikon
Live cell chamber	Okolab

Table 2.3: Laboratory materials

Materials	Commercial supplier
1.5 ml reaction tubes	Axygen
0.5 ml reaction tubes	Axygen
Falcon tubes	BD Biosciences
TC dishes	Fisher Nunc
Imaging petri dishes	IBIDI
4-well µ slides	IBIDI
Petridish	Scientific Laboratory Supplies Ltd
Gloves	Microflex
General glassware	Fisherbrand
Parafilm	Pechiney Plastic Packaging
Filter tips	Starlab
Tissues	Kimwipes
1.8 ml Cryotube	Fisher Scientific UK Ltd
8 ml polystyrene round bottom tubes	BD Biosciences
14 ml round bottom tubes	BD Biosciences
Cell strainers	Corning
CellTrics® cell filters	Sysmex
96-well PCR plates	VH Bio Ltd
96-well LoBind PCR plates	Eppendorf
PCR strips and lids	Axygen
Scalpel	Scientific Laboratory Supplies Ltd
High-Sensitivity DNA chips	Agilent

Table 2.4: Reagents for tissue culture

Chemical	**Commercial supplier**
DMEM	Gibco
DMEM/F12	Gibco
Neurobasal medium	Gibco
Fetal Bovine Serum (FBS)	Gibco
Penicillin-Streptomycin	Gibco
L-glutamine	Gibco
Non-essential Amino Acids (NEAA)	Gibco
β-mercaptoethanol	Sigma
N-2 supplement	ThermoFisher Scientific
B-27 supplement	ThermoFisher Scientific
mouse LIF (mLIF)	Stem Cell Institute, Cambridge
PD0325901	Stem Cell Institute Cambridge
CHIR99021	Stem Cell Institute Cambridge
0.05% Trypsin-EDTA	Gibco
0.25% Trypsin-EDTA	Gibco
TrypLE^{TM} Express (1X)	Gibco
Phosphate Buffered Saline (PBS)	ThermoFisher Scientific
Gelatine	Sigma
Optimem	Gibco
FuGENE®	Promega
G418 antibiotic	ThermoFisher Scientific
DMSO	Sigma

Table 2.5: Chemicals and Reagents I

Chemical	Commercial supplier
Propan-2-ol	VWR Chemicals
Ethanol	VWR Chemicals
Glycerol	VWR Chemicals
Agarose	Melford
TritonX-100	Sigma
Tween® 20	Sigma
EDTA	Sigma
EGTA	Sigma
SDS	Sigma
Ampicillin	Life Technologies
DH5α bacteria	ThermoFisher Scientific
SOC media	ThermoFisher Scientific
RNase A	Thermo Fisher Scientific
Collagenase D	Roche
DNAse I	Roche
FCS	Invitrogen
PFA	Sigma
DAPI	Sigma
SlowFade Gold	Thermo Fisher Scientific
BSA	New England Biolabs
Alexa Fluor conjugated secondary antibodies	Molecular Probes
Immersion oil	Nikon
Immersol^{TM} immersion oil	Zeiss
Sucrose	Sigma
HEPES Buffer	Gibco
$MgCl_2$	Life Technologies
KCl	Sigma
Tris-HCl pH 8.3	Sigma
DTT	Life Technologies

Table 2.6: Chemicals and Reagents II

Chemical	**Commercial supplier**
Agencourt AMPure XP beads	Beckman Coulter
MspI	ThermoFisher Scientific
HiFi HotStart Uracil+ ReadyMix	Kapa Biosystems
RLT Plus buffer	Qiagen
RNAse inhibitor	Ambion
RNAse inhibitor	SUPERasin, Life Technologies
Streptavidin-coupled magnetic beads	Dynabeads, Life Technologies
SuperScript II reverse transcriptase	Life Technologies
Superscript II First-Strand 5x buffer	Life Technologies
Betaine	Sigma
Template-Switching oligo	Exiqon
Elution buffer	Qiagen
Ultrapure, nuclease-free H_2O	Life Technologies
Proteinase K	Sigma
dATP	New England Biolabs
TE buffer	Qiagen
NEBuffer 2	New England Biolabs
T4 Polynucleotide Kinase	New England Biolabs
HC T4 DNA ligase	New England Biolabs
Shrimp Alkaline Phosphatase	Fermentas
Exonuclease I	New England Biolabs
Klenow Fragment, exo-	Fermentas
5x Phusion HF buffer	New England Biolabs
dNTPs	New England Biolabs
Phusion polymerase	New England Biolabs
BP Clonase II	ThermoFisher Scientific
LR Clonase II	ThermoFisher Scientific
HyperLadders 100 bp & 1 kb	Bioline
Orange G dye	Sigma

2.2 Methods

Since there is no overlap in methods between chapters, and to simplify access to methods relevant to particular chapters, the Materials and Methods section is broken down by chapter.

2.3 Methods for The Ageing Clock

The methods detailed in this section are adapted and expanded from the published methods section found in Stubbs et al., 2017.

2.3.1 Derivation of unique molecular identifier (UMI) adapters

Cytosine-methylated primers used to derive the adapters were ordered from IDT. The sequences of the two primers were:

Top primer: A/5mC/A/5mC/T/5mC/TTT/5mC/5mC/5mC/TA/5mC/A/5mC/GA/ 5mC/G/5mC/T/5mC/TT/5mC/5mC/GA*T*/5mC/*T

Bottom primer: 5'-Phos/A/5mC/TGNNNNNNNNAGAT/5mC/GGAAGAG/5mC/G GTT/5mC/AG/5mC/AGGAATG/5mC/5mC/*G*A*G

Where N is A, G, T or 5mC.

These primers were annealed at equimolar 100 µM concentration in TE buffer. Annealing was performed by heating the primer mixture to 95°C for 5 mins to remove any secondary structure, then ramp cooling to 16°C at a rate of 1°C per minute. 10 µL of fill-in master mix containing 1x NEB Buffer 2, 1 mM dNTP mixture, 5 µL of Klenow Exo- and ultrapure water. This solution was incubated at 37°C for 60 minutes. This solution was precipitated for 10 minutes with 1x AMPure XP solution and 2x propan-2-ol, then was placed on a magnet and the supernatant removed. The precipitated product on beads was washed 2x with 80% EtOH. The beads were dried for 10 minutes before elution into an A-tailing master mix. This master mix contained 1x NEB Buffer 2, 1 mM dATP and ultrapure water. Once in solution, 40 µL of Klenow Exo- was added (making the solution up to 400 µL) and the solution was incubated at 37°C for 45 minutes. This final product was precipitated over 10 minutes with 1x AMPure XP solution and 2x propan-2-ol, before the solution was placed on a magnet and the supernatant removed. The precipitated product on beads was washed 2x with 80% EtOH. The beads were dried for 10 minutes prior to elution into TE buffer. The adapter product was eluted to a final concentration of 15 µM, with the concentration quantified using a NanoDrop. Adapters were aliquoted and stored at -20°C.

2.3.2 Sample collection - Babraham dataset

C57BL/6-BABR male mice were kept under standard conditions in the Babraham Animal Facility. For the initial dataset derived from these BABR mice (that used for the age-association study), cortex, heart, liver and lung samples were collected at 4 different ages: newborn (<1 week), 14 weeks, 27 weeks and 41 weeks. All tissues were snap frozen directly after isolation. Genomic DNA was isolated from ∼10 mg frozen tissue using the DNeasy Blood & Tissue Kit. A total of 62 samples were collected, processed and further analysed. The resulting dataset is referred here to as the Babraham dataset.

2.3.3 UMI-RRBS library preparation

RRBS libraries were prepared from isolated DNA following published protocols (Meissner et al., 2005). Briefly, RRBS libraries were prepared by *MspI* digestion of 100-500 ng genomic DNA, followed by end-repair and T-tailing using Klenow Exo-. T-tailing was performed instead of the conventional A-tailing as my UMI-adapters are more efficiently derived using an A-tailing protocol. Adapter ligation (UMI-adapters) was performed overnight using HC T4 DNA Ligase, followed by a clean-up step using AMPure XP beads (0.9x). Subsequently, libraries were bisulfite treated according to the manufacturer's instructions (Sigma Imprint Kit; 2 step protocol) and purified using an automated liquid handling robotic system. The libraries were amplified using KAPA HiFi Uracil^{+} HotStart DNA Polymerase, indexing the samples with individual primers. These indexes were defined by the Sanger Institute, referred to as Sanger indexes. All amplified libraries were purified (AMPure XP beads, 0.8x) and assessed for quality and quantity using High-Sensitivity DNA chips on the Agilent Bioanalyzer. High-throughput sequencing of all libraries was carried out with a 75 bp paired-end protocol on a HiSeq 2000 instrument. 75 bp paired-end sequencing was performed based on a cost-benefit calculation conducted on an *in silico MspI* digested mouse genome (GRCm38). Paired-end sequencing was conducted to provide us with double the UMI diversity.

2.3.4 Babraham UMI-RRBS data processing

All Babraham UMI-RRBS datasets had their raw paired-end FastQ files pre-processed to remove the first 13 bp from the 5' ends containing the UMI sequence tags. Both Read 1 and Read 2 UMIs and fixed sequences were written into the read IDs. All samples were subjected to adapter and quality trimming with Trim Galore (`http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`; v0.4.2; options: –paired –three_prime_clip_R1 15 –three_prime_clip_R2 15) to remove potential UMI and fixed tag sequences from the 3' ends. The trimmed files were then aligned to the mouse genome (GRCm38) using Bismark [35] (v0.16.3, default parameters). The mapped sequences were deduplicated by chromosomal position combined with the UMI sequences of both Read 1 and Read 2 (no mismatches tolerated) using the tool UmiBam (`https://github.com/FelixKrueger/Umi-Grinder`; v0.0.1; options: –bam –dual_umi). This deduplication step cannot be achieved without the UMI sequences, since the very nature of RRBS precludes this. These UMI-deduplicated BAM files were then further processed using the Bismark Methylation Extractor (default parameters) to yield Bismark coverage files.

2.3.5 Non-Babraham RRBS data processing

Datasets were processed in the following manner: Raw FastQ files were trimmed with Trim Galore (v0.4.2; parameters: –rrbs) and then aligned to the mouse genome (GRCm38) with Bismark (v0.16.3; default parameters). The aligned BAM files did not undergo deduplication but were processed directly with the Bismark Methylation Extractor (default parameters) to yield Bismark coverage files.

2.3.6 Calling of methylation at single CG sites

For the analysis of age-associated changes in DNA methylation and their subsequent use in the generation of the epigenetic clock (Stubbs et al., 2017), calling of methylation was conducted on an individual cytosine basis. Briefly, mean methylation levels of each cytosine in a CG context that was covered in each sample was calculated from the Bismark coverage files. In addition, a read count was performed for each cytosine in a CG context in each sample, so that filtering could be done based on this information in downstream analysis.

For the derivation of the improved epigenetic predictor described in Results section 3.3, calling of DNA methylation was conducted over individual CpG sites in the genome. Briefly, mean methylation levels of each CpG site that was covered in each sample was calculated from the Bismark coverage files. In addition, a read count was performed for each CpG site, so that filtering could be done based on this information in downstream analysis. Information was calculated over the dinucleotide in this instance in an attempt to reduce issues associated with missingness and to allow more samples to be included.

2.3.7 Statistical analysis of age association at single CG sites

For the statistical analysis conducted on the 62 sample Babraham dataset, I first filtered out positions that had a mean coverage of less than 2 reads or more than 100 reads. This was done to remove spurious reads from library preparation and potential mapping artefacts. For the remaining positions, any positions that were covered with less than 5 reads in a sample were replaced with NA. Before calculating age association, I further filtered such that positions were covered in at least 90% of samples (number of sites = 1,921,569). Ages in days were used when computing the Spearman correlation for each site using the R implementation of the Spearman correlation, and multiple testing correction was performed using the Q-value package (Dabney et al., 2010). For the tissue-specific analysis, a further filtering step was conducted to ensure that there were at least 4 samples being considered for each correlation test. For data exploration, I used PCA analysis of sites that were covered in all samples at 5x (number of sites = 729,785).

2.3.8 Genomic enrichment analysis of significant age-associated sites

Normalised likelihood was calculated as:

$$Normalised\ likelihood, at\ x = (\frac{s}{b} \times \frac{B}{S}) - 1 \tag{2.1}$$

Where:

s= number of significant sites at a given x

S = total number of significant sites

b = number of background sites at a given x

B = total number of background sites

2.3.9 CG scarcity

CG scarcity was calculated as:

$$CG\ scarcity = \frac{200}{(no.\ of\ CG\ sites\ within\ a\ 200bp\ window, centered\ on\ a\ CG\ of\ interest)} \tag{2.2}$$

2.3.10 Gene ontology (GO) analysis of neighbouring genes

Neighbouring genes were defined for single cytosine positions that were within 4 kb of a gene. GO terms were defined using the gprofiler online software (Reimand et al., 2007). For the GO enrichment analysis, a background gene list was made consisting of the neighbouring genes (max distance of 4 kb) for all sites considered in that analysis. Significant GO terms were ordered by p-value and the top six GO terms are shown.

2.3.11 Human-comparative analysis

I defined 1 kb windows around the 21 k CpG sites that are interrogated by both the 27 k and 450 k MethylArray (Horvath, 2013), to ensure that the sites could be faithfully lifted over. These sites were then lifted over from the human genome to the corresponding regions in the mouse genome (GRCm38). Of note, 91 % of the sites selected by Horvath (Horvath, 2013) for the human clock were successfully lifted, i.e. 329 of 353. To be able to compare the Horvath human clock sites (Horvath, 2013) to other sites in the mouse genome, I chose to use all ∼21 k sites, of which I was able to lift over 19 k regions. 175 regions corresponding to the 353 clock sites and 10 k regions corresponding to all ∼21 k sites were covered in the initial 62 sample Babraham dataset. Adding additional datasets (e.g. the Reizel dataset; Reizel et al., 2015) reduced the number of regions covered dramatically. As such, the comparison analysis with the Horvath clock was conducted with the initial Babraham dataset alone.

Using these sites, I first assessed age association by comparing the correlation to age of the Horvath clock regions versus that of random selections of all lifted-over regions. I then built an age prediction model based on the 175 covered regions corresponding to the Horvath clock sites. For this, I built a ridge model as implemented in glmnet by fixing the α parameter to 0. The predictor reaches a median absolute error (MAE) of 11.2 weeks. To compare this to background, I built 1000 random models, picking a random set of 329 regions, regardless of coverage, from the 19 k regions I could lift over. The average MAE was 10.65 weeks in these random models.

	Babraham	**Reizel**	**Cannon**	**Zhang**	**Schillebeeckx**	**Total**
Samples	62	143	36	4	3	248
Male	62	87	36	4	0	189
Female	0	56	0	0	3	59
Ages						
1W	14	14	-	-	-	28
3W	-	18	-	-	-	18
7W	-	-	-	4	-	4
8W	-	10	-	-	-	10
9W	-	-	36	-	-	36
13W	4	-	-	-	-	4
14W	8	-	-	-	-	8
16W	-	-	-	-	3	3
20W	-	84	-	-	-	84
26W	4	-	-	-	-	4
27W	16	-	-	-	-	16
28W	-	13	-	-	-	13
31W	-	4	-	-	-	4
41W	16	-	-	-	-	16
Tissues						
Liver	15	92	36	4	3	150
Lung	16	-	-	-	-	16
Cortex	16	-	-	-	-	16
Heart	15	-	-	-	-	15
Muscle	-	33	-	-	-	33
Cerebellum	-	8	-	-	-	8
Spleen	-	10	-	-	-	10

Table 2.7: Ages and tissue of the samples used to define the mouse epigenetic clock

2.3.12 Defining the published epigenetic predictor of age in mice

Dataset overview for published epigenetic clock

For defining the published epigenetic mouse clock (Stubbs et al., 2017), I included four additional external RRBS datasets, which were downloaded from the GeneExpressionOmnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/): Reizel (Reizel et al., 2015; GSE60012; n = 173); Cannon (Cannon et al., 2014; GSE52266; n = 40); Zhang (Zhang et al., 2016a; GSE80761; n = 4); and Schillebeeckx (Schillebeeckx et al., 2013; GSE45361; n = 23) datasets. A short description of these datasets is provided.

	Babraham	**Reizel**	**Cannon**	**Zhang**	**Schillebeeckx**	**Total**
Maternal HF & Offspring HF	-	-	8	-	-	8
Maternal HF & Offspring LF	-	-	10	-	-	10
Maternal LF & Offspring HF	-	-	9	-	-	9
Maternal LF & Offspring HF	-	-	9	-	-	9
Castration (YM)	-	10	-	-	-	10
Sham castration (YM)	-	8	-	-	-	8
Castration (OM)	-	9	-	-	-	9
Sham castration (OM)	-	8	-	-	-	8
Testorone control (YM)	-	4	-	-	-	4
Ovariectomy (F)	-	3	-	-	-	3
Testosterone ovariectomy (F)	-	4	-	-	-	4
Vehicle ovariectomy (F)	-	6	-	-	-	6

Table 2.8: Treatments used to assess changes to biological age

Reizel dataset (Reizel et al., 2015):

The Reizel dataset was generated from 173 samples originating from four different tissues (liver, muscle, cerebellum and spleen) and collected at six time points ranging from 1 to 31 weeks. The original study investigated gender and tissue specificity of demethylation during ageing. Additionally, a perturbation based on castration and restoring testosterone levels after castration was performed. For the development of my epigenetic mouse clock, the perturbations were not taken into the training but were kept for the test-set. Further information can be found in the original publication (Reizel et al., 2015). 143 samples remained after QC.

Cannon dataset (Cannon et al., 2014):

The Cannon dataset was generated from 40 samples, all liver at the age of nine weeks. The original study investigated the effect of maternal diet on the metabolism of adult offspring. For the published epigenetic clock, I selected part of the data to be in the training set to reflect the nine-week time point (n = 5). The other part of the data was used to assess the effect of diet upon ageing. Further information can be found in the original publication (Cannon et al., 2014). 36 samples remained after QC.

Zhang dataset (Zhang et al., 2016a):

The Zhang dataset was generated from four samples all originating from liver aged between six and eight weeks. The original study investigated methylation differences between different strains of mice and between mouse and zebrafish. For the published epigenetic clock, these samples were used as a validation to see how the predictor works for an unobserved time point. The age of these mice was set to seven weeks. Further information can be found in the original publication (Zhang et al., 2016a). Four samples remained after QC.

Schillebeeckx dataset (Schillebeeckx et al., 2013):

The Schillebeeckx dataset was generated from 23 samples all originating from the liver, the adrenal gland and from endometrial cancer. The mice were ovariectomised at the age of three to four weeks then samples were collected following an additional three months. The original study introduced a laser capture microdissection RRBS method. For the published epigenetic clock I selected the liver samples, which were generated using normal RRBS; three samples remained after QC. These samples were used as a validation to see how the predictor works for an unobserved time point. The age of these mice was set to 16 weeks. Further information can be found in the original publication (Schillebeeckx et al., 2013).

Age prediction

To predict mouse age, I adopted a similar approach to those utilised in human studies (Hannum et al., 2013 and Horvath, 2013), namely an elastic-net regression model. Firstly, I selected the cytosine positions with more than five-fold coverage in all training and test samples used, totalling to 17,992 positions. By selection of the positions available in all datasets, I hope to have a set of methylation sites that will be present in most RRBS studies, irrespective of size selection and data handling. In addition, I filtered out both sex chromosomes (X and Y) and the mitochondrial genome to ensure that the model would be neither sex-specific nor hampered by the unreliability of mitochondrial genome bisulfite conversion. After selection of the sites and samples, I used a quantile normalisation to normalise methylation values, followed by a standardisation that put the mean methylation per site to 0 and the standard deviation to 1.

For the predictor, I used the elastic-net generalised linear model as implemented in the glmnet package (Friedman et al., 2010). In order to optimise the α, which defines the elastic net mixing parameter (from 1 for lasso to 0 for ridge), and to optimise the λ, the regularisation parameter, I used a double-loop cross-validation setup. This setup is described in Ronde et al. (de Ronde et al., 2014). I trained the model to predict the log-transformed mouse age (in weeks); three weeks were added before log transformation of the ages in order to be able to predict sample ages pre-birth.

For the training set, I selected 129 healthy samples from the Babraham, Reizel and Cannon studies. By using an internal ten-fold cross-validation in the inner loop, the optimal α (0.05) and optimal λ (0.93) were identified. The actual performance of the predictor was scored (as assessed by the mean squared error) in the outer loop. After this cross-validation in the training set, I built the final model on all 129 samples. To reach the final model, I took the β values as derived from glmnet for the selected sites (329) and trained a quadratic function using the nls function in R to transform the raw prediction scores (sum of the product of the β weights multiplied by their respective methylation level) to the log age in weeks. This quadratic regression was performed to correct for the age bias in the training and test datasets. The final function used was:

$$\log(age) = 0.1207x^2 + 1.2424x - 2.5440 \tag{2.3}$$

where x is the summed β score per sample.

A set of healthy and treated samples originating from the same three studies, as well as the Schillebeeckx and Zhang samples, were used to assess the usability of the final model. The MAE of the prediction was found to be 3.33 weeks. Furthermore, the model has been used to assess the influence of diet on methylation age using the Cannon training samples, as well as the influence of female castration on the methylation age.

MouseEpigeneticClock script

I have generated an easy-to-use R project to predict methylation age from new samples and deposited it under GNU General Public License at Zenodo (Stubbs et al., 2017) and as a GitHub project: `https://github.com/EpigenomeClock/MouseEpigeneticClock`.

2.3.13 Defining the improved epigenetic predictor of age in mice

Dataset overview for improved epigenetic clock

For defining the improved epigenetic mouse clock, I included the four external RRBS datasets that were utilised in the published epigenetic mouse clock (Materials and Methods section 2.3.12). In addition, I incorporated a further 6 publically available datasets containing predominantly RRBS data, and two datasets of population and single-cell WGBS data. These were also downloaded from the GeneExpressionOmnibus (GEO) database (`https://www.ncbi.nlm.nih.gov/geo/`) or the SRA database (`https://www.ncbi.nlm.nih.gov/sra/`): Cimmino (Cimmino et al., 2015; GSE65919; n = 6); Ghahramani (Ghahramani et al., 2014; GSE50218; n = 24); Kemp (Kemp et al., 2014; GSE48975; n = 15); Auclair (Auclair et al., 2016; GSE71499, n = 10); Petkovich (Petkovich et al., 2017; GSE80672; n = 255); Gravina (Gravina et al., 2016; SRP069120, n = 34); ENCODE mouse WGBS (`https://www.encodeproject.org/matrix/?type=Experiment&replicates.library.biosample.donor.organism.scientific_name=Mus+musculus&assay_`

slims=DNA+methylation&assay_title=WGBS; n = 72) datasets.

A short description of these datasets is provided.

Cimmino dataset (Cimmino et al., 2015):

The Cimmino dataset was generated from 6 samples all originating from lineage negative, Sca1 positive, c-kit negative (LSK) cells derived from mouse bone marrow at the age of 26 weeks. The original study characterised DNA methylation in wild type and TET1 knockout LSK cells. The age of these mice has been set to 25.8 weeks. Further information can be found in the original publication (Cimmino et al., 2015). Four samples remained after QC.

Ghahramani dataset (Ghahramani et al., 2014):

The Ghahramani dataset was generated from 24 samples all originating from bed nucleus of the stria terminalis/preoptic area and striatum (BNST/POA) at 4 and 60 days of age. The original study investigated DNA methylation differences owing to prenatal exposure to testosterone. There were three experimental groups: males, females and females exposed to testosterone. The age of these mice has been set to 1 and 9 weeks respectively. Further information can be found in the original publication (Ghahramani et al., 2014). 24 samples remained after QC.

Kemp dataset (Kemp et al., 2014):

The Kemp dataset was generated from 15 samples all originating from lung tissue from 48 to 55 weeks of age. The original study investigated methylation differences resulting from CTCF haploinsufficiency, of interest due to the susceptibility of CTCF heterozygous mice to neoplasias. Both males and females were assessed in the study. Female mice showed significantly increased susceptibility to neoplasias compared to male mice upon knockdown of CTCF. The ages of these mice were defined individually, between 48 and 55 weeks. Further information can be found in the original publication

(Kemp et al., 2014). 15 samples remained after QC.

Auclair dataset (Auclair et al., 2016):

The Auclair dataset was generated from 10 samples all originating from mixed tissue from E8.5 mouse embryos. The original study investigated the role of the lysine methyltransferase G9a in the control of DNA methylation during embryogenesis. The age of these mice was defined as -1 weeks of age. Further information can be found in the original publication (Auclair et al., 2016). 7 samples remained after QC.

Petkovich dataset (Petkovich et al., 2017):

The Petkovich dataset was generated from 255 samples all originating from whole blood across a wide range of ages. In the original study, this data was used to define a blood-specific predictor of epigenetic age using 90 CpG sites in the genome. Further information can be found in the original publication (Petkovich et al., 2017). 236 samples remained after QC.

Gravina dataset (Gravina et al., 2016):

The Gravina dataset was generated from 34 samples originating from hepatocytes (young=4 months; old=26 months), or fibroblasts from E13.5 mouse embryos. In the original study, this data was used to probe epigenetic heterogeneity at the single cell level. However, bulk datasets of relatively high coverage (more than 15x) were also defined. Further information can be found in the original publication (Gravina et al., 2016). Zero samples remained after QC.

ENCODE

The ENCODE dataset was generated from 72 samples originating from multiple different tissues including: forebrain, heart, hindbrain, midbrain, liver, embryonic facial prominence, limb, neural tube, intestine, kidney, lung and stomach. The ages of these samples range from embryonic day E10.5 to day 0 (post-natal). This data is very high coverage WGBS data that has been made available for the research community as part of the ENCODE project. Further information can be found on the ENCODE website in the mouse WGBS section (`https://www.encodeproject.org/matrix/?type=Experiment&replicates.library.biosample.donor.organism.scientific_name=Mus+musculus&assay_slims=DNA+methylation&assay_title=WGBS`). After QC, all 72 samples were remaining.

Age prediction

To predict mouse age using this new model, a very similar computational workflow to that for the published mouse epigenetic clock (described in Section 2.3.12) was employed. The differences between the two approaches will be described here.

Instead of calling methylation at unique cytosine positions in a CG context, methylation was called over individual CpG sites (i.e. two cytosines). These CpG sites were filtered for sites that exhibited more than five-fold coverage in all training and test samples used, totalling 34,721 sites. In addition, as previously, both sex chromosomes (X and Y) and the mitochondrial genome were filtered out. Following this selection, normalisation and standardisation were performed, similarly to that described previously (Section 2.3.12).

For the predictor, again the elastic-net generalised linear model as implemented in the glmnet package (Friedman et al., 2010) was used and a double-loop cross-validation setup performed (de Ronde et al., 2014).

For the training set, 427 healthy samples from the Auclair, Cannon, Babraham, Reizel, ENCODE and Petkovich studies were selected. By using an internal ten-fold cross-validation in the inner loop, the optimal α (0.1) and optimal λ (0.327168) were identified. Similar to what was conducted for the initial mouse epigenetic clock, the prediction of a log-transformed, linear transformation versus a Horvath-style log-linear transformation (Equations 2.4 and 2.5) was tested. In the Horvath-style approach, *adult.age* is optimised during cross-validation (Section 2.3.12). For this new model, I found the best fit for the data was a Horvath-style log-linear model, where *adult.age* is defined as 46 weeks. As such I utilised this transformation for the final model derivation. If age $<=$ *adult.age*:

$$F(age) = \log(age + 1) - \log(adult.age + 1) \tag{2.4}$$

Else if age $>$ *adult.age*:

$$F(age) = \frac{(age - adult.age)}{adult.age + 1} \tag{2.5}$$

After the cross-validation in the training, the final model was built using all 427 samples. The final model was defined as the β values derived from glmnet for the selected 275 sites. For this model, there was no need to fit a quadratic function using the nls package in R. This is likely due to the offset being removed as a result of increased sample numbers used for training for this predictor.

A set of healthy and treated samples from the same original studies and from Kemp, Ghahramani, Cimmino and Zhang studies were used to assess the usability of the final model. The MAE of the prediction was found to be 5.33 weeks.

2.4 Methods for Single Cell Ageing

2.4.1 Mice

Animals were handled according to national and European Community guidelines, and an ethics committee of the Institut Pasteur (CTEA) in France approved protocols.

2.4.2 Isolation of satellite cells

Six to eight-week-old (young) and 104 to 110 week-old (old) Tg:Pax7-nGFP mice (Sambasivan et al. 2009) were sacrificed by cervical dislocation. *Tibialis anterior* (TA) muscles (Stuelsatz and Yablonka-Reuveni, 2016) were dissected and placed into cold

DMEM. Muscles were then chopped and put into a 50 mL Falcon tube containing 10 mL of DMEM, 0.1% collagenase D, 0.25% trypsin, 10 µg mL^{-1} DNaseI at 37°C under gentle agitation for 30 minutes. Digests were then allowed to stand for 5 minutes at room temperature and the supernatants were collected into 5 mL of FBS on ice. The digestion was repeated 4 times until the muscle was completely digested. The supernatants were subsequently filtered through a 70 µm cell strainer. Cells were spun at 515*g* for 15 minutes at 4°C and the pellets were resuspended in 1 mL freezing medium (10% DMSO in fetal calf serum (FCS)) for long term storage in liquid nitrogen.

Before FACS isolation, samples were thawed in 50 mL of cold DMEM, spun at 515*g* for 15 minutes at 4°C. Pellets were resuspended in 300 µL of DMEM 2 % FCS and filtered through a 40 µm cell strainer. Cells were isolated based on size, granulosity and GFP expression levels using a FACS MoFlo Astrios.

GFP$^{\text{high}}$ subpopulations (henceforth referred to as TA-Hi MuSCs) were FACS-isolated, collected in 300 µL cold DMEM 2% FCS, and re-sorted as single cells into 2.5 µL cold RLT Plus buffer containing 1 U/µL RNAse inhibitor in 96-well LoBind plates. These were then flash-frozen on dry ice and stored at -80°C.

2.4.3 Single-cell combined methylome and transcriptome sequencing (scM&T-Seq) - library preparation

TA-Hi MuSCs that had been single-cell sorted into RLT Plus lysis buffer containing RNase inhibitor in 96 well plates were processed using the scM&T-Seq method (Angermueller et al., 2016), which was developed using existing scG&T-Seq, scBS-Seq and Smart-Seq2 protocols (Macaulay et al., 2015, Smallwood et al., 2014 and Picelli et al., 2013). An adapted description from the methods papers is provided here for clarity, since this scM&T-Seq method has been derived from four separate methods papers and has not previously been fully described itself (Angermueller et al., 2016, Macaulay et al., 2015, Picelli et al., 2013 and Smallwood et al., 2014).

Genomic DNA and mRNA were separated using the Agilent Bravo liquid-handling robot. A modified oligo-dT primer (5'-biotin-triethyleneglycol-AAGCAGTGGTATC

AACGCAGAGTACT30VN-3', where V is either A, C or G, and N is any base; IDT) was conjugated to streptavidin-coupled magnetic Dynabeads as per the manufacturer's instructions. To capture polyadenylated mRNA, 10 µL conjugated beads were added to each well containing cell lysate and this was incubated for at least 15 minutes at room temperature with mixing to prevent the beads from settling. The mRNA was then collected to the sides of the well using a magnet and the supernatant, containing the genomic DNA (gDNA), was transferred to a fresh plate. To maximize gDNA capture, the beads were washed three times with a wash buffer (50 mM Tris-HCl pH 8.3, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT, 0.5% Tween-20, 0.2x SUPERasin RNAse inhibitor) at room temperature. The solution from each wash was pooled with the original supernatant. To minimize sample loss, the same tips were used for all wash steps.

Directly following the last wash, 10 µL of a reverse-transcription master-mix (0.5 µL SuperScript II reverse transcriptase (200 U/µL), 0.25 µL RNAse inhibitor (20 U/µL), 2 µL 5x Superscript II First-Strand Buffer, 0.25 µL DTT (100 mM), 2 µL betaine (5 M), 0.9 µL $MgCl_2$ (1 M), 1 µL Template-Switching Oligo: 5'-AAGCAGTGGTATCAACGC AGAGTACrGrG+G-3' where 'r' indicates a ribonucleic acid base and '+' indicates a locked nucleic acid base; 10 µM, Exiqon), 1 µL dNTP mix (10 mM) and 3.6 µL nuclease-free water) were added to each well. Reverse transcription was performed on a PCR machine for 60 minutes at 42°C followed by 30 minutes at 50°C and 10 minutes at 60°C. The plate was vortexed during this process every 20 minutes, to ensure the beads were maintained in suspension. PCR was then performed immediately by adding the PCR master-mix (12.5 µL KAPA HiFi HotStart ReadyMix with 0.25 µL PCR primer: 5'-A AGCAGTGGTATCAACGCAGAGT-3', 10 mM)) to the 10 µL of reverse-transcription reaction mixture. The sample was then vortexed and thermally cycled as follows: 98°C for 3 minutes, then 24 cycles of [98°C for 15 seconds, 67°C for 20 seconds, 72°C for 6 minutes] and finally 72°C for 5 minutes. Amplified cDNA was cleaned up with 0.8x solution of AMPure Beads and eluted with 25 µL of elution buffer (Buffer EB, Qiagen). The quality of the cDNA libraries was assessed using the Agilent Bioanalyzer. RNA-sequencing (RNA-Seq) libraries were prepared from cleaned up single-cell cDNA libraries using the Nextera XT kit according to the manufacturer's instructions, with the exception that the volumes were reduced to one-fifth of the specified amounts. 384 RNA-Seq libraries were pooled per lane of 125 bp paired-end sequencing on a HiSeq2500.

To avoid batch effects masking biological effects, libraries were prepared from young and old individuals in parallel, in three separate batches. To control for contamination, empty positions in the plates were also prepared and sequenced.

2.4.4 Single-cell combined methylome and transcriptome sequencing (scM&T-Seq) - processing of sequencing data

Libraries were sequenced on the Illumina HiSeq2000 platform using the default RTA (v1.9) analysis software. The sequencing data generated from TA-Hi MuSC libraries includes both DNA methylation data and transcriptomic data. These data types were processed separately, and the link between the datasets was maintained within the sample labelling.

DNA methylation - initial processing of sequencing data

Raw Bisulfite-Seq reads were trimmed to remove poor quality calls, read through adapter contamination and first 6 bp from the 5' end of all reads to remove the sequence bias introduced by the random priming step during the Post-Bisulfite Adapter Tagging (PBAT) library preparation using Trim Galore (v0.4.2, parameters: –paired –gzip –phred33 –clip_r1 6 –clip_r2 6, `www.bioinformatics.babraham.ac.uk/projects/trim_galore/`, Cutadapt version: 1.9.1).

DNA methylation - subsequent processing of sequencing data

Trimmed reads were aligned to the mouse genome in paired-end mode using Bismark v0.16.3 (Krueger and Andrews, 2011) with default parameters plus: –pbat (Bowtie2 v2.2.9). Reads were then deduplicated with deduplicate_bismark, selecting a random alignment for each position that was covered more than once. CpG methylation calls were then extracted using the Bismark methylation extractor (v0.16.3) with the following parameters: –no_overlap –bedGraph.

Filtering steps were performed as described in Results section 4.2.6.

Transcriptome - initial processing of sequencing data

RNA-Seq reads were trimmed using Trim Galore (v0.4.2; `http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`) using default parameters to remove the Nextera adapter sequence and poor quality base calls. Trimmed reads were mapped to the mouse genome (build GRCm38, to which the ERCC spike-ins had been added) using HISAT2 (v2.0.5; options –sp 1000,1000 –no-mixed –no-discordant), guided by Ensembl gene models (release 74).

Transcriptome - subsequent processing of sequencing data

Normalised gene expression counts per cell were defined as: reads per million of mapped reads, using featureCounts in R (Liao et al., 2013) with ensembl GTF file. Filtering steps were performed as described in Results section 4.2.2.

2.4.5 Mapping of public ChIP dataset

The publically available MuSC ChIP dataset from Liu et al., 2013 was downloaded from GEO (accession GSE47362), and uniquely aligned to the mouse genome (build GRCm38) using Bowtie (v1.1.0; options: -m 1 –phred33-quals –strata –best).

Peaks were called using Model-based Analysis of ChIP-Seq 2 (MACS2; v2.1.0.20140616; options –callpeak -f SAM –GSIZE mm –PVALUE 0.001). Peaks were called for H3K4me3, H3K27me3, and H3K36me3.

2.4.6 Cell Cycle Analysis

Classification of TA-Hi MuSCs with respect to cell-cycle was performed using a previously described classification algorithm (Scialdone et al., 2015).

2.4.7 t-Distributed Stochastic Neighbour Embedding (tSNE)-based transcriptome clustering

tSNE dimensionality reduction was performed on the normalised gene expression counts using the *Rtsne* package in R (van der Maaten and Hinton, 2008). Lowly-expressed genes, defined as being expressed in less than three cells, were removed from this analysis.

2.4.8 Transcriptional variability

Genes that were lowly- or highly-variable within a given individual were defined as previously described in Mohammed et al., 2017. Briefly, the squared coefficient of variation was plotted against the mean expression for each gene within an individual. The rolling median was then calculated. Distance from this rolling median was computed for each gene, with directionality considered. Genes were then ordered by this distance and selections made, based on a gene number cut-off, to define highest and lowest variable genes within a given individual.

2.4.9 Cell-to-cell correlation of transcriptional variability

Cell-to-cell coordination of highly-variable genes was defined as previously conducted in Mohammed, H. et al (Mohammed et al., 2017). Briefly, Spearman correlations were calculated in R, using the base *cor* function. For these comparisons, any variable genes that were not expressed in either of the cells being correlated were excluded. These correlation matrices were then visualised as heatmaps. Genes were considered if their mean expression was >1. To calculate heterogeneity from these correlations (ρ) this transformation was performed:

$$d = \sqrt{\frac{(1 - \rho)}{2}} \tag{2.6}$$

2.4.10 Gene Ontology analysis of variable genes

Highly- and lowly-variable genes were assessed for GO enrichment. This was conducted using the gprofiler software (Reimand et al., 2007) and a background list of genes. This background list was defined as all genes that could have been deemed highly- or lowly-variable within a given individual. If there were many significant GO terms, the top six were shown; where there were six or less all were shown.

2.4.11 Conventional DNA methylation variability analysis

Conventional DNA methylation variability analysis, not shown, was performed as previously described in Smallwood, S. et al (Smallwood et al., 2014). Analysis was conducted defining binary methylation values.

2.4.12 Hamming distance (HD) DNA methylation variability analysis

Before computing this Hamming distance (HD) measure of DNA methylation variability, the methylation state at each cytosine within a CG context, within each cell, was defined in binary. To compute the HD, a given set of genomic features had to be defined. For every cell-cell pairwise comparison, the cytosines in a CG context that intersect the feature list and are present in both cells are identified. The difference in methylation state for each single cytosine comparison is then computed and the mean derived from computing the sum of the comparisons. The directionality of the differences is not computed. In addition to the HD, mean methylation across the feature set is defined for each cell within the comparison. This was implemented using a customised PERL script.

For comparison to expectation, a random derivative of this analysis was defined. The workflow for this was the same as that conducted for the actual cell-cell comparison, except that the vectors containing the methylation statuses for each cytosine within a CG context for each cell within the comparison were randomised. This was conducted using the random package within python (Van Rossum and Drake, 2011), 1000 times for each cell-cell comparison. This was implemented using a customised python2 script.

Parallelisation of this analysis was organised using bash.

2.4.13 Subsetting based on DNA methylation

To enable the HD variability analysis to be conducted on feature lists subsetted based on DNA methylation levels, methylation needed to be defined over individual cytosine positions from a 'bulk' sample. This was achieved by counting forward and reverse reads from the coverage files output from bismark (Krueger and Andrews, 2011) for each cell from all individuals, old and young included together, and counting the number of cells that contained information at each position. This was performed using a customised python2 script. For HD variability analyses, this 'bulk' cov file was filtered for positions that contained information from at least 5 cells and each cytosine position in CG context in the genome was binned into 10% intervals of DNA methylation. These bins were then used for feature list subsetting before HD variability calculations were conducted as described in section 2.4.12.

2.4.14 Defining CG-density for maximal coverage

To ensure that as much as possible of the available DNA methylation information could be utilised for the interrogation of CG-density-related behaviour, a customised approach to the calling of CG-density was defined. To ensure CG-density was defined over regions with coverage and not blindly from the genome fasta file, the 'bulk' cov file defined in the subsetting script (Section 2.4.13) was utilised. Using this cov file, regions of CG-density were defined over fixed numbers of 'seen' cytosine positions. Using these fixed coordinates, the true CG-density was then defined from the genome fasta file. Ns in the fasta file were removed so as not to confound the calculation of CG-density. In addition, a maximal size was defined (default=100kb). This was implemented using a customised python2 script.

A secondary python2 script was defined to enable these CG-density regions to be segregated into bins of CG-density as defined by the user in either log10 or linear space. The defaults are 10 bins and log-transformed. In the paper, the values used are specified if they deviate from these defaults.

2.4.15 DNA methylation patterning

A computational approach was developed to assess neighbourhood similarity within a given list of features, within a single cell. This approach computes the absolute difference for every cytosine position relative to every other cytosine position within a given feature. This calculation of distance is subsetted into two based on methylation status of the site being used as the reference (or comparison) site. These metrics are then computed for every single feature within the list of features. This produces a dataframe, with columns corresponding to sum of absolute differences and number of observations, and rows representing distance from the reference position. A similarity score is then calculated for each distance from the reference using this dataframe. The similarity score is defined as:

$$Similarity\ Score = \frac{\sum |difference|}{\sum observations} \tag{2.7}$$

This was implemented using a customised python2 script. This method is agnostic to the strand of the cytosine, although future iterations of the software could take this into account. In addition, the cov files being provided are presently not allele-split, as such a 1-n genome is assumed.

For comparison to expectation, a randomised background was computed. This was computed in the same way as for the real data, except that for each feature within each cell, the positions were randomised. This was a relatively computationally expensive process, therefore in this analysis only ten random comparisons were calculated. This was not a poor background comparison set, since within a given feature there are a relatively small number of cytosine positions covered. Randomisation was performed using the random package in python. This was implemented using a customised python2 script.

Parallelisation of this analysis was organised using bash.

2.4.16 DNA methylation-transcription correlation analysis

The association between DNA methylation at promoters or gene-bodies and expression of the corresponding gene was assessed. This analysis was conducted on a filtered set of genes. The filtering removed lowly-expressed transcripts. This ensured that there were enough cells with coverage for the analysis. The minimal criteria that were chosen for the analysis detailed in Results section 4.2.12 were: 20 cells with methylome coverage (where coverage is defined as >2 cytosines in CpG context present (within the promoter)). For this analysis promoters were defined as +/-2 kb from the transcription start site (TSS). Both Pearson and Spearman correlations were computed using the cor package in R and visualised as volcano plots. A multiple-testing correction was applied using the p.adjust package in R and a false discovery rate (FDR) of 10%.

2.4.17 DNA methylation-transcription feature-based analysis

The relationship between DNA methylation variability (measured using the HD metric) and transcriptional variability was calculated. To conduct this analysis, genes were classified into states of variability. For each of these bins, feature lists were defined for the promoters or gene bodies of these genes and DNA methylation variability calculated.

2.5 Methods for Live Cell DNA Methylation

2.5.1 Phusion PCR

The standard Phusion PCR reaction was: 10 µL 5x Phusion HF buffer, 2 µL 5 mM dNTPs, 1 µL 10 µM primer mix, 1 µL DNA (1:1000 dilution of miniprep if used), 0.5 µL Phusion polymerase (2,000 U/ml), 36.5 µL H_2O. Variations of the following cycling conditions were used: 30 seconds at 98°C followed by 35 cycles of (10 seconds at 98°C, 10 seconds at temperature 3°C above the lower primer T_m, 3 minutes at 72°C) then a final 5 minute incubation at 72°C before holding at 4°C.

2.5.2 BP and LR cloning

For BP or LR cloning, the following were added to a 1.5 mL microcentrifuge tube at room temperature and mixed: 1-7 µL attB-PCR product (>10 ng µL^{-1}; final amount 15-150 ng), 1 µL Donor vector (150 ng µL^{-1}), and TE buffer pH 8.0 to 8 µL total volume. The BP or LR Clonase II enzyme was thawed and vortexed. 2 µL of the enzyme was added to each reaction and mixed well. The reaction was then incubated at 25°C for 1 hour or overnight for larger fragments. The reaction was stopped by addition of 1 µL of Proteinase K and incubation at 37°C for 10 minutes. Product was then transformed into bacteria.

2.5.3 Bacterial transformation

Plasmids were transformed into DH5α bacteria. 50 µL of bacteria were incubated with 10 µL of DNA on ice for 10 minutes. A 40-second heat shock was performed at 42°C. Samples were incubated on ice with 200 µL SOC for a further 5 minutes, followed by incubation in a shaker at 37°C for 1 hour. Bacteria were then plated on agar plates containing the relevant antibiotic and incubated at 37°C overnight. Bacterial colonies were picked and grown up in LB liquid culture for plasmid DNA extraction, validation and use.

2.5.4 Plasmids defined for future applications

A number of plasmids were cloned that could be useful for future experiments measuring more targeted DNA methylation changes in living cells. A brief description of these plasmids is provided in Table 2.9.

Protein	**Backbones**	**Linker lengths (aa)**	**Fluorophores**	**Original constructs**
MBD of MBD1	pDONR and pDest RESISTANCE	5, 10, 15	Venus, Cerulean	(Koushik et al., 2006 and Yamagata, 2010)
Histone H2A	pDONR and pDest RESISTANCE	5, 10, 15	Cerulean	(Koushik et al., 2006)
dCas9	pDONR and pDest RESISTANCE	5, 10, 15	Cerulean	(Chen et al., 2013a and Koushik et al., 2006)

Table 2.9: Plasmids generated for future FRET-style experiments

The starting plasmids for this cloning were predominantly gifts. mVenus N1 was a gift from Steven Vogel (Addgene plasmid # 27793), and mCerulean N1 was a gift from Steven Vogel (Addgene plasmid # 27795). MBD1-eGFP was a gift from Kazuo Yamagata. mCerulean-H2A-10 was a gift from Michael Davidson (Addgene plasmid

55373). pSLQ1658-dCas9-EGFP was a gift from Bo Huang & Stanley Qi (Addgene plasmid # 51023).

2.5.5 Mouse embryonic stem cell culture

Mouse embryonic stem cells (mESCs) and mutant derivatives thereof were either grown in 15% serum media (450 mL DMEM, 75 mL fetal bovine serum (FBS), 5 mL 100x Penicillin-Streptomycin antibiotic, 5 mL 100x L-glutamine, 5 mL 100x NEAA, 500 µL β-mercaptoethanol, 500 µL 1000x mLIF) or 2i/LIF media (250 mL DMEM/F12, 250 mL Neurobasal, 2.5 mL N-2 Supplement, 5 mL B-27 Supplement, 5 mL L-Glutamine, 5 mL 100x Penicillin-Streptomycin antibiotic, 500 µL β-mercaptoethanol, 500 µL 1000x mLIF, 3 µL mL^{-1} of 10 mM CHIR99021 and 1 µL mL^{-1} of 10 mM PD0325901.

When passaged, serum-grown mESCs were washed once with 1x PBS following removal of media. PBS was removed and cells were incubated with trypsin at 37°C for 5 minutes or until cells had detached. Four volumes of media were added to inactivate the trypsin. Appropriate volumes of the cell suspension were then transferred to new plates to achieve the desired cell density. New plates had been pre-coated with 0.1% gelatin for 15 minutes , then the gelatine removed prior to use.

When passaged, 2i/LIF-cultured mESCs were washed once with 1x PBS following removal of media. PBS was removed and cells were incubated with TrypLE for 1 minute at room temperature. Cells were then resuspended in PBS and spun down at 200g for 3 mins. The supernatant was removed and the cell pellet was washed with PBS and spun down again. Following supernatant removal, the cell pellet was then resuspended in 2i/LIF media and replated at the desired density. Similar to serum-cultured mESCs, 2i/LIF-cultured mESCs were cultured on gelatin.

2.5.6 Freezing and thawing mESCs

Cells were prepared for freezing as described for passage procedures, with the exception that for the last resuspension step cells were resuspended in 0.5 mL of media. To this

0.5 mL, 0.5 mL of 2x freezing media (for serum-cultured mESCs 50% serum media, 30% FBS and 20% DMSO; for 2i/LIF-cultured mESCs 80% 2i/LIF media and 20% DMSO) was gently added. 2x freezing media was stored at 4°C and was used within 2 weeks. Following addition of freezing medium, samples were transferred to labelled cryovials before being placed at -80°C in a Mr Frosty, allowing the cells to cool slowly. After 2-3 days, cryovials were transferred to liquid nitrogen for long-term storage.

Thawing of cells was conducted in a timely manner to minimise damage to the cells. Cryovials were thawed at 37°C for one minute, then 4 ml of appropriate media was added and the samples transferred to gelatinised plates. The following day the media was changed. Cells were passaged at least once before being used for experiments.

2.5.7 Transfection

Transfections were performed using FuGENE® in a 6-well plate format. 250 µL of Opti-MEM™ was added to tubes containing 2 µg of DNA in total. In separate tubes 250 µL of Opti-MEM™ was incubated with 6 µL of FuGENE®. Both sets of tubes were left to incubate at room temperate for 5 minutes, then the two mixes were combined and incubated together for a further 20 minutes. This transfection mix was then added to the respective wells of cells.

2.5.8 Stable cell line derivation

To generate stable cell lines, cells were selected for 10-14 days with G418 post-transfection. Selected cells were sorted to enrich for highly-expressing GFP-positive cells in order to facilitate imaging experiments. In preparation for analysis by fluorescence-activated cell sorting (FACS), cells were washed with PBS, resuspended in PBS 0.1% BSA, and filtered through a 50 µm filter to ensure a single-cell suspension. GFP-positive cells were identified by comparison to a wildtype mESC line.

2.5.9 Preparation of cells for live cell experiments

Cells were split as described in section 2.5.5, except for a few subtle differences. Firstly, cells were filtered through a 50 µm filter to ensure a single-cell suspension prior to re-plating. Secondly, only a very thin coating of gelatin was applied to the plate surface, to ensure that this did not interfere with the imaging. Lastly, cells were re-plated onto IBIDI dishes or 4-well µ-slides. These formats were chosen due to their compatibility with high-quality imaging at high magnifications and with the microscope oils utilised for high magnification imaging. Upon re-plating, cells were allowed to adhere to the gelatin surface in the standard cell culture incubators for one hour before they were taken to the microscope. The cells were then given a further 30 minutes to adjust to the environment of the live cell chamber on the microscope prior to imaging.

2.5.10 Immunofluorescence and imaging

Antibody staining was performed as previously described (Santos et al., 2003) on mESCs grown on coverslips, after fixation with 2% PFA for 30 minutes at room temperature. Briefly, cells were permeabilised with 0.5% TritonX-100 in PBS for 1 hour; blocked with 1% BSA 0.05% Tween-20 in PBS (BS) for 1h; incubated with the appropriate primary antibody diluted in BS; washed in BS; and incubated with secondary antibody for 30 minutes. For simultaneous detection of DNA methylation, after the first round of antibody staining samples were washed in PBS, post-fixed in 2% PFA for 10 minutes, treated with 2N HCl for 30 minutes at 37°C and washed extensively with PBS before incubating with anti-5mC diluted in BS. All secondary antibodies were Alexa Fluor conjugated, diluted 1:1000 in BS. Incubations were performed at room temperature unless otherwise stated. DNA was counter-stained with 5 µg/ml DAPI in PBS. Single optical sections were captured with a Zeiss LSM780 microscope (63x oil-immersion objective) and the images pseudo-coloured using Adobe Photoshop CS4. For visualization, images were corrected for brightness and contrast, within the recommendations for scientific data. ImageJ 1.51p (NIH) was used for fluorescence semi-quantification and the Colocalization_Finder plugin (https://imagej.nih.gov/ij/plugins/colocalization-finder.html) for colocalisation analysis.

Antibody	**Cat. no.**	**Company**	**Dilution**
Anti-Dnmt1	ab87654	Abcam	1:1000
Anti-5mC	BI-MECY-0100	Eurogentec	1:500

Table 2.10: Antibodies used for immunofluorescence

2.5.11 Fluorescence Recovery After Photobleaching (FRAP) imaging and primary analysis

MBD1-eGFP mobility was assessed in a number of different mESC lines. The regions chosen to be photobleached were non-peripheral regions of the nucleus to ensure that recovery could occur unhindered in all directions.

Cells were imaged using an Andor Revolution spinning disk confocal microscope, comprising Nikon Ti-E frame, Nikon 100x 1.4 NA plan apochormat lens or Nikon 60x 1.45 plan apochromat lens, Yokogawa CSU-X scanhead, Andor laser combiner, Andor FRAPPA photobleaching unit and Andor iXon 897 EM-CCD camera. The system was configured and controlled using Andor iQ software. GFP was imaged with 100 ms exposure (EM gain set to 300) using 488 nm laser excitation (set at 6%) with emitted light filtered using a 525/40 nm bandpass filter. The photobleached regions chosen included regions containing heterochromatic foci and regions without. The photobleached area was 15x15 pixels (where the diameter of one pixel is 0.166 667 µm at 100x magnification and 0.266 µm at 60x magnification), except where specified in the control experiments. Twenty images were acquired prior to photobleaching using continuous capture. Photobleaching was performed using 50% laser power and a pixel dwell time of 40 µs, repeated twice. A further 480 images were then acquired using the imaging settings described above to assess recovery. Two FRAP experiments were conducted per cell to verify robustness of the method. A live cell chamber surrounds the microscope, and temperature and CO_2 levels were maintained at 37°C and 5% respectively.

The fluorescence of the photobleached region, of the entire nucleus, and of a background region outside of the cells in the image was quantified for each frame in the image stack. This was conducted using FIJI (Schindelin et al., 2012). Briefly, regions of interest (ROIs) were annotated and metrics extracted across each image stack. These three ROI metrics were stored in a .txt file and exported. The Full Normalisation detailed in Results section 5.2.3 was performed on these to ensure independence from the total fluorescence present in the nucleus of a given cell, and to enable us to disregard differences in recovery due to differences in the proportion of protein in the immobile fraction. The contribution of newly-synthesised protein to the recovery was assumed to be negligible, owing to the short time window of the experiment. The recovery was fit as described in Results section 5.2.3.

2.5.12 Gaussian Mixture Model (GMM) to describe DNMT1-iKO

To model the behaviour of the DNMT1-iKO, a Gaussian Mixture Model was defined from the data. Four states were defined. The parameters for each state were initialised using a K-means based clustering approach and these initialised parameters were then refined by the GMM and state-likelihoods computed. The most likely state was then assigned to each cell.

2.5.13 Differential Dynamic Microscopy (DDM) imaging

MBD1-eGFP mobility was assessed in a number of different mESC lines. Fluorescence images were taken using a Nikon Ti-E equipped with a 100x 1.4 NA plan apochromat lens, Lumencor Spectra-X LED light source and Hamamatsu Orca Flash 4.0 CMOS camera. GFP excitation used the 470 nm LED, with a 488/10 bandpass excitation filter and 525/50 bandpass emission filter. The system was configured and controlled using Nikon Elements software. It was imperative that each image contained only one single cell. One thousand frames were acquired for each image of one cell. These frames were acquired at 50 Hz, i.e. 20 images per second. This frame rate was achieved by reducing the field of view to 512 x 512 pixels, with the camera attached to the PC using CameraLink. Images were saved from the microscope in .nd2 files. An OKO lab live cell chamber surrounded the microscope, and temperature and CO_2 levels were maintained at 37°C and 5% respectively, except where detailed in the results.

Nikon Elements JOBS software was used to enable multiple cells to be imaged in the same experiment. This software marked cell positions using x,y,z coordinates and allowed for stage movement to different locations between image acquisition bursts.

For experiments where single cells were imaged over time, Nikon Elements JOBS software was utilised again. However, it was run separately for each time interval due to cell movement between image intervals. As such, cellular positions were recalculated prior to DDM imaging using a reduced intensity LED excitation and longer exposure time (100 ms). This was done to ensure minimal quenching prior to the DDM imaging.

Images were converted from .nd2 files to .tiff files using FIJI (Schindelin et al., 2012). Primary and secondary analysis were developed as part of this thesis and as such is described in Results section 5.2.11.

Chapter 3

The Ageing Clock

3.1 Introduction

DNA methylation has been shown to be an accurate readout of chronological age and possibly biological age in humans. At present, however, there is no method to predict chronological and biological age in an experimentally tractable model system. This means that we are both unable to probe the mechanisms of this epigenetic predictor but also that it is not presently possible to utilise the time-saving benefits that such a clock can produce in ageing studies in a mammalian model system setting. Such as the ability to rapidly test ageing interventions in an *in vivo* model. In this chapter, I will define an epigenetic age predictor for mice using data from a modified form of reduced representation bisulfite sequencing (RRBS).

RRBS is a simple means of genomic enrichment based on the premise that the genome has a non-random base composition. By judicious choice of restriction enzyme, these base composition biases can be exploited to enrich for sites of interest. The most commonly used RRBS method for enriching cytosine-guanine dinucleotide sequences (CpGs) in the genome, is to digest isolated DNA with the methylation insensitive restriction enzyme *MspI*. Enzymes such as *BglII*, *XmaI* and *TaqαI* have also been used (Lim et al. (2016), Meissner et al., 2005 and Tanas et al., 2017). *MspI* cuts at the recognition sequence C|CGG. This *MspI* digestion of the genome can capture 10% of

CpGs in the genome, whilst reducing the number of fragments sequenced by ∼30 fold compared to whole genome bisulfite sequencing (WGBS). This cost reduction means that it is possible to use RRBS to obtain a far higher coverage per sample than is achievable using WGBS. This is important for epigenetic age prediction, because the magnitude of changes observed at single CpG sites in these predictors is often small and as such would require high depth sequencing to observe the changes. The use of RRBS also makes sequencing more efficient, since reportedly 70-80% of standard sequencing reads contain no CpGs and thus no relevant information.

There are many additional approaches to enrich genomic regions of interest (Martin-Herranz et al., 2017). However, I chose RRBS over these for a number of reasons. Firstly, RRBS is a depletion-based method, rather than a positive selection-based approach. This means that the sites of interest are not actively acquired in the selection step and as such fewer biases are introduced, making these approaches more readily quantifiable. Secondly, unlike some of the other methods such as antibody pulldown based approaches (e.g. using an anti-5mC antibody), RBBS can provide information at single base resolution. Thirdly, RRBS is able to provide access to a large number of CpG sites (typically >2M sites) in a reproducible manner. This is not the case for amplicon-based polymerase chain reaction (PCR) approaches, which can only assay a small number of sites, or antibody-based approaches, which do not provide any information on the regions that will be sequenced. Fourthly, there are already high quality RRBS datasets generated that are well-documented and include aged mouse samples in the public domain. As such, in new RRBS experiments, these datasets can be incorporated into the analysis to reduce biases in mice or sample preparation. Lastly, although mouse samples have previously been run on the human MethylArray systems, a MethylArray was not utilised, because a dedicated instrument is not currently available for the mouse genome and as such the number of CpGs that could be interrogated would be greatly reduced due to lack of homology. However, would such an array-based system have been available, this would have been preferable owing to the higher reproducibility. This is due to the Poisson distribution of reads sequenced, inherent in the sequencing process, which reduces the number of sites that can be utilised in a linear modelling-based approach without requiring imputation.

3.2 Results

3.2.1 Enabling deduplication in the RRBS datasets

In the human epigenetic predictors, it is common for very small changes in the individual methylomes to result in large changes in the predicted age. As such, I decided to refine the current RRBS protocol to incorporate unique molecular identifiers (UMIs) into the library preparation, to ensure that I could subsequently remove any duplicate reads that could mask subtle changes in DNA methylation. A schematic of this is shown in Figure 3.1.

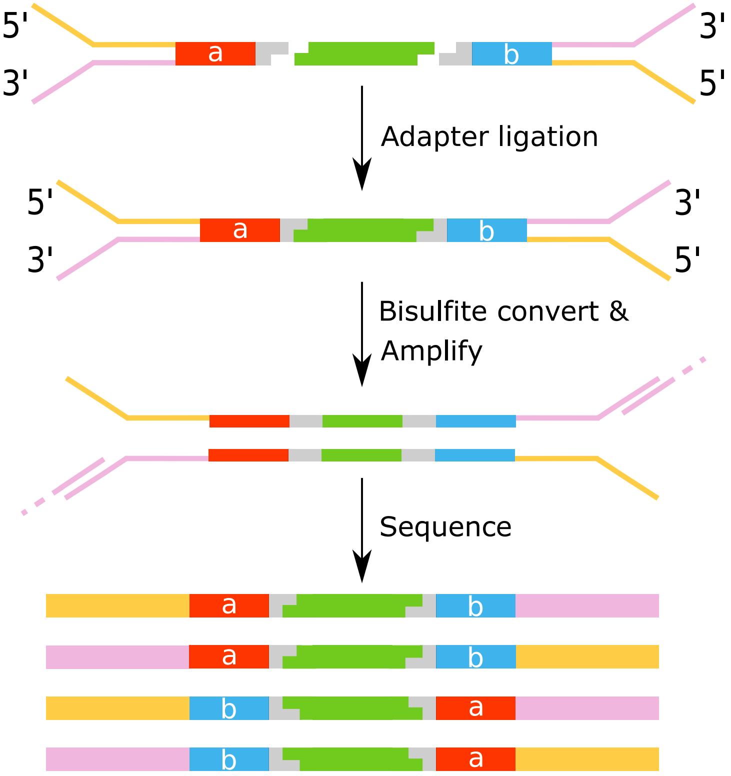


Figure 3.1: Deduplication from paired UMIs: Schematic of the library process utilising UMIs and paired-end next generation sequencing. In brief, adapters containing UMIs are ligated to genomic DNA. This adapter ligated DNA can then be bisulfite treated and amplified in a stranded manner. Upon sequencing this will result in 4 independent types of information that can be deciphered from the directionality of the PCR and the UMI. Two example UMIs are shown, labelled A and B.

In WGBS, deduplication is usually achieved by removing reads that start and end at the exact same coordinates in the genome. In RRBS, this is not a feasible method for deduplication, since all fragments that cover the same region of a genome will likely all originate from digested fragments that start and end at the exact same positions. The problem of deduplication is greater in an RRBS setting than a WGBS setting, owing to the increased number of PCR cycles performed in the RRBS setting in comparison to the WGBS setting, inherent in the reduced number of fragments that will be sequenced from the same starting amount of DNA.

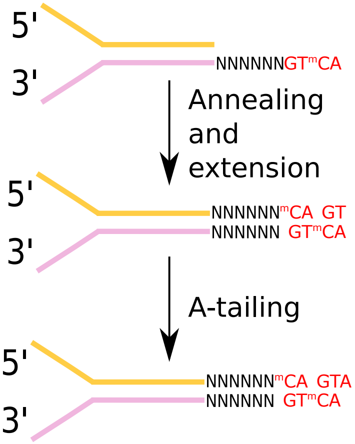


Figure 3.2: Derivation of double-stranded UMI adapters: Schematic of the steps involved in deriving double-stranded UMI adapters. Oligonucleotides are annealed together. This double-stranded oligo is then filled-in ensuring that the two strands containing the UMI (depicted with Ns) are complementary. These double-stranded adapters can then be A-tailed to improve ligation efficiency.

To uniquely label all fragments in an RRBS library, I modified the adapters that are ligated to the ends of the restriction digested DNA in such a way that they included eight random bases from the nucleotides adenine, guanine, thymine and 5'-methylcytosine (A, G, T, 5mC, respectively). I chose 5'-methylcytosine over cytosine, because these ligated adapters will subsequently be bisulfite converted, which would convert any cytosines in the adapter to uracils and these would then be amplified as thymines in turn. To ensure that the adapters would contain complementary UMIs on both strands of the adapter, DNA oligonucleotides (oligos) were ordered that represented the top and bottom strands of the future adapters. The top DNA oligo was designed to only contain the standard illumina adapter sequence, whereas the bottom DNA oligo was designed to contain the standard illumina sequence, followed by an 8N sequence, followed by a 4-base constant CAGT sequence. The purpose of the 4-base constant sequence was to reduce any ligation bias during the ligation of the 8N sequence. Such biases could be dependent on the sequence of the genomic DNA being ligated and could affect the complexity of the subsequent library. These oligos were slowly annealed to favour the correct formation of the partially double-stranded adapter. Subsequently, the top strand of the adapter was filled in using Klenow exo- using A, 5mC, G and T in order to ensure that the sequence complementary to the 8N barcode was obtained on the top strand and that the complementary nature would remain following bisulfite treatment. Following the fill in reaction, an A-tailing reaction was performed on the adapter to improve the ligation efficiency with the T-tailed genomic DNA. A schematic is shown in Figure 3.2.

Preliminary sequencing experiments showed that the 8N barcode was mostly random, with a slight A-bias. This A-bias disappeared when subsequent oligo batches were utilised, suggesting that it could have been due to a bias in base availability during oligo synthesis. Next, I wanted to ensure that there was no bias in the sequence content of the barcode in relation to the sequence content of the genomic fragment that it was ligated to. I therefore compared the GC-content of the DNA and adapter and found no sequence bias (Figure 3.3).

Following these initial whole genome experiments, I tested the adapters in an RRBS experiment and found that I could generate libraries. Importantly, quantification of the libraries showed that from a similar number of cycles these libraries were equivalent to those generated with the standard illumina adapters, suggesting that my A and T-tailing reversal was no less efficient than their approach.

In collaboration with Felix Krueger, creator of Bismark (Krueger and Andrews, 2011), I next derived a computational approach for calling and removing duplicates. In essence, fragments were first compared by their start and end positions in the genome. Then these putative duplications were assessed to see whether their read 1 (R1) and read 2 (R2) barcodes (16N in total) were highly similar (which I defined as identical or one base different) both in a direct match setting and in a reverse complement setting. A one base mismatch was chosen from empirical evidence as this showed that any more

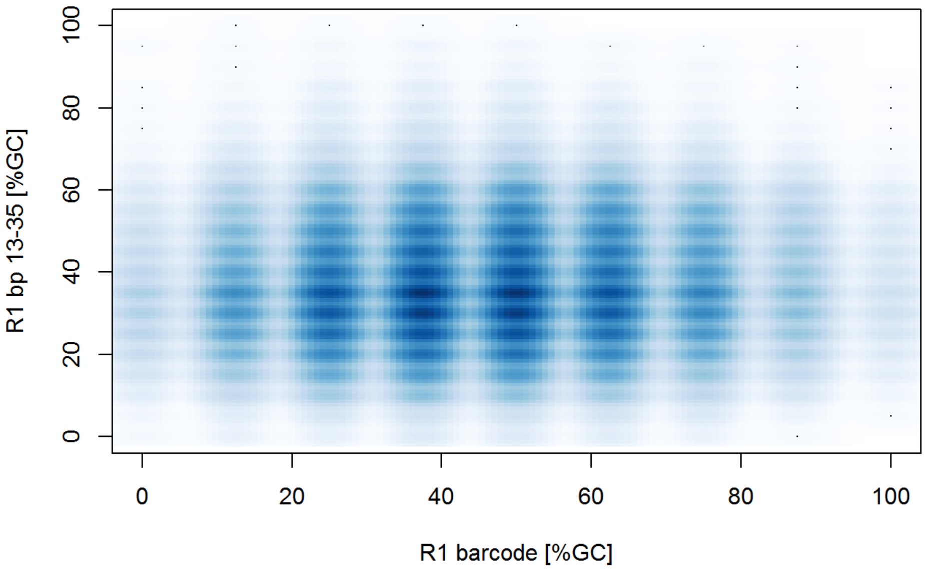


Figure 3.3: There is no GC bias in UMI incorporation: Scatterplot depicting the GC-content of the UMI against that of the neighbouring ligated genomic sequence. There was no relationship detected.

subsequent mismatches would likely be the result of a separate barcode.

Overall, these observation show that I have successfully developed a method that is capable of deduplicating RRBS data. Additionally, though not detailed in this thesis, I have shown that these adapters can be used to call SNPs, rare base modifications, and hemi-methylation.

3.2.2 DNA methylation correlations with age in mice

Before deriving an epigenetic predictor in mouse, I first set out to confirm and expand upon what is known about age-associated changes in DNA methylation in mice. To this end, I collected liver, lung, heart and brain (cortex) samples from newborn to 41-week old mice. To ensure that the effects I was seeing were related to age and not an additional variable, such as genetic or hormonal variability within the dataset, I restricted our cohort of mice to inbred C57BL/6 BABR male mice and sampled 3-5 animals per time point. In total, I successfully collected 62 samples and extracted genomic DNA for methylation analysis from these samples (Figure 3.4).

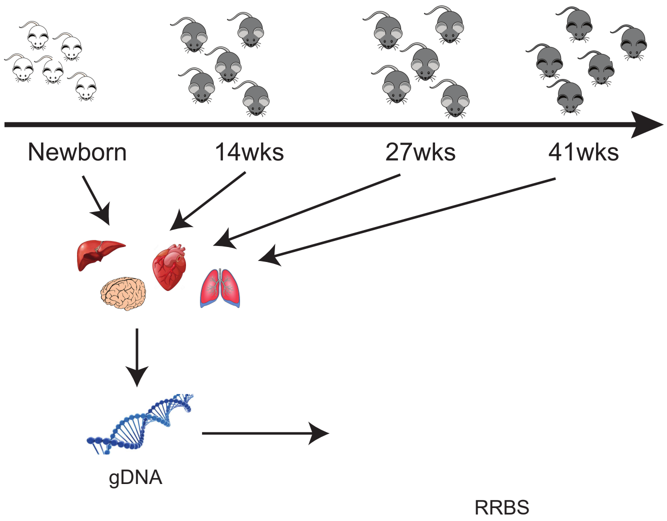


Figure 3.4: Schematic of the Babraham dataset: Graphic summary of the Babraham dataset, with mouse age differentiated by colour. DNA was isolated from liver, lung, heart and cortex of newborn mice and mice aged 14, 27 and 41 weeks, and reduced-representation bisulfite (RRBS) libraries made from this.

I subsequently generated UMI-style RRBS libraries of all samples to assess DNA methylation changes at a wide range of CpG sites. To ensure that there was no batch effect in the sample preparation that could be misconstrued as ageing-related methylation changes, samples were prepared in one batch. These libraries were sequenced to an average of 15x genomic coverage. This level of coverage was a compromise between the cost of the sequencing, the number of samples that I wanted to interrogate, and the magnitude of changes that I was hoping to be able to detect. Per sample I achieved an average coverage of more than 1.23 million CpG sites with at least 5-fold coverage and of these 0.73 million CpG sites had more than 5-fold coverage in all samples analysed.

I next analysed the global CpG methylation levels of the samples and found that the newborn samples had average global methylation levels of (43%) methylation (Figure 3.5A). I did not observe any differences between the global methylation levels found in the different newborn tissues. In contrast, in the samples from adult mice, I found that the global methylation levels were slightly higher than those of the newborn samples at 45%. Again, there were no major differences among the different adult mouse tissues nor were there difference between the methylation levels of the mice across the age groups. In WGBS datasets global levels of methylation are typically between 70 and 80% (Ehrlich et al., 1982). The reason for the low levels of global methylation seen in these samples is due to the fact that RRBS enriches for CpGs in CpG Islands, which tend to be hypomethylated (Meissner et al., 2005). To check that the bisulfite conversion of the samples was successful, I assessed the levels of non-CpG methylation. I defined poorly converted samples as having non-CpG methylation levels greater than 10%. This highlighted that a couple of samples had been poorly converted and these were subsequently discarded from my further analyses. Overall, I observed very low levels of non-CpG methylation in the liver, lung and heart samples (∼0.4%) with slightly higher levels of non-CpG methylation in the brain (∼1.4%). In addition, when the cortex samples were grouped by age, I found that newborn cortex samples had far lower levels of non-CpG methylation than the older samples (Figure 3.5B). This is consistent with the idea that de novo methylation in non-dividing cells results in accumulation of CHH methylation (H=C, A, or T; Lister et al., 2013). Lastly, to validate that all samples were correctly assigned and that there had not been any confusion in the annotation prior to sequencing, I performed a Manhattan distance-based hierarchical clustering of the most variable sites in the dataset (Figure 3.5C). This approach selects for sites that would likely fall within CpG Islands and that are differentially methylated dependent on the tissue type. From this cluster-based analysis, I found that I could correctly group the majority of the samples by their key tissue-specific methylation signatures. The only samples that did not cluster correctly were the newborn lung samples, which grouped together with the heart samples. The reason for this discrepancy in the newborn lung samples could be a result of some contamination during the tissue extraction, or it could reflect the additional time required for the lung to fully acquire its adult methylation signature (Christensen et al., 2009).

Having verified the expected global and tissue-specific properties of the DNA methylomes in our dataset, I then wanted to assess whether I could identify not merely regions of the genome that were hypo- or hypermethylated with age, but whether I could identify single cytosine positions in the genome that were hypo- or hypermethylated with age. The main reason for wanting this single position resolution was to ensure that any

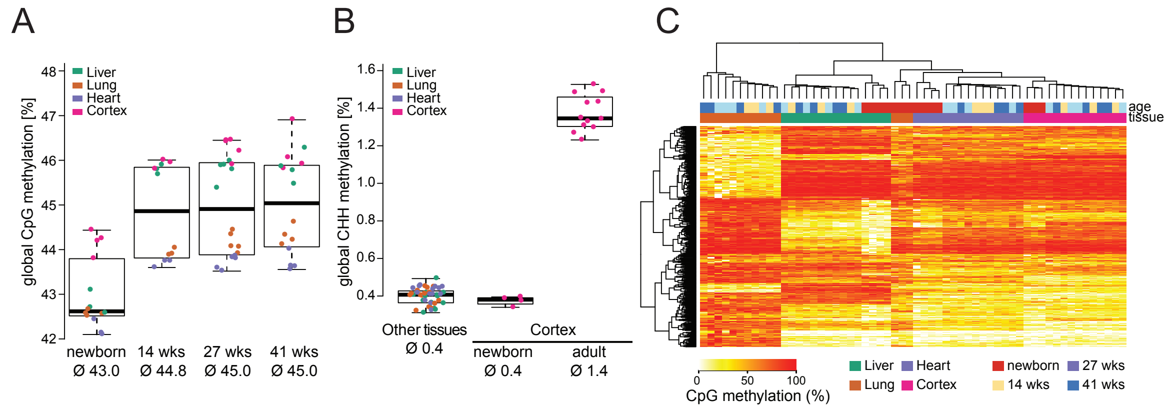


Figure 3.5: QC of the Babraham dataset: (A) Percentage global CpG methylation level of different tissues at different ages, ordered by age (weeks). Mean CpG methylation levels were calculated for each sample, with colour indicating sample tissue. (B) Global levels of DNA methylation in a CHH context in different tissues at different ages. 'Other tissues' includes heart, liver and lung samples of all four ages whilst 'Cortex' was segregated into newborn (less than one week old) and adult (14, 27 and 41 week old) samples. (C) Hierarchical clustering using Manhattan distance clusters samples of different ages predominantly by tissue of origin. Age and tissue type of each sample is colour labelled.

age-associated methylation changes that I identified were not confounded by coverage. This would occur due to the spacing of the CpG sites in the genome compounded by the frequency of *MspI* restriction sites in regions of high CG density resulting in varying coverage of neighbouring cytosines in CpG context. In addition, I wanted to be sure that I wasn't diluting any potential site-specific changes by assessing changes regionally. From a correlation analysis of DNA methylation changes with age, I was able to identify a substantial number of sites that were significantly correlated with age in a tissue independent manner (Spearman's correlation, with a multiple testing corrected p-value <0.05). I conducted this analysis in a tissue independent manner, because I wanted to assess the likelihood of defining a multi-tissue epigenetic predictor of age in the mouse. I conducted this analysis both with and without the newborn samples to ensure that any developmental-specific changes were abrogated. Unfortunately, owing to the reduced number of time points, I was only able to identify sites that were nominally significantly associated with age, but none of these passed my multiple testing cut-off. An example of a tissue-independent correlation is shown for the correlation analyses conducted with and without newborn samples (Figure 3.6A and B respectively).

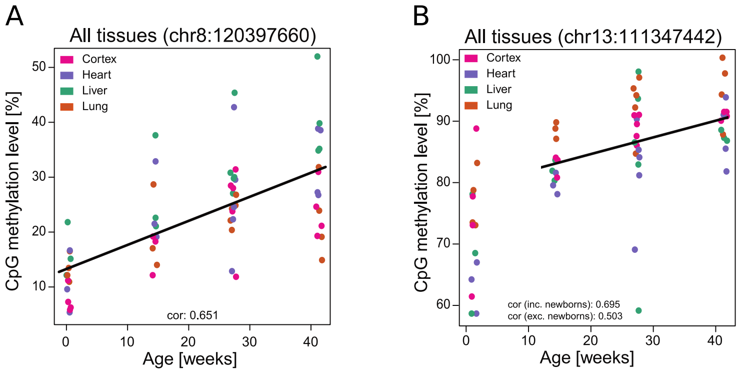


Figure 3.6: Example tissue independent correlations: (A) An example single CG site, chr8:120397660, whose methylation level has a Spearman correlation with age of 0.651. Tissue of origin is specified by colour and jitter is purely for aesthetic purposes. (B) shows the same for a CG site, chr13:111347442, that is identified as correlated with age when newborn samples are excluded from the analysis. Spearman correlations when newborn samples are included or excluded are detailed.

In addition, I conducted a correlation analysis in a tissue-dependent manner to assess whether sites were also correlated with age in a tissue specific manner. This analysis was again conducted with and without newborn samples. In this analysis, I found a large number of sites that were exclusively tissue specifically correlated with age (Figure 3.7). From Figure 3.7B, it is also interesting to note that some of the sites that are correlated with age in a newborn-excluded manner tend to exhibit a change in correlation directionality between postnatal-development (<1 to 14 weeks of age) and adulthood (14 to 41 weeks of age).

Looking more generally at the tissue-independent correlations I find a skew in the newborn-inclusive correlation analysis for sites that are gaining methylation or becoming hypermethylated with age (Figure 3.8A and B). This skew is removed when the newborns are excluded from the analysis and this is likely a reflection of the global increase in methylation that is occurring in the first 14 weeks after mice are born (Figure 3.8C and D).

Interestingly, when the correlation coefficients for all sites that were present in both the newborn inclusive and exclusive, tissue-independent correlation analyses are compared, I observe a strong tendency for the directionality of the coefficient to be maintained, as is highlighted by the shape of the distribution (Figure 3.9). However, it is interesting to note that similar to the tissue-dependent correlations, there are sites that exhibit a

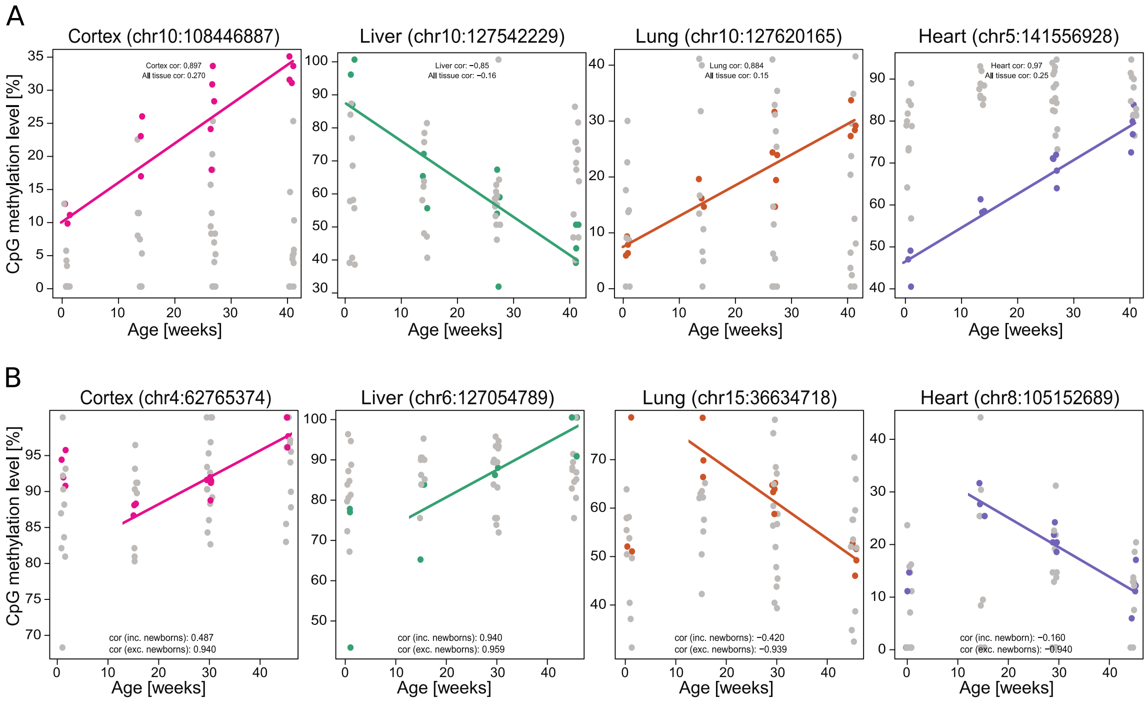


Figure 3.7: Example tissue dependent correlations: (A) Examples of tissue-specific Spearman correlations with age in cortex, liver, lung and heart. Percentage CpG methylation level in all samples is shown, with those of the relevant tissue highlighted. Associated Spearman correlations are shown for both all tissues and for the relevant tissue. Jitter is purely for aesthetic purposes. (B) shows the same for tissue-specific correlations identified when newborn samples are excluded from the analysis. Spearman correlations when newborn samples are included or excluded are detailed.

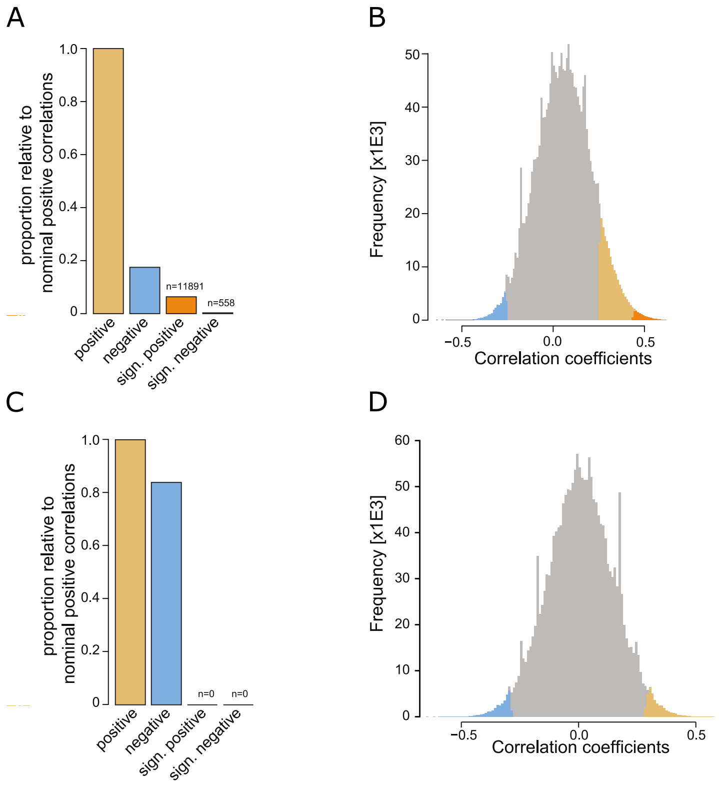


Figure 3.8: Summarised tissue−independent correlations: Summary of Spearman correlations calculated across all tissues. The bar plot (A) depicts proportionate numbers of correlations and the histogram (B) depicts distributions of correlation estimates. Nominal correlations (p value<0.05) are coloured light orange (positive) and light blue (negative), whilst significant correlations (q-value<0.05) are coloured orange (positive) and blue (negative). N numbers above the bar plot signify the number of CG sites within a particular category. (C) and (D) show the same for Spearman correlations calculated excluding newborn samples from the analysis. There are no significant correlations when newborn samples are excluded.

change in correlation directionality between postnatal-development (<1 to 14 weeks of age) and adulthood (14 to 41 weeks of age). This correlation dichotomy, found both tissue-dependently and independently, intriguingly fits well with the theory of antagonistic pleiotropy (Williams, 2001), whereby misregulation of integral developmental processes with age are thought to be responsible for ageing.

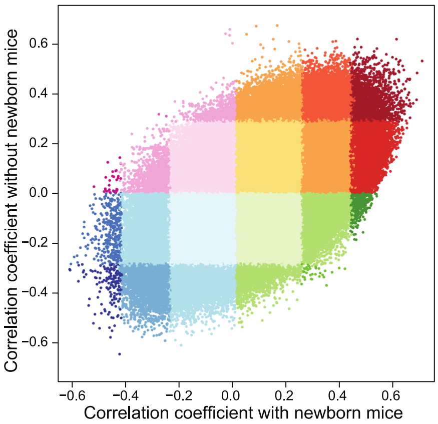


Figure 3.9: Comparison of tissue−independent correlations with and without newborns: Scatterplot of conservation of Spearman correlation estimates when newborn samples are included or excluded from analysis. Correlations that are either positive or negative in both cases are highlighted red or blue respectively. Correlations that differ in directionality with newborn-inclusive correlations being positive are coloured green, and with newborn-inclusive correlations negative pink. Significance of correlation is signified by shading.

To understand whether there was any simple explanation for the significantly positively or negatively correlated tissue-independent sites in the genome (newborns included), I assessed whether they were enriched for any genomic element specifically. From the analysis shown in Figure 3.10A, I found that both positively and negatively correlated sites were enriched and depleted for relatively similar features, suggesting that it is not the features themselves that are determining the directionality of the changes with age. In particular, I found that CpG Islands and CpG Island-rich promoters were depleted for age-associated sites, whereas CpG Island shores and CpG Island shelves were strongly enriched for age-associated sites. To ensure that this test took into account the background distribution of potential sites that could have been significantly associated with age (owing to the inherent feature biases of RRBS), I performed a binomial test to determine which enrichments were significant (multiple testing corrected p-value <0.05). CpG Island shores were defined as the surrounding 0-2 kb upstream and downstream of CpG Islands, and CpG Island shelves were defined as regions 2-4 kb upstream and downstream of CpG Islands. There was also minor enrichment for other genomic elements, such as introns. This analysis suggests that tightly controlled regulatory regions of the genome, such as CpG Islands, are less prone to ageing-associated changes in DNA methylation, suggesting that they are better maintained than the rest of the genome. In addition, the depletion of CpG Islands could have suggested that there was a CpG density-based association to the changes that I was seeing. To test the above hypothesis, I calculated the inverse of the CpG density, or CpG scarcity for the significantly age-associated sites and compared these to what is seen from a background (random) distribution of sites (Materials and Methods section 2.3.9; Figure 3.10B). A two-tailed t-test with Bonferroni correction was performed to assess the significance of any CpG scarcity difference in these sites. Using a p-value of 0.05 as significance cut-off, I found in this analysis that both CpGs that were positively and negatively associated with age were significantly more likely to be found in regions with higher CpG density (lower CpG scarcity) than would be expected by chance. This is perhaps not surprising, since I was measuring the CpG density of the surrounding regions and I had already seen from the previous analysis that these sites were enriched in CpG Island shores and shelves, i.e. regions that would by definition be proximal to regions of high CpG density. However, it should be noted that although these sites are significantly associated with higher CpG density, high CpG density is not strongly predictive of a site changing its DNA methylation status with age (AUC = 0.58 and 0.61).

In addition to assessing simple genomic features and CpG density, I also wanted to assess whether the age-associated DNA methylation changes were enriched in regions surrounding genes that were associated with any specific process or ontology. As such, I performed a gene ontology (GO) analysis of the genes closest to the cytosine of interest, with a maximum cut-off of 4 kb away from the transcript its self. Positively and negatively correlated sites were assessed separately. From this analysis, I found a number of GO processes that were highly significantly associated with positively correlated DNA methylation changes (i.e. sites in the genome that are associated with gains in

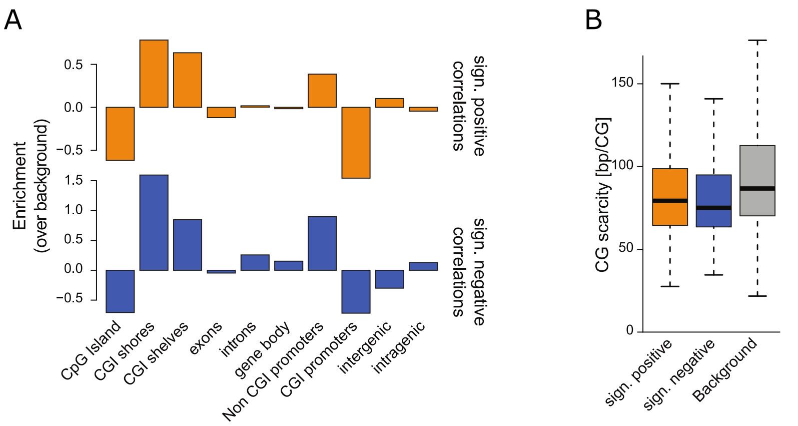


Figure 3.10: Genomic feature enrichment of age-associated sites: (A) Enrichment of both positively and negatively age-correlated CG sites according to genomic feature. This was tested using a binomial test, *p value<0.05. The background used was all CG sites with five-fold coverage in 90% of samples. (B) CpG content 500 bp either side of significantly age-correlated sites. The background was calculated using all CG sites with five-fold coverage in 90% of samples; CpG scarcity is defined as the average distance to a CpG within this 1 kb region. *Bonferroni corrected p value<0.05.

methylation; Figure 3.11A). These GO terms included "anatomical structure morphogenesis", "anatomical structure development" and "developmental process" (Figure 8). These GO terms fit with the concept that sites that gain methylation are associated with potential PRC2-associated genes, and with the idea that an attempted restriction of developmental pathways is at least partially driving the ageing process. In the case of genes neighbouring negatively correlated cytosines, I found a significant enrichment for processes including nucleotide and enzyme binding, which is suggestive of potentially more metabolically associated changes (Figure 3.11A). These findings suggest that the age-associated changes that I observe could alter important biological processes.

In order to determine whether the age-associated changes that I observe for the tissue-independent newborn analysis are truly tissue-independent, I intersected these sites with the tissue-dependent correlation analysis to determine whether the sites were significantly correlated in all tissues independently, and to ensure that there wasn't a single tissue driving any specific correlation. What I found (Figure 3.11B) is that almost all of the sites are shared amongst the four tissues with a subset of sites being only shared amongst three of the tissues. The reason for this lack of overlap of all sites could be due to a lack of statistical power owing to a reduced depth of the sites or a reduced sample number. Overall these results suggest that these ageing associated sites that have been identified are truly tissue independent.

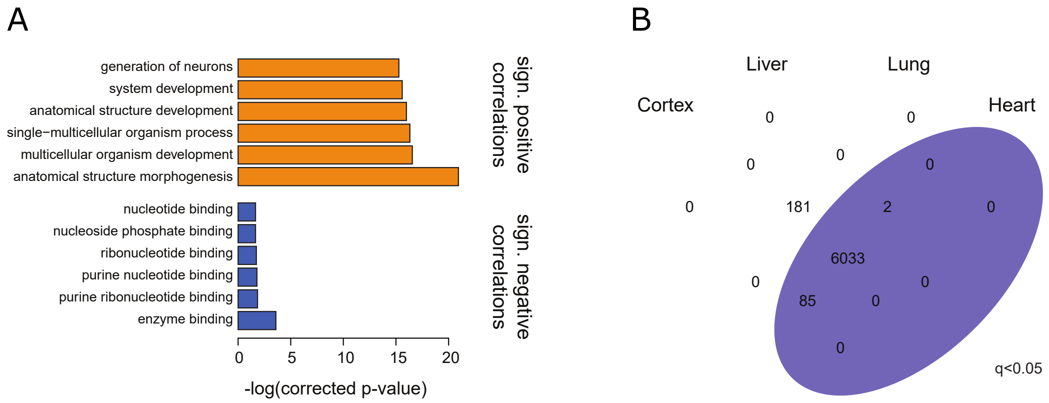


Figure 3.11: Tissue−independent age−associations assessed: (A) The six most significant Gene Ontology (GO) terms for significantly age-correlated CG sites. Terms are plotted against -log(corrected p value), with positive correlations in orange and negative in blue. (B) The overlap between tissues of tissue-independent age-associated Spearman correlations with a corrected p-value (q-value) of <0.05.

In addition, I conducted a similar correlation conservation analysis for the tissue-independent correlations conducted without the newborn samples. However, owing to the lack of statistical power, this analysis was performed with highly nominally significant correlations (p-value <0.005) and not multiple-testing corrected significant correlations. As such, this analysis could contain false positives. Moreover, owing to the arbitrary nature of the p-value threshold, this analysis could also contain false negatives. From this analysis, I found that contrary to the high levels of conservation seen for the newborn-inclusive correlations there was very little tissue-dependent conservation when newborns were excluded (Figure 3.12). In other words, the majority of tissue-independent correlations are seemingly being derived from single tissues and are not truly tissue-independent. These results have to be taken with some scepticism, owing to the inherent limitations of the dataset. However, they suggest that the age-associated changes that are being called with the newborns included are more reflective of the regulation of a core-developmental programme, whereas the changes that are seen with the newborns excluded more likely reflect tissue-specific processes.

Lastly, I assessed whether correlations that were called in a tissue-dependent fashion were conserved across the different tissues (Figure 3.13). This analysis highlighted that tissue-specific correlations, both with (multiple-testing corrected p-value <0.05; Figure

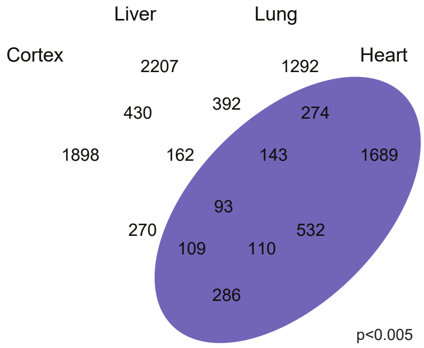


Figure 3.12: Tissue−independent age−associations assessed (newborns excluded): Overlap between tissues of tissue-independent age-associated Spearman correlations with a corrected p-value (q-value) of <0.005.

3.13A) and without the newborns (p-value <0.005; Figure 3.13B) present, were truly tissue-specific changes that are not readily seen in the other tissues.

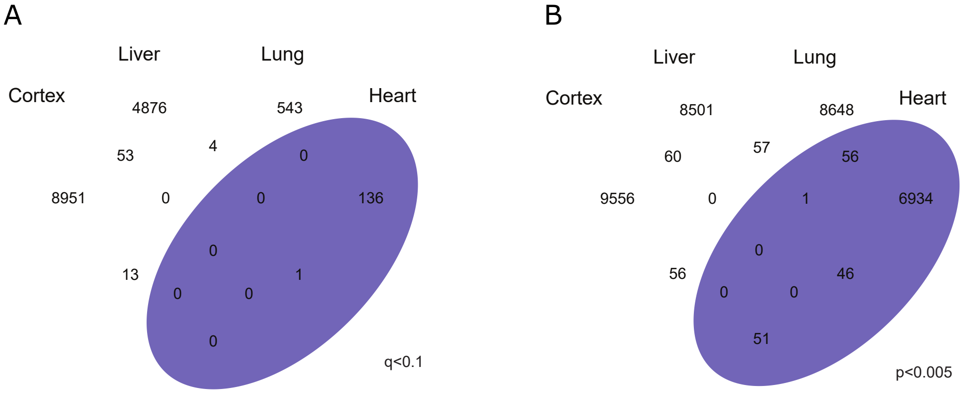


Figure 3.13: Tissue−dependent age−associations assessed: (A)Overlap between tissues of tissue-specific age-associated Spearman correlations, with these defined as correlations with a corrected p-value (q-value) of <0.1 for any of the four tissues. (B) The same representation of tissue-specific correlations with a p-value of <0.005.

This fact is further validated by the GO analysis in Figure 3.14, which has been conducted on the tissue-specific significantly correlated changes (newborns included) and highlights that tissue-specific changes are associated with tissue-specific processes, such as neuronal processes in the cortex. Interestingly, although there is very little overlap between these tissues from the perspective of the cytosines themselves, and although the precise processes are different, again there seems to be a common theme of developmental processes appearing. This is once more suggestive of ageing being associated with regulation of a core developmental programme.

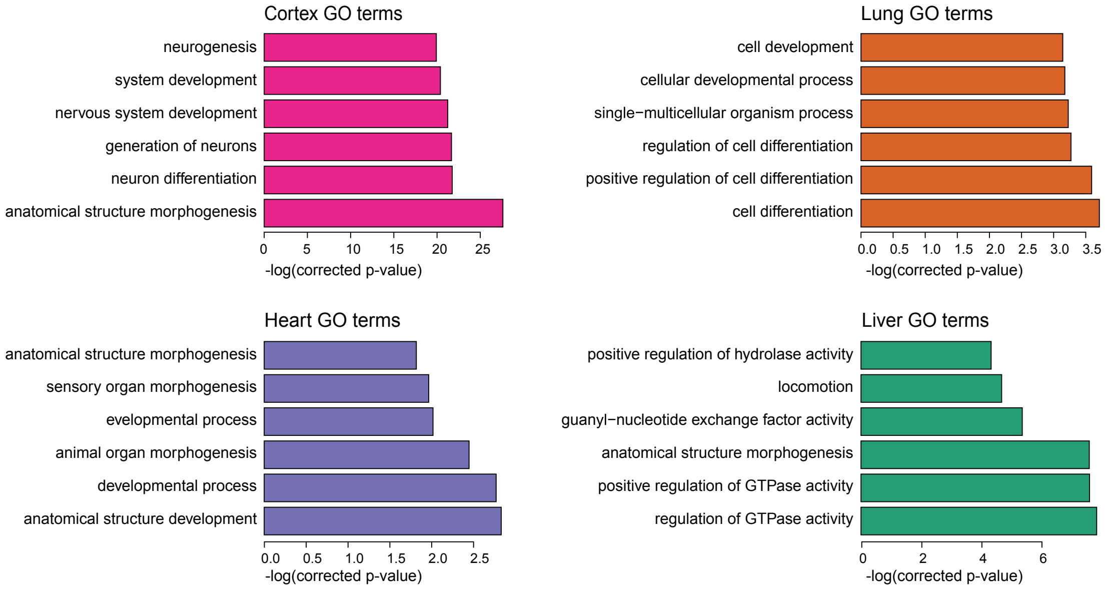


Figure 3.14: Tissue−dependent age−association GO−analysis: The six most significant GO terms for significantly tissue-specific, age-correlated CG sites, plotted for each tissue. Terms are plotted against -log(corrected p value).

3.2.3 DNA methylation levels at a discrete set of CpGs are predictive of age

Having seen that many individual cytosines in a CpG context correlate with age, I set out to define an epigenetic predictor of age in mice. In keeping with the work done in humans, I decided to define my predictor using an elastic-net regression model implemented within the glmnet package in R (Figure 3.15; Friedman et al., 2010). In addition, to our own dataset, I also included a number of publically available RRBS datasets in my analysis. These datasets contained samples from liver, lung, muscle, spleen, and cerebellum and were derived from both male and female C57BL/6 mice, with ages ranging from newborn to 31 weeks of age (Cannon et al., 2014, Reizel et al., 2015, Schillebeeckx et al., 2013 and Zhang et al., 2016a). The rationale behind the inclusion of additional samples was predominantly fourfold. Firstly, more samples would mean that a larger training dataset would be available to define a predictor from, improving the model. Secondly, having samples from mice kept under slightly different conditions, with libraries made using slightly different approaches would mean that the model would be less prone to overfitting, and my model would potentially be more generalizable. Thirdly, additional datasets mean that it would be possible to keep some datasets completely external from the modelling process and so these could be used for an independent validation of the predictor. Lastly, these datasets contained treatments that could be utilised to assess whether the model would be able to measure biological age in addition to chronological age. A more detailed description of the samples can be found in the Materials and Methods section.

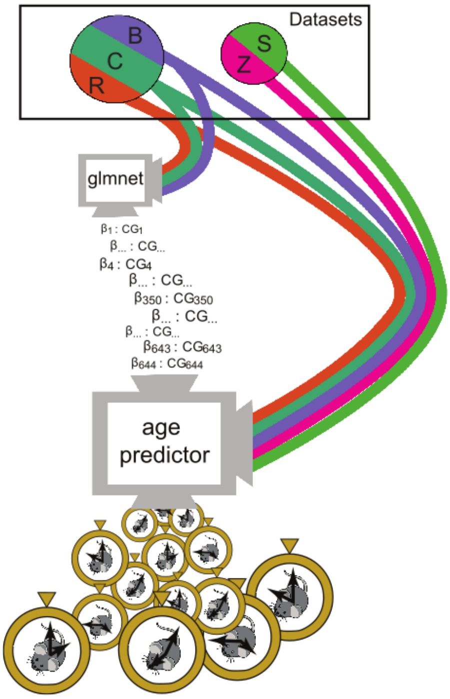


Figure 3.15: Schematic of the modelling procedure: Graphical representation of model definition and testing. Different datasets make up circles and are coloured in agreement with later figures, namely: Reizel (R) in brown, Cannon (C) in green, Babraham (B) in purple, Zhang (Z) in pink and Schillebeeckx (S) in light green. The two datasets not used in training are shown in a different circle and lines represent the processing of DNA methylation data. Training of the model is represented by the screen with caption 'glmnet' and resulting prediction sites and their corresponding weighting are transferred to the 'epigenetic age predictor'. Test data is then fed into the predictor and age is predicted, represented by the pocket watches.

In summary, 129 healthy samples were utilised for the training set and the remaining 189 samples were utilised for the test set, including two datatsets (Zhang and Schillebeeckx datasets), both derived from different experimental settings, that had no samples present in the training set and as such could be used as truly independent test sets, to assess the robustness of the model. All samples were processed as described in Materials and Methods section 2.3. Importantly, all cytosines in CpG context on the sex chromosomes or found in mitochondrial DNA were excluded from the analysis. This was done for a number of reasons. Firstly, to ensure that the model would be sex independent. Secondly, to ensure that it would be compatible with library preparation approaches that may result in loss of the mitochondrial DNA. Thirdly, the mitochondrial DNA was removed because it is commonly poorly converted during bisulfite treatment.

To define the model I decided to utilise sites that were covered by at least 5 reads in all samples. This resulted in there being only ∼18k sites that could be used, which represents about 2% of the total number of sites present in all datasets. One of the reasons that there were so few usable sites was because there was not 100% convergence between the sites of the different datatsets (Figure 3.16).

The reason for this lack of overlap is due to the size selection step in the RRBS protocol being different for the different RRBS datasets. The reason for setting this threshold of at least 5 reads in all samples was to ensure that I could be confident that there were at least six-states that each cytosine in a CpG could be in (i.e. 0, 20, 40, 60, 80 and 100% methylated). In addition, it ensured that there was no need to attempt imputation-based approaches. I had previously attempted both mean and median based imputation, whereby I chose all sites that were represented in 80% of samples with greater than or equal to five reads per sample. However, I found in cross-validation experiments that the results were similar to or worse than starting with fewer sites for which no such imputation was required.

Additionally, I found in the initial cross-validation testing experiments that there was a consistent over prediction at young ages and a consistent under prediction at old ages. This suggested that perhaps the age required a mathematical transformation prior to training. This had previously been done for the multi-tissue predictor defined by Horvath in humans. In humans, the transformation that was applied was a split transformation, as described in the introduction, with a logarithmic transformation from pre-birth to 20 and a linear transformation after 20. As such I decided to test whether a pure logarithmic transformation, a pure linear transformation or a split transformation (similar to the Horvath transformation (Equations 1.1 and 1.2),

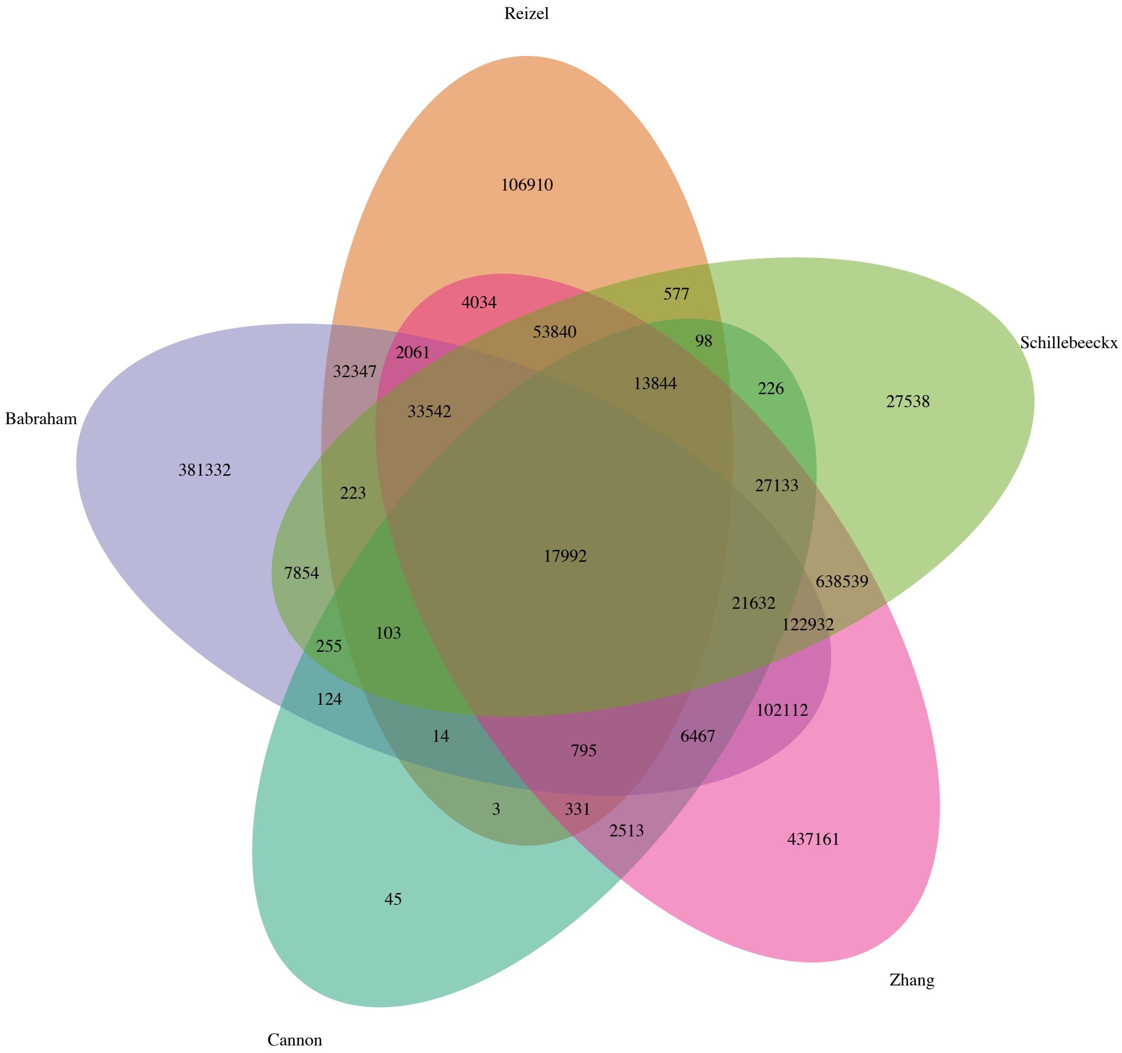


Figure 3.16: Overlap of the datasets used to define the mouse predictor: Venn diagram of the site overlap between the datasets used to define and test the mouse predictor. The dataset is highlighted next to its segment.

but with *adult.age* set to 12 weeks) would be best. I found from these experiments that the pure logarithmic transformation performed slightly better than the log-linear transformation, with both performing better than the linear transformation. As such I decided to define the model using a logarithmic transformation of the data. Perhaps with data spread more evenly across the time course, I would also have found that a log-linear transformation would be similarly optimal but this will only be possible to assess as more time points are assayed.

The model was defined using cross-validation to optimise the model parameters, details can be found in the Materials and Methods section 2.3.12. In addition to the optimisation performed by glmnet for lambda (λ), I performed alpha (α) optimisation as part of the cross-validation. Previously, in the human tissue-specific models (Florath et al., 2013, Hannum et al., 2013 and Weidner et al., 2014) the alpha parameter was defined arbitrarily, whereas in the case of the multi-tissue predictor alpha was set to 0.5 (halfway between Lasso and Ridge; Horvath, 2013). By contrast, I wanted to define alpha empirically using the training dataset. Following this cross-validation, the model I defined contained 329 CpG sites. The model will hereto be referred to as the mouse clock. Coincidentally, the number of sites in the mouse clock is similar to that defined for the human multi-tissue predictor of Horvath (Horvath, 2013).

This mouse clock performed well in the training dataset, as was to be expected, with a correlation of 0.977 between the predicted ages and the actual ages (Figure 3.17A). In addition, in the training set the model had a median absolute error (MAE) of 0.97 weeks. In the test set the model performed worse than in the training set, but was still able to relatively accurately predict age (Figure 3.17B). In the test set the correlation between predicted and actual age was 0.839 and the MAE was larger at 3.33 weeks (Figure 3.18A). This test set error corresponds to a 8.5% measurement error at the oldest age tested (41 weeks). I calculated that my epigenetic predictor was accurate to within a similar margin of error as the human predictors when expressed as a proportion of lifespan. This calculation was made assuming that the expected lifespan for a mouse is (>100 weeks) and that of a human is (>85 years). In addition to defining the error across all samples, I also split the MAE into two groups based upon the ages of the

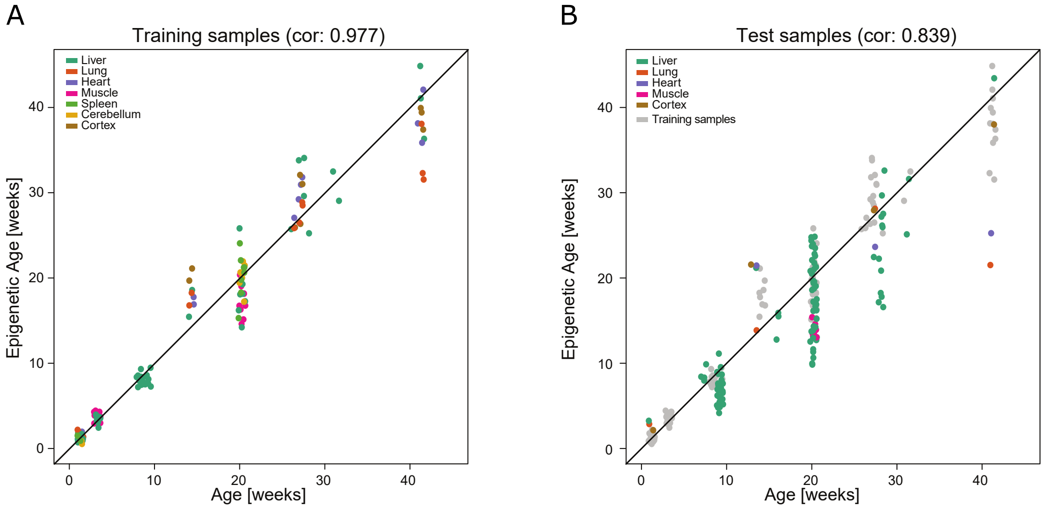


Figure 3.17: Age predictions for training and test data: (A) The ages of samples from the training set, predicted by the model, with actual age in x and predicted age in y. Data points are coloured by tissue and jitter shows experimental error in estimates of age. (B) The ages of samples from the test set, predicted by the model. Training set data is additionally shown.

samples and found that with increasing age there was an increasing error associated with the predictions made by the model (2.14 weeks in the <20 weeks test set, and 4.66 weeks in the >20 weeks text set; Figure 3.18B). Interestingly, this was found both in the training and in the test samples, suggesting that it is a phenomenon not just associated with poor test predictions at increased ages.

Importantly, I find that the samples from the two independent datasets are very well predicted, suggesting that this model isn't simply over-fitting to the datasets themselves (Figure 3.19). This was a legitimate concern in the derivation of this model, because each dataset has very characteristic ages associated with it, and so any dataset-specific difference could manifest itself as a potential confounding ageing signature (for instance the Cannon dataset only has samples at 9 weeks of age).

In addition, I wanted to assess whether there were any additional potential confounders or biases in my model. As such, I wanted to assess whether there was any dataset that was making the model less well calibrated , but saw that this did not appear to be the case (Figure 3.20). In addition, I wanted to see whether there was any relationship between the error in the prediction and the median depth of coverage of the given sample, but found none (Figure 3.20). Lastly, I wanted to assess whether a gender

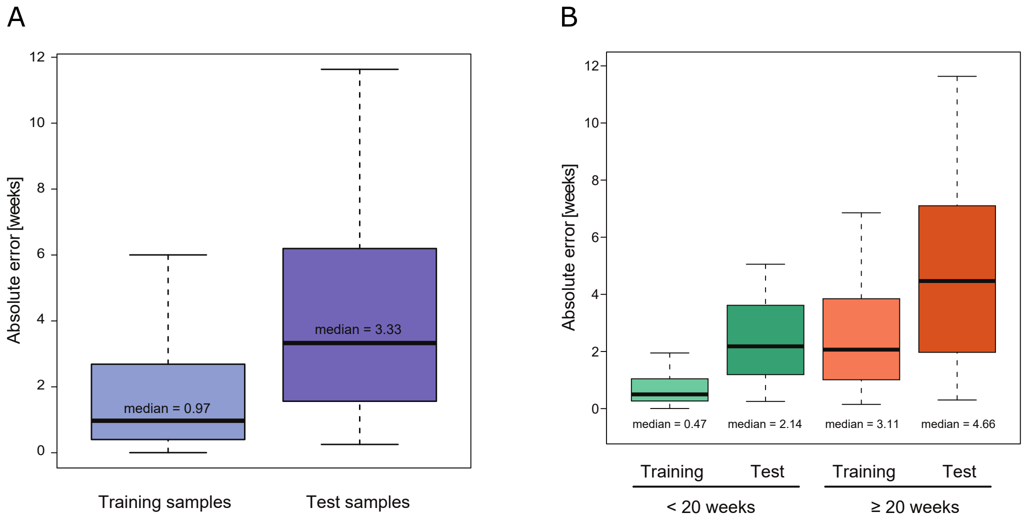


Figure 3.18: Median absolute deviation of age predictions: (A) The absolute error in age prediction for training and test samples, with median error labelled. (B) The absolute error in age prediction for training and test samples by age group (under and over 20 weeks of age), with median error labelled.

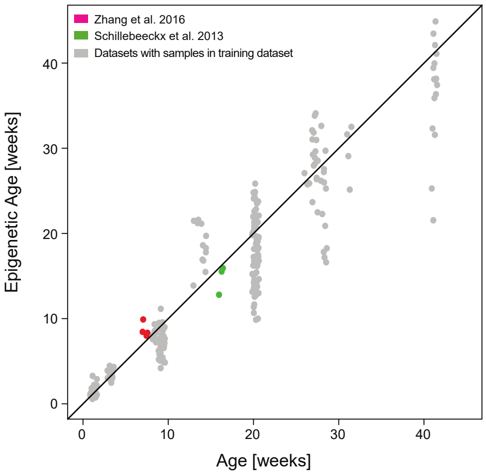


Figure 3.19: Independent validations of age prediction: Age prediction by the model of age of samples from two unobserved datasets (Schillebeeckx et al., 2013 and Zhang et al., 2016a), with training data also shown.

bias could be affecting the accuracy of prediction (Figure 3.20). Again this could have been expected owing to the two genders not being equally represented in the training dataset. However, I found no significant difference between the prediction error for males and females.

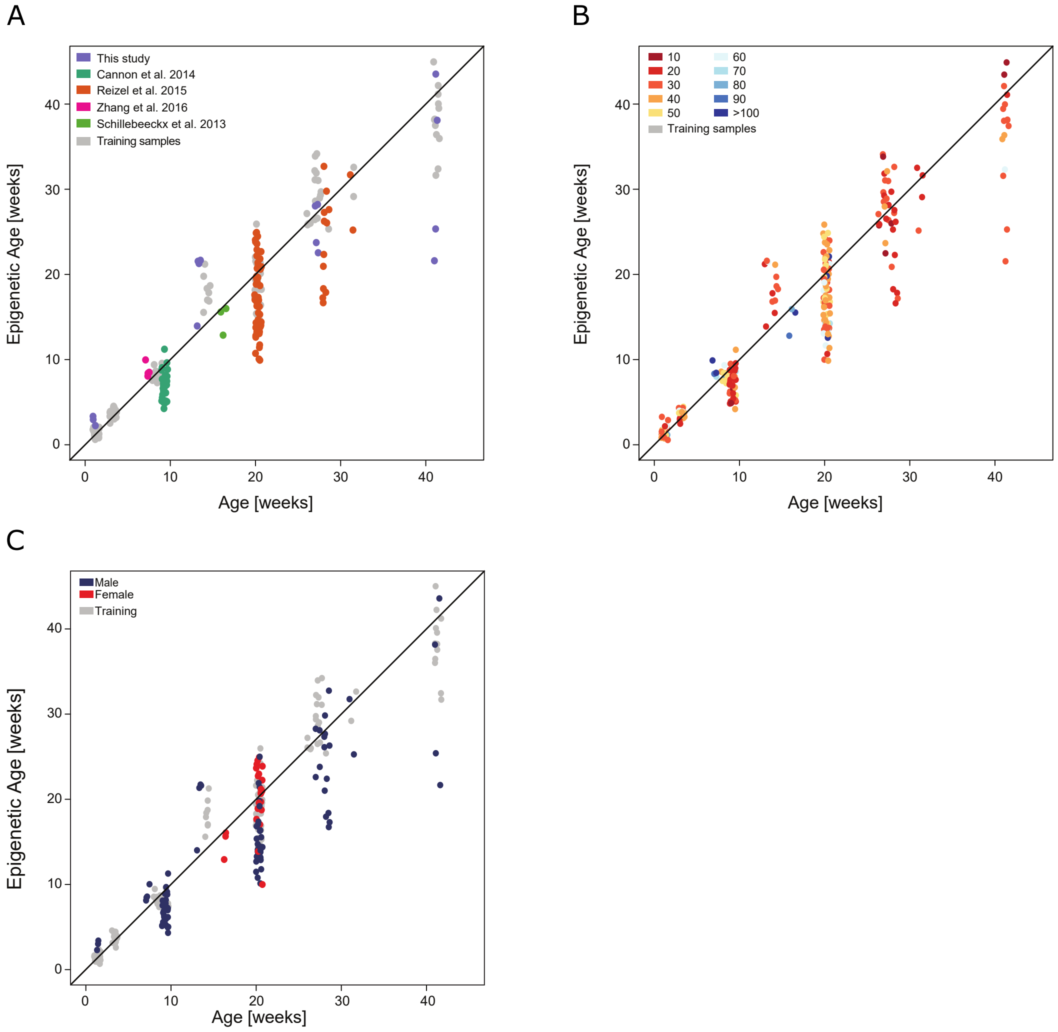


Figure 3.20: Controlling for biases in age predictions: (A) Age prediction of test samples by the model, with samples coloured based on dataset. Training data also shown. (B) Age prediction of test samples by the model, with samples coloured based on average sequencing coverage of prediction sites. (C) Age prediction of test samples by the model, with samples coloured based on sex.

To attempt to understand the nature of the variation within the samples with relation to the 329 sites that define the mouse epigenetic predictor, I performed a principle component analysis (PCA) on the training set. Ninety percent of the observed variability was explained by 69 principal components (PCs), of which 2 PCs (PC1 and PC13) displayed a clear age relation ($p<0.05$; Figure 3.21).

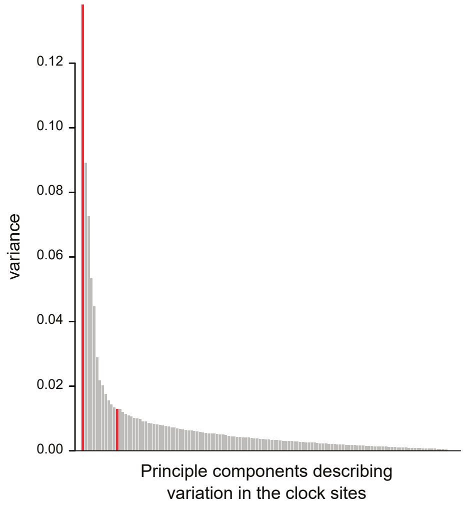


Figure 3.21: Variance explained by age: Percentage explained variation of principal components that segregate samples using the chosen clock sites. 1 and 13 (red) are nominally significant related to age.

Shown in Figure 3.22 are PC1 and PC2. PC1 captures age-dependent variation, whereas PC2 is able to differentiate well between liver samples, which made up a vast proportion of the training set and the other tissues. This analysis of variation at these sites highlights that the major variation within the mouse clock sites in the training set is due to a number of factors, including tissue and age and importantly that factors such as dataset and other technical variables such as bisulfite reagent are not.

An elastic-net regression model is a multiple linear regression model and as such will select sites that are most informative when combined, whilst allowing some redundancy and ensuring robustness. One implication of this is that the clock sites do not necessarily represent the most strongly age-associated sites, but instead reflect those that will be the most informative when combined. As such I wanted to assess what the relationship between age-association and weight of the clock sites was (Figure 3.23).

This figure shows that although there are a number of clock sites that are highly correlated with age and have a proportionately large weight associated with them, the majority of the sites are not strongly correlated with age. In addition, it is possible to see that a number of sites are negatively correlated with age, but have a positive

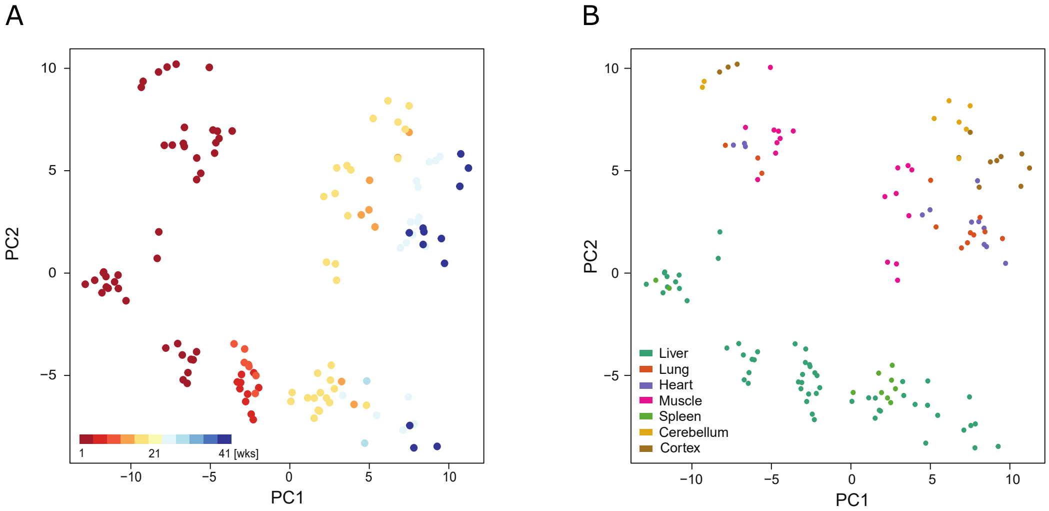


Figure 3.22: PCA visualisation of predictive sites: (A) PCA of training samples using clock prediction sites, with samples coloured by age (weeks). (B) PCA of training samples using clock prediction sites, with samples coloured by tissue.

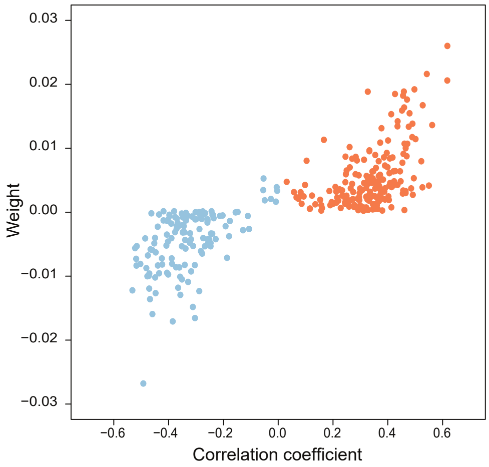


Figure 3.23: Predictive weighting is correlated with age-association: Scatter plot illustrating the weights of clock sites against their age-associated correlation, with negatively correlated sites in blue and positively correlated sites in orange.

weight. One interpretation of this could be that such sites are reflective of the sites that exhibit opposing behaviour during development and in adulthood. In addition, I find that there are slightly more positively weighted clock sites than negatively weighted clock sites and that the weightings themselves are relatively uniform. This suggests that there is no single site that is vastly more important than the rest, which hints at this model describing a more systemic, genome wide effect (Figure 3.23, Figure 3.24 and Figure 3.25A).

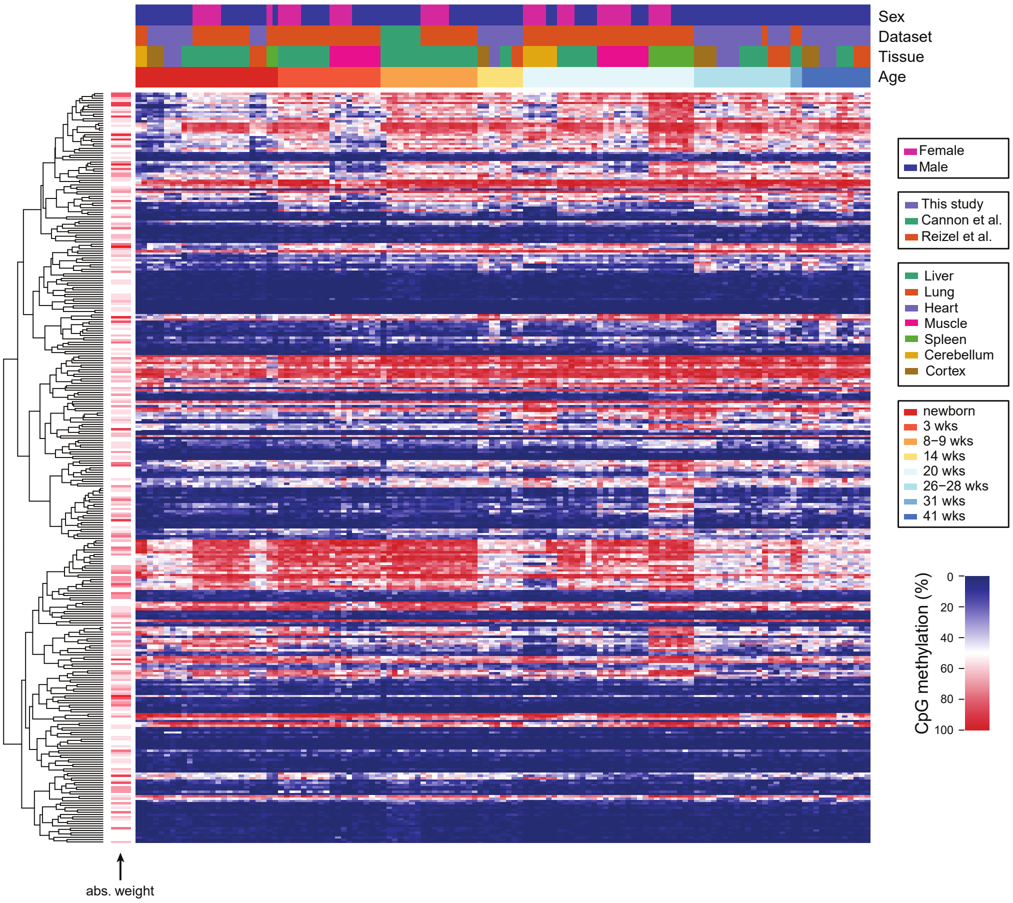


Figure 3.24: Methylation levels of predictive sites across the training set: Heatmap showing the percentage methylation of clock sites. Specified by column side bars are: sex (female pink; male blue), dataset (Babraham purple; Cannon green; Reizel orange), tissue (liver green; lung orange; heart purple; muscle pink; spleen yellow-green; cerebellum mustard; cortex brown), and age (a spectrum from red (<1 week of age) to blue (41 weeks of age)). The clock sites are clustered by Euclidean distance and samples are ordered primarily by age but then by tissue, dataset and sex.

Next, I assessed how the methylation at the clock sites behaved across the samples, visualised in the heatmap in Figure 3.24. I find that there are tissue- and dataset-specific changes in the levels of methylation at any given clock site. However, this does not appear to be the case when gender is considered. In addition, it is possible to see from the heatmap that the changes with age are mainly the result of gradual gains or losses in methylation. This suggests that the ageing process that is depicted by this epigenetic model is gradual and not the result of any sudden changes in methylation at specific sites. This is in keeping with the hypothesis that these changes are the "cumulative effect of an epigenetic maintenance system" as defined by Horvath (Horvath, 2013).

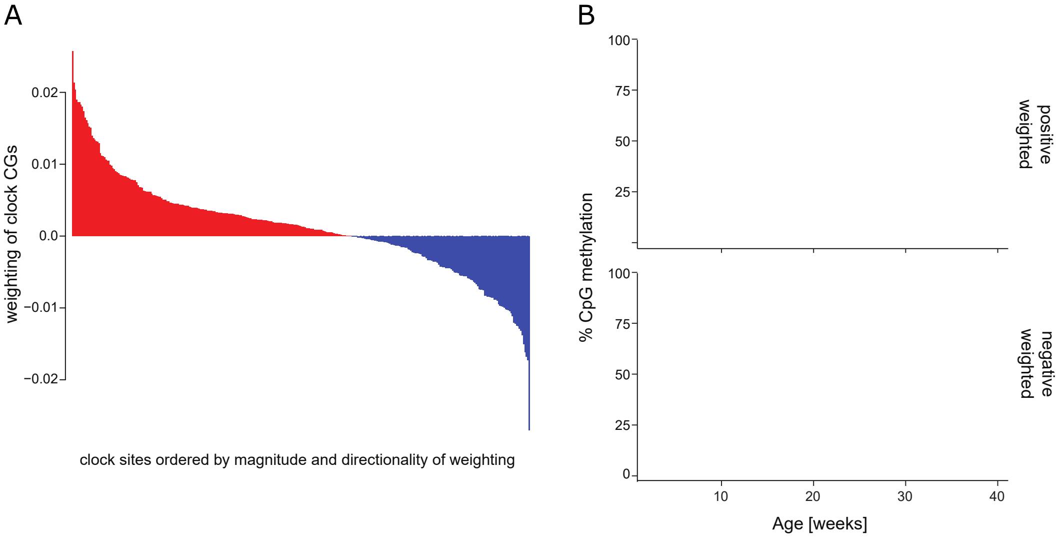


Figure 3.25: Weighting and methylation levels of predictive sites: (A) Barplot illustrating the weighting and directionality of clock sites, with sites ordered from highest to lowest based on this measure. Positively weighted sites are coloured red and negatively weighted blue. (B) A regression plot illustrating overall directionality of methylation changes at clock sites. Positively weighted sites are coloured red and negatively weighted blue. The initial width of each regression line is based on the standard deviation for newborn samples and the final width on the standard deviation for samples of 41 weeks of age.

I also wanted to assess whether, similar to what Hannum saw, the ageing related changes that I see in the clock sites are a result of a tendency towards increased entropy (Figure 3.25). I found that clock sites that were highly methylated on average lost methylation and that clock sites that were lowly methylated on average gained methylation. In other words, the starting level of methylation in the newborn is strongly predictive of the direction that the change in methylation will occur in. This is consistent with the hypothesis that ageing related changes are a result of increased entropy in the system, i.e. a tendency to become 50% methylated.

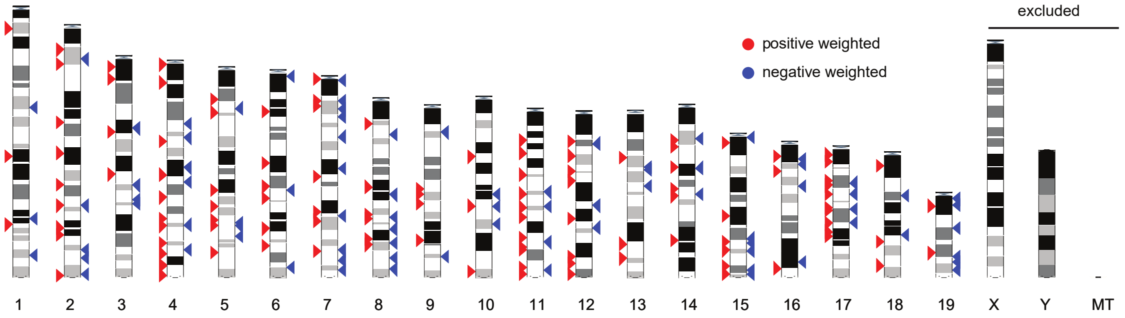


Figure 3.26: Ideogram of the location of the predictive sites: Ideogram illustrating clock site genomic locations, with positively weighted sites indicated in red and negatively weighted sites in blue.

Having assessed the weights of the clock sites, I next wanted to understand whether I could infer any biological mechanism from their location in the genome. To get a first look at the location of the sites in the genome, I generated an ideogram of the clock sites (Figure 3.26). From this it was not possible to identify any specific isochore, chromosome or other large genomic feature that the clock sites seemed to be associated with. In addition, there did not seem to be a difference, in terms of distribution of the sites, between the positively and negatively weighted sites (Figure 3.26). Next I decided to assess whether the clock sites were enriched in any broad genomic annotation, such as CpG Islands (Figure 3.27A). I found that similar to the significantly age-associated sites there was a significant enrichment of the sites in CpG Island shores and non-CpG Island promoters (Binomial test, corrected p-vaue < 0.05). Similarly, I observed a depletion in CpG Islands for both positively and negatively weighted sites similar to age-associated sites (Binomial test, corrected p-vaue < 0.05). However, I see that although there is a significant enrichment of the sites in CpG Island shelves for negatively weighted sites, there is no detectable enrichment in either direction for the positively weighted sites. Interestingly, I observe a significant enrichment for both positively and negatively weighted sites in intergenic regions (Binomial test, corrected p-vaue < 0.05). This is in contrast to the age-associated sites where negatively correlated sites are actually depleted in intergenic regions, but are mildly enriched in positively correlated sites. Additionally, I wanted to see whether there was a significant enrichment for the sites to fall within regions of low or high CpG scarcity (Figure 3.27). However, I found that there was no difference either between positively weighted sites, negatively weighted sites, or random sets of sites defined from the background available set of filtered RRBS sites utilised in model derivation. Lastly, I wanted to assess whether there were any specific GO terms associated with the clock sites in general or defined separately as positive and negatively weighted sites, but found that the sites were not significantly associated with any specific GO term. One interpretation of this is that the clock sites represent an ensemble of many different biological processes that are occurring. Another interpretation would be that the number of mouse clock sites is so low as to preclude any GO enrichment.

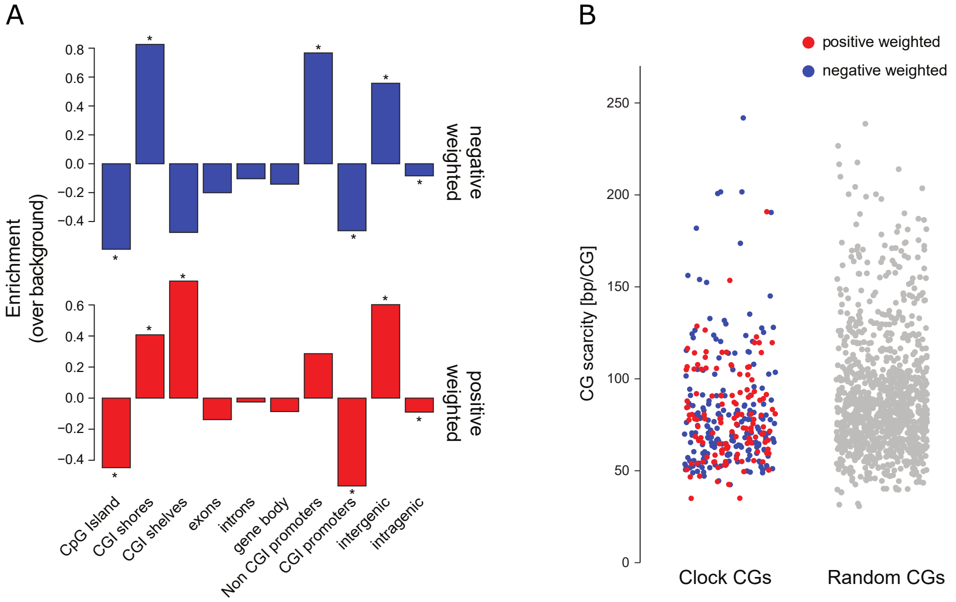


Figure 3.27: Feature enrichment of predictive sites: (A) Enrichment of positively (red) and negatively (blue) weighted clock sites according to genomic feature. Enrichment tested using a binomial test with (*) indicating a p-value of <0.05. (B) CG scarcity of the chosen clock sites compared to a random set of CG sites, with positively weighted sites indicated in red and negatively weighted sites in blue.

3.2.4 DNA methylation age is altered in ovariectomised females and by diet

Following on from the design of the mouse predictor, validation that it can function to predict chronological age and assessment of the clock sites themselves, I next wanted to assess whether I could predict biological age differences. Fortunately, in the publicly available datasets that were available to me, I could assess whether gender, ovariectomy or diet resulted in measurable age differences. From an initial inspection I found strong agreement between the chronological age and epigenetic age of the mice independent off treatment or control, suggesting that any change that I would see would be relatively minor. This is in keeping with what has been seen both for Hannum's cancer validation and for the validations conducted using the Horvath predictor, all of which required large sample sizes to detect these biological age effects.

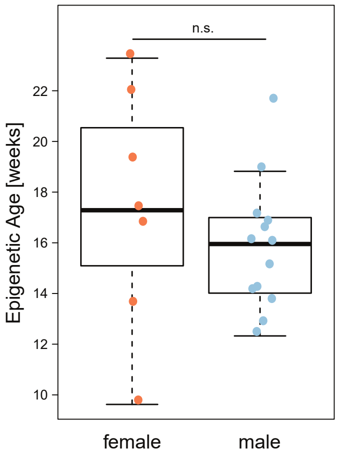


Figure 3.28: There are no gender biases in prediction accuracy: The ages predicted by the model of 20-week-old liver samples (Reizel et al., 2015), with samples segregated by sex. A *t*-test generated a p-value of 0.58.

In the case of gender, I tested whether there was a significant difference between the predictions of age in females and males from the test dataset (Figure 3.28). In the literature, it has been reported that there are gender-specific DNA methylation patterns (Reizel et al., 2015). In addition, it is known that males and females have differing life expectancies, with female mice commonly outliving their male counterparts, although this finding is strain dependent (Austad and Fischer, 2016). In the samples that I tested I found no significant difference between males and females (Figure 3.28). This highlights that the model is able to accurately predict both genders, and suggests that there is no skewing in the predictions, derived from a bias in the training set genders. However, it was noticeable that the female mice had a greater variability in their predictions (Figure 3.28). One explanation for this is that there were very few female samples in the training set and as such the model was better calibrated to male mice.

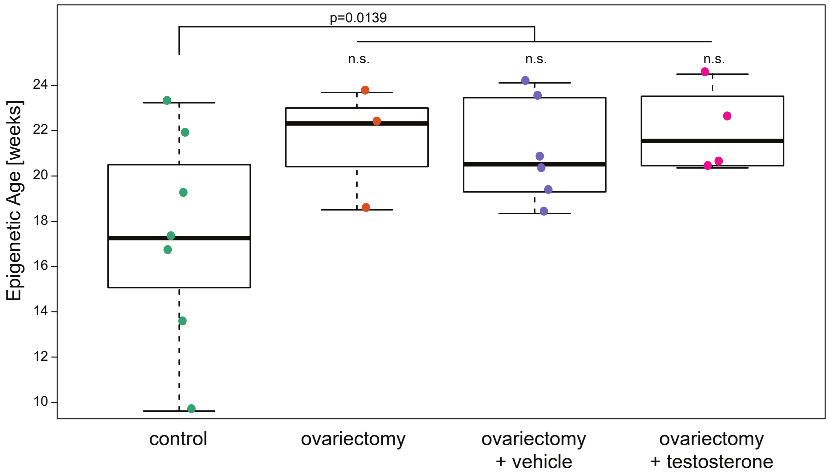


Figure 3.29: Epigenetic age predictions are elevated upon ovariectomy: Age prediction by the model of liver samples from normal females and from female mice that underwent ovariectomy and were administered either vehicle or testosterone (Reizel et al., 2015). An unpaired two-tailed *t*-test was performed to assess the impact of ovariectomy and gave a p-value of 0.014.

From the literature, it has been known for many years that ovariectomy in mice and rats results in a considerable decrease in average lifespan (Asdell et al., 1967). It has also been shown that, in the case of rats, this lifespan phenotype is not reversed when the ovariectomised individuals are treated with testosterone (Asdell et al., 1967). As such I wanted to assess whether the ovariectomised mice in the Reizel dataset appeared epigenetically younger than control mice and in addition whether this difference remained when these mice were treated with either testosterone (5mg/ml) or vehicle (Figure 3.29). I found that ovariectomy resulted in a significant increase in epigenetic age in mice. Additionally, I found that mice that were ovariectomised and then treated with testosterone or vehicle did not have a significantly different epigenetic age from the ovariectomised mice. These results are in keeping with an accelerated epigenetic age (or increased biological age) that is associated with a decreased lifespan. In turn, this result suggests that hormonal differences in mice can result in biological age differences. This is in keeping with what is seen in human breast and endometrial tissue where epigenetic age is known to be poorly calibrated relative to other tissues (Horvath, 2013).

The last treatment that I could assess from the test dataset that I had available was diet. In the literature there is a huge body of evidence that calorie restriction can

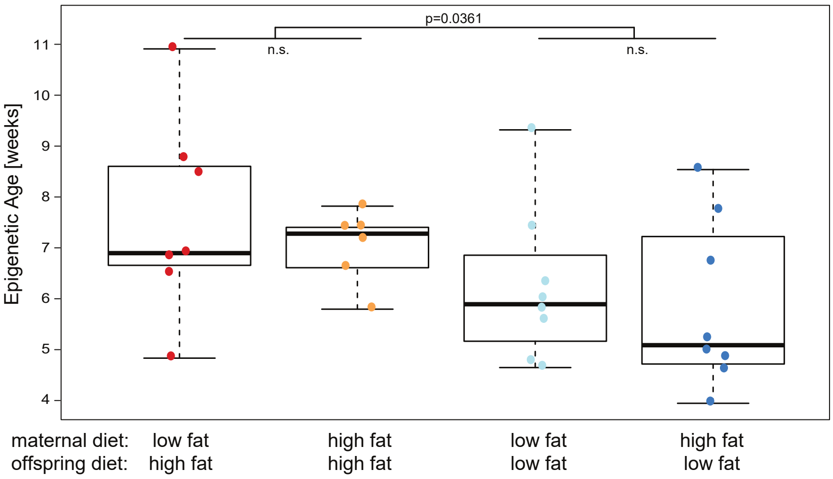


Figure 3.30: Epigenetic age predictions are elevated by high fat diet: Age prediction by the model of samples from a diet alteration study (Cannon et al., 2014). Liver samples were from mice that had undergone the following diet perturbations: maternal high fat diet and subsequently offspring high fat diet, maternal high fat diet and subsequently offspring low fat diet, maternal low fat diet and subsequently offspring high fat diet, and maternal low fat diet and subsequently offspring low fat diet. A two-way ANOVA was performed, with p-values displayed where significant.

alter lifespan, most commonly by increasing it. In the case of high fat diet there is also evidence that it results in an increased likelihood of medical complications and concomitantly a shortened lifespan. In addition, it is known from human studies that dietary alterations to the pregnant mother can result in metabolic phenotypes in the offspring. This has been noted in human studies of natural disasters such as famines, but has also been documented in more controlled experimental settings (Cannon et al., 2014 and O'Rourke, 2014). The study by Cannon *et al.*, characterised the effects of lipid content in the maternal and offspring diet on a range of variables such as physiology, body weight and DNA methylation levels in the liver. In this study they found that there was a significant increase in the likelihood of the offspring suffering from metabolic disease and obesity if the mother was fed a high fat diet (Cannon et al., 2014). In addition, they found that the greatest adverse effect was observed when a maternal low fat diet was combined with an offspring diet. In contrast they found that the least adversely effected were offspring fed a low fat diet following a maternal high fat diet. These results are in keeping with calorie restriction improving health and the notion that a high fat diet is inherently bad for health (Partridge et al., 2005). I assessed whether there was firstly a significant difference between the ages of the offspring fed on high and low fat diet independent of the diet that the mother was fed. I found that indeed there was a significant epigenetic increase in the offspring that were fed a high fat diet, i.e. an accelerated epigenetic age or increased biological age (Figure 3.30). However, although it appears as though there is a minor tendency for the maternal low fat diet to perform worse than the maternal high fat diet my results were not significant (Figure 3.30). This suggests that diet is able to effect epigenetic age and that potentially even maternal diet can also have an impact, though more samples would be need to test this hypothesis.

These results are of particular interest because both of these studies are performed on mice that are far younger than their median lifespan. This is particularly true in the case of the diet study where the mice were only 9 weeks of age and suggests that this epigenetic predictor could be incredibly useful for speeding up the iteration time between ageing experiments conducted in the mouse. Although validation of key experiments would clearly still be required.

3.2.5 Age predictions using the human clock sites in mouse

Having seen that it is possible to define an epigenetic predictor for age in the mouse with a similar error to that of the human predictors (as a proportion of lifespan), I next wanted to assess how well conserved, if at all, the human predictor is in mouse.

This comparison was made more difficult owing to the difficulty in lifting over single base coordinates in the genome into mouse. As such, 1kb regions surrounding each of the 353 Horvath epigenetic predictor sites were defined, to aid the liftover. Three hundred and twenty-eight of the 353 regions that I defined could be lifted over to the mouse genome (release mm10). Of these 328 regions, I found that only 175 overlapped with regions that were covered within our RRBS dataset, when only the Babraham RRBS data was considered. The reason I chose to conduct the analysis solely on the Babraham samples was to maximise the number of regions that would overlap. These 175 regions will hereto be referred to as the Horvath clock regions in mouse. It is important to note that just because these 175 regions contained at least one cytosine in a CpG context, they did not necessarily contain the exact cytosine that existed in the human genome. Additionally, it should be noted that often there was more than one cytosine in a CpG context within each of these regions that were covered. In these instances, the methylation level of the region was defined as the sum of the methylated counts divided by the total number of counts over the region. This could lead to biases in depth between the different regions, and may also result in averaging biases due to the different cytosines being covered by different fragments within the RRBS dataset, something that I avoided by using single positions in the mouse predictor. Knowing these caveats, I assessed how well correlated the human clock regions in mouse were compared to a random set of matched regions (Figure 3.31A). From this analysis, I found that the absolute mean correlation of the Horvath clock regions in mouse were weakly correlated with age, and were not significantly more correlated with age than the random region comparison, although they were shifted to the more correlated side of the distribution. This was not unexpected, since the human clock sites were also weakly correlated with age in humans, owing to the nature of the elastic-net regression method.

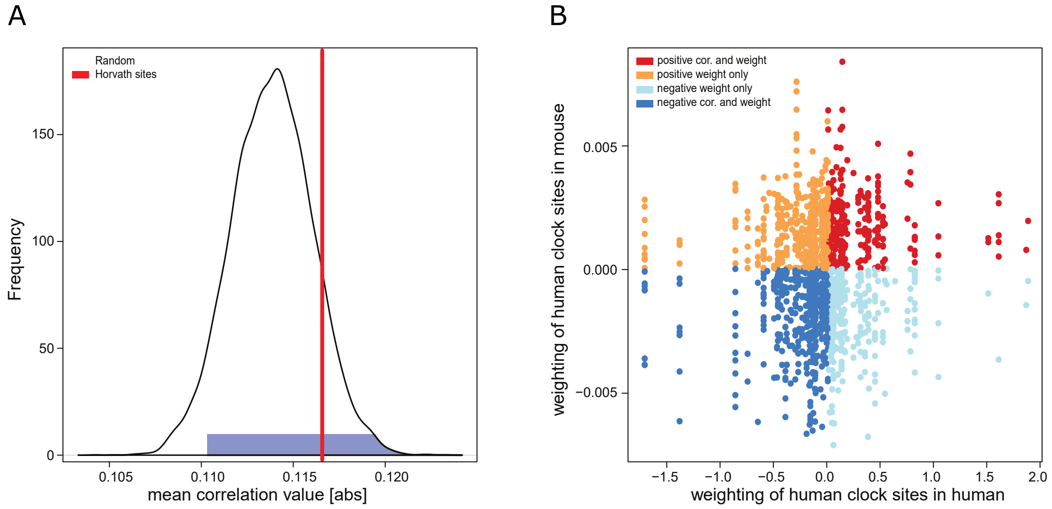


Figure 3.31: Conservation of the Horvath predictive sites in mouse: (A) The mean absolute age correlation of the 175 Horvath clock regions in mouse (red line) and the average absolute age distribution of 1000 random sets of 175 regions (blue). (B) A comparison of the weightings of the predictive sites in the Horvath human clock (Horvath, 2013) with the weightings of the corresponding regions in the epigenetic predictor built using the Horvath clock regions in mouse.

Having assessed the association of these regions with age, I next wanted to assess whether the Horvath sites, when used to predict age in mice, exhibited similar behaviour within the model, i.e. had weights that had the same directionality as those seen in humans. As such I defined a ridge model of the Horvath clock regions in the mouse. From this analysis, I found that the sites did not have a similar directionality in the two organisms (Figure 3.31B). This suggests that the behaviour of these sites are potentially different in these two organisms. However, it again should be noted that these results are comparing the regional weights in one organism against single site weights in another organism and as such no strong conclusions can be drawn.

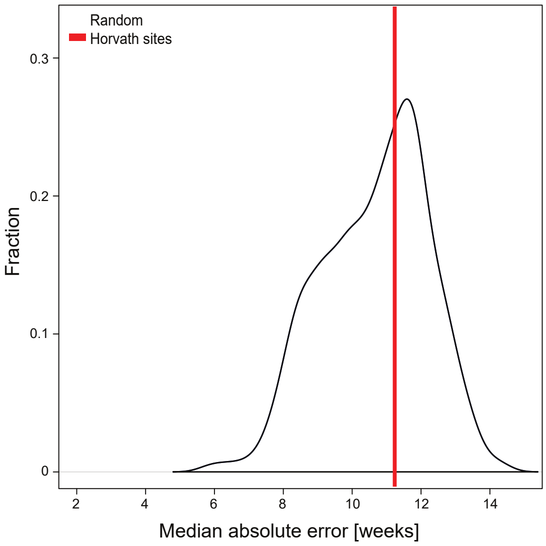


Figure 3.32: The whole genome is predictive of age: Prediction evaluation of the Horvath human prediction sites in mouse. The red line illustrates the MAE of the age prediction model built using the human clock sites (Horvath, 2013). The distribution illustrates the MAEs of 1000 age prediction models each generated using 329 random regions defined in the mouse genome.

This epigenetic predictor, defined from the Horvath clock regions in mouse, was able to predict age with an MAE of 11.2 weeks (from cross-validation), indicating that these sites are in some way predictive of age (Figure 3.32). Unfortunately, it was not possible to directly compare my epigenetic predictor of age in mice with this Horvath predictor, owing to a number factors. One such factor is that my predictor was based on a training set that included samples from a number of publicly available datasets. Another factor was the fact that my predictor was defined from single cytosines in CpGs and not regions containing multiple cytosines averaged. As such to get an impression of the predictive power of the Horvath predictor in comparison to other potential models, I defined random sets of regions, each containing 175 regions, that were matched, as much as possible, for regional size and number of cytosines contained within each region. Ridge models were then defined for each random set of regions and these were compared to that obtained from the Horvath predictor (Figure 3.32). Predictions that were made in cross-validation were then used as the test set to assess predictive accuracy (MAE). This analysis showed that the Horvath predictor was not significantly more predictive of chronological age in the Babraham dataset than the predictors defined from random sets of regions. In fact, the average MAE from the predictors defined from random sets of regions was 10.6 weeks, slightly better than the Horvath predictor. This was interesting because it highlighted not that the Horvath predictor was not predictive of age in mouse, but, more curiously, that the whole genome its self appears to be predictive of age to a greater or lesser extent.

In summary, these results assessing the predictive power of the Horvath clock regions in mouse, have highlighted that there are potential differences between the regions of the genome that are most predictive of age in the two species. This result could speculatively suggest that the specifics of the two processes are different, which is not so surprising owing to the different time scales that they are occurring over. More importantly though, they suggest the intriguing possibility that DNA methylation genome wide is changing with age, and as such all regions of the genome are to some extent predictive of age. The exact nature of this predictive power at any given cytosine in a CpG is potentially reflective of context-specific properties and will potentially provide insights into the mechanism behind these age-related changes.

3.2.6 Attempts to include WGBS samples

I wanted to assess whether my predictor would work in WGBS samples as well as RRBS samples. This was for a number of reasons. Firstly I wanted to see how generalizable my model was to different experimental protocols for measuring DNA methylation levels. Secondly and perhaps more importantly, I also wanted to assess the reliability of my model in mice of older ages, for which there was only WGBS data available

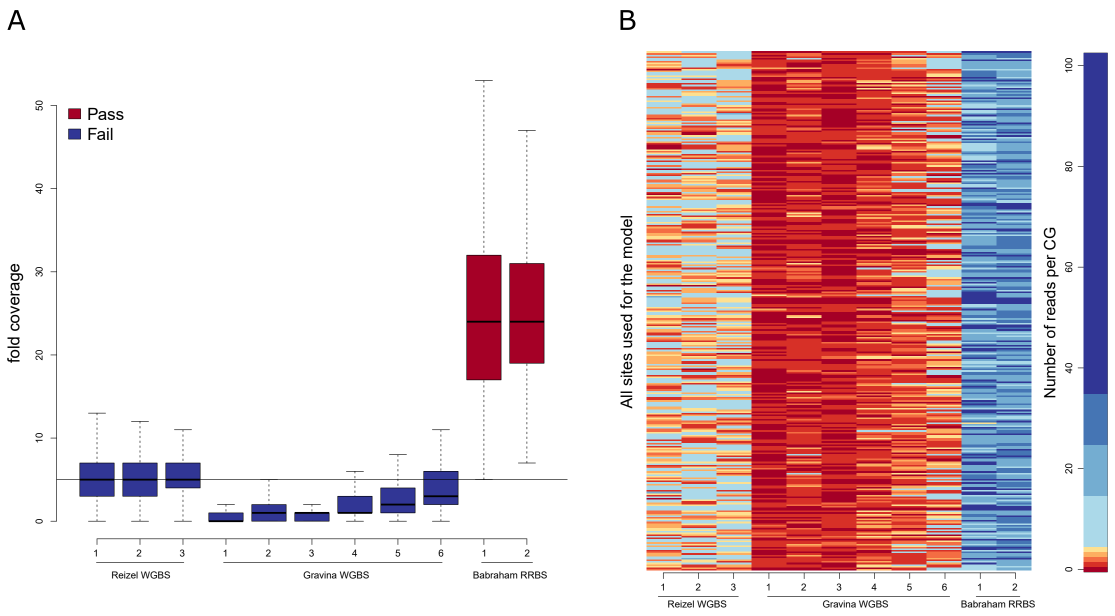


Figure 3.33: QC of WGBS datasets: (A) Boxplot of fold coverage of sites required by the model. (B) Heatmap of fold coverage of the sites required by the model. Shown are publically available WGBS datasets from Gravina et al., 2016 and Reizel et al., 2015, for comparison two Babraham RRBS samples are shown.

(Gravina et al., 2016).

As such I attempted to include publicly available WGBS samples from Reizel *et al.* and Gravina *et al.*. In particular, I was interested in the Gravina samples because they contained samples derived from 104 week old animals. Unfortunately, these samples were not of high enough coverage over the predictive sites in my model. In fact even when I allowed for 20% of sites to be of insufficient coverage, still none of the WGBS samples passed. This result is detailed in Figure 3.33 alongside two passing example Babraham RRBS samples for reference. The boxplot in Figure 3.33A, displays the coverage across all sites to highlight the genome wide depth of these samples. Figure 3.33B highlights the depth at the specific CG sites that are required by the model itself. In the future, it will be very interesting to see how my model performs in a WGBS setting, should there be any samples of sufficient coverage, and in an amplicon setting. The fact that I could not include the samples from the Gravina dataset also means that I am still unsure as to how well my model would perform at ages far greater than the 41 weeks that I have data for.

3.3 Discussion

In this work I have defined a very comprehensive set of matched single base resolution methylomes of mice across multiple ages and tissues. Using this dataset, I was able to determine that there are many thousands of cytosines throughout the genome that exhibit changes in DNA methylation that are significantly linearly correlated with age, both positively and negatively. In addition, I could derive some biological insight into the processes that these sites are likely to affect. However, future experiments will be required to validate whether such processes are truly being altered in an age- and methylation-dependent manner. For instance, it will be interesting to determine how well correlated the DNA methylation changes that I see are in relation to changes to the transcriptome.

I then went on to derive a mouse epigenetic predictor of chronological age that utilises information from 329 cytosines in CpG context. I have validated that this clock is accurate and able to determine chronological age in samples from independent datasets. However, I have only been able to validate that my model works up to 41 weeks of age, which is still relatively young in terms of the lifespan of a mouse (∼30% of the median lifespan of the Babraham mice). As such, future work will be incredibly important to further refine this model such that it is accurate across a far greater range of ages. In addition, I have shown that similar to the human epigenetic clocks that have been defined I am also able to determine aspects of biological age from manipulations or interventions, including diet and ovariectomy at very young ages. This is very exciting, because it means that the experimental iteration time for mammalian ageing experiments could be greatly reduced. This will make mechanistic manipulations and blue sky experimentation far cheaper and mean that these sorts of experiments in ageing studies will be far more readily undertaken.

3.3.1 Other epigenetic clocks

At the same time that the work from this chapter was being published in Genome Biology, two additional mouse clocks were defined (Petkovich et al., 2017 and Wang et al., 2017). In addition, to this a Wolf clock was also defined (Thompson et al., 2017). These studies further validate the hypothesis that such a mechanism is highly conserved, at least among placental mammals.

These two mouse clocks were both defined in specific tissues. The Wang *et al.* epigenetic predictor was defined using publicly available liver data and the Petkovich *et al.* predictor was defined in whole blood. These predictors were both defined using elastic-net regression models. However, they differed slightly in the specifics of the way in which sites were chosen for the modelling. For instance, in the coverage thresholds that were set.

The Wang predictor utilised 107 samples to define and test their predictor. The predictor itself was defined from 148 CpG sites in the genome and was accurate to within 4.2 months (∼18 weeks). This is in contrast to my predictor, which was defined from 329 cytosines in CpG context and for which there was an error of 3.33 weeks. However, it should be noted that their error measure includes mice up to 104 weeks of age whereas ours only includes mice up to 41 weeks of age, which could in part explain this discrepancy in recorded measurement error. Another reason for this discrepancy could be due to the difference in size of the training set that was used to define the model and the number of folds of cross-validation, together with the lack of optimisation of the alpha parameter in their model.

In their study, Wang *et al.* assessed whether there was any detectable differences in biological age for the well-studied, long-lived genetic mutant Ames Dwarf mouse, for the life extending intervention of calorie restriction, and for the life extending treatment of rapamycin. They found that all resulted in a decrease in epigenetic age relative to chronological age, consistent with the idea of these animals ageing at a slower rate.

The Petkovich predictor was defined using newly generated RRBS data from 141 male mice. The Petkovich predictor utilises 90 CpG sites in the genome and is highly correlated with chronological age both in the training and test samples. The error in the predictions for the samples assessed are 4 to 14 weeks up to 43 weeks of age and 21 to 30 weeks of age for mice older than 43 weeks of age. This lower bound of error is comparable with my error measurement of 3.33 weeks of age, and suggests that this model is relatively precise.

In this study Petkovich *et al.* assessed whether there were any detectable differences in biological age for two well-studied long-lived genetic mutants: the Snell Dwarf and the growth hormone receptor knockout (GHR-KO) mice. In addition, they assessed biological age upon calorie restriction in two different mouse strains, in fibroblasts and in reprogrammed iPSCs. They found that for the Snell Dwarf mice, the GHR-KO mice, and the calorie restricted mice, independent of strain, there was a reduced epigenetic age relative to the epigenetic ages of the control samples. In addition, they found that iPSCs had a reset epigenetic age when compared to the epigenetic age of the lung and kidney fibroblasts from whence they were derived. This is similar to what is seen in the human setting.

Interestingly, both these studies and ours found that ageing-associated changes tended to result in an increase in entropy of the site under study. Curiously these studies also suggest that this epigenetic clock phenomenon is a genome wide one. Another piece of evidence in support of this is that there is very little overlap between the three models. However, this could also be due to the filtering criteria and the nature of the models being derived in different tissue settings. This is particularly evident from the work conducted by Petkovich *et al.* who defined two independent models that were both strongly predictive of age and our study where I found that even randomly selected sets of regions were predictive of age.

In addition, all found that it was possible to predict biological age in a number of different settings. In contrast to both of these studies, ours was the only study to report accelerated epigenetic age relative to controls. However, these results taken together highlight that it is possible for epigenetic age to both increase and decrease upon treatment, but also to tick slower more generally in the case of the genetic mutants. Importantly, our study was the only one that defined an epigenetic predictor in a tissue independent fashion. This is important because it reduces the likelihood of the predictor being cell composition derived and suggests that similar to the human setting, an epigenetic predictor can be defined solely from cell intrinsic changes. Although it is likely that in the other studies the predictors are also being derived from such cell intrinsic changes too. This is particularly likely to be the case for the Petkovich *et al.* study since they are able to show that they can predict the “expected” epigenetic ages for non-blood samples, namely iPSCs and fibroblasts derived from two separate origins.

Unfortunately, it was not possible to directly compare the predictive power of the various models against one another. This was due to a number of technical reasons. The main constraint being coverage, this is due in part to the sequencing depth achieved in the different datasets but also due to the different library preparation protocols that were performed. In an attempt to counter this I tried performing a number of different imputation-based approaches. These included mean and median imputation, alongside imputation derived from a deep learning approach called DeepCpG (Angermueller et al., 2017). Unfortunately, there were so many sites missing that the predictions that came out were more a reflection of the imputation than of the samples themselves. This was compounded by the other limitation in this comparison, which is that I would want to ideally compare the two other clocks in samples that they did not utilise from public data but that came from the same or similar tissue. This further reduced the available samples that could meet the coverage criteria. In the case of the Wang study another issue was that they derived their clock from samples that I also utilised, as such any prediction that did hold would not be strictly from an independent dataset.

Due to the now increased number of available samples from these studies and from additional public studies, together with newly generated data from ourselves, which now spans ages up to and including 104 weeks (Figure 3.34), I decided to focus on defining a more general epigenetic predictor that encompasses all of these datasets rather than spending time trying to imperfectly compare the different predictors against one another. This model now contains data from almost 1000 samples including RRBS and WGBS, from a multitude of different treatments, strains and ages. With ages ranging from early pre-birth to over 3 years of age. This model is still being derived and will form the basis for future work.

However, from what I have seen from preliminary cross-validation experiments it is

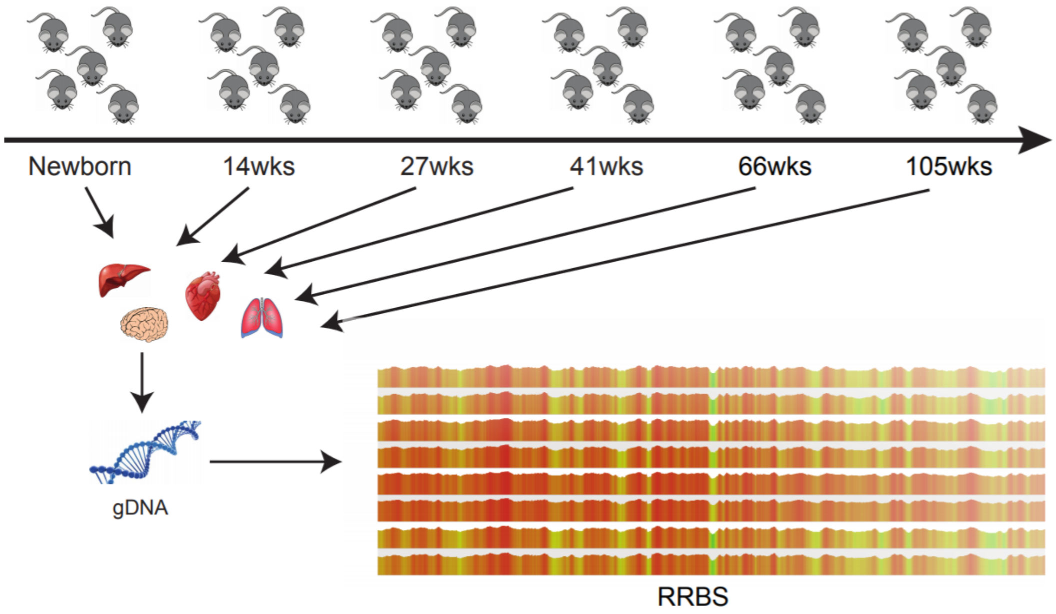


Figure 3.34: Schematic of the Babraham dataset extended: Graphic summary of the Babraham dataset. DNA was isolated from liver, lung, heart and cortex of newborn mice and mice aged 14, 27, 41, 66 and 105 weeks, and reduced-representation bisulfite (RRBS) libraries made from this.

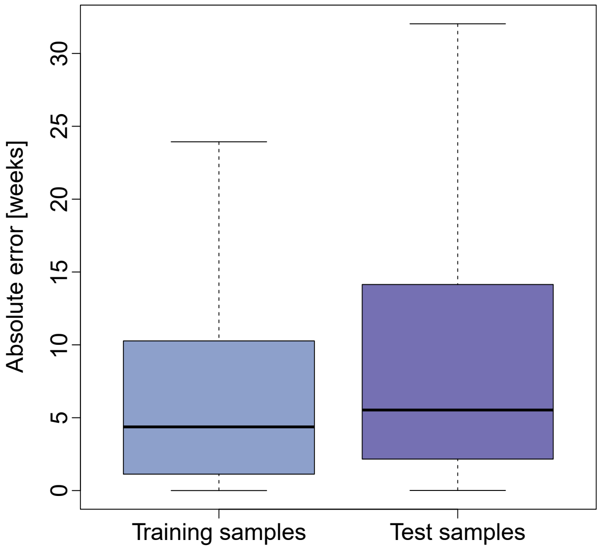


Figure 3.35: Median absolute deviation of age predictions from extended model: The absolute error in age prediction for training and test samples.

clear that the model encompassing all of these datasets performs far better than any of the models individually with an MAE across all ages of 5.33 weeks (Figure 3.35). I hope that future work on this model will validate the findings pertaining to biological age that the previous mouse epigenetic predictors demonstrated. In addition, I hope that the model will provide a solution to the issue of missingness and open the door for use of this predictor in a more general ageing research setting.

Chapter 4

Single cell ageing

4.1 Introduction

Insights into the ageing process that are made at the single cell level will improve our understanding of the true nature of ageing and how it is occurring. Studying ageing at this level of detail will enable us to interrogate hypotheses to examine whether ageing is the result of a passive decay process, or an actively procured state. Amongst other hypotheses, we could ask for instance whether the aged state results from changes in cell type composition or stem cell exhaustion. In addition to enabling us to ask these perhaps more abstract questions regarding ageing, studying the process of ageing at this level of detail in specific cell types and tissues will also help us understand the root causes of age-related pathologies in said cell types and tissues and why they increase with age.

In this present study, I wanted to ask the question of whether I could detect and characterize the process of ageing in a predominantly quiescent stem cell population at the single cell level from epigenetic and transcriptomic information. There were many reasons for why a quiescent stem cell population was chosen. Firstly, in order to simplify the analysis, I wanted to study a population of cells that were diploid and not polyploid, such as cardiomyocytes, which exhibit a range of ploidy that has been shown to be age dependent (Laflamme and Murry, 2011). Secondly, I wanted to study a quiescent population of cells to mitigate as much as possible against any confounding heterogeneity that could be the likely result of cell cycle differences between young and old cells. Thirdly, I wanted a population of cells that were very homogeneous, again, to remove any confounding heterogeneity associated with variation in cell composition. Fourthly, I wanted to study a population of stem cells that represented the greatest differentiation potential within their lineage(s). This is because it is known that many stem cells are able to divide to differing extents in an asymmetric fashion (Conboy et al., 2007, Kuang et al., 2007 and Rocheteau et al., 2012), as such I wanted to determine whether detrimental epigenetic effects have already manifested themselves in this stem cell pool and as such could they be passed on to more differentiated/functional cells upon division. Fifthly, a strong link has been made between ageing and the loss of methylation of histone H3 at lysine K27, (H3K27me3), an epigenetic modification that is important for ensuring correct lineage specification early in development and is maintained in adult stem cells. Sixthly, I wanted to assess a population of cells in an *in vivo* context. In addition, I wanted to study stem cells that had a simple differentiation trajectory, such that future work studying cell types downstream of our chosen stem cells could be assessed comparatively easily. The system that I chose to study that fit all these criteria were the quiescent stem cells of the mouse skeletal muscle system.

In the skeletal muscle system, ageing is accompanied by a decline in muscle mass and strength (Brack and Muñoz-Cánoves, 2016 and Evans and Campbell, 1993) as well as a hampered regenerative capacity (Bernet et al., 2014, Chakkalakal et al., 2012, Cosgrove et al., 2014, Price et al., 2014, Sousa-Victor et al., 2014 and Tierney et al., 2014). This commonly results in physical incapacitation of individuals within the elderly population (Jang et al., 2011 and Renault et al., 2002). Skeletal muscle homeostasis and regeneration are ensured by tissue-specific Pax7-expressing satellite (stem) cells (Lepper et al., 2011, Sambasivan et al., 2011 and von Maltzahn et al., 2013; Figure 4.1). In adult mice these cells are largely quiescent in homeostatic muscles. However, upon muscle injury satellite cells are activated, proliferate and either differentiate to form new muscle fibers or self-renew to reconstitute the stem cell pool (Brack and Rando,

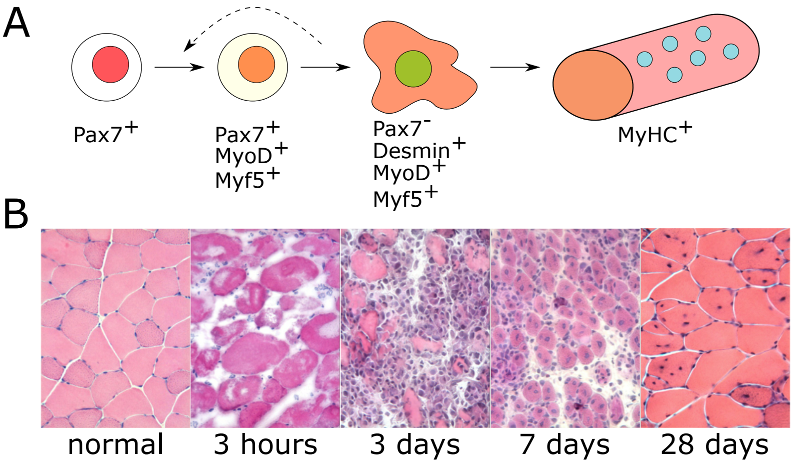


Figure 4.1: Schematic of the muscle lineage: (A) Diagram of the muscle lineage; from Pax7 positive quiescent muscle satellite cells, to activated muscle satellite cells, to Pax7 negative myoblasts and onto fused muscle fibers. (B) Images of mouse TA muscle upon injury. Images are taken at defined time points displayed under each image. Blue staining indicates the position of nuclei within the muscle tissue. Images courtesy of Dr Brendan Evano.

2012, Conboy and Rando, 2002, Olguin and Olwin, 2004 and Zammit et al., 2004). Age is associated with a reduction in the regenerative potential of muscles and has been attributed to impaired intrinsic regenerative potential of the satellite cells (Bernet et al., 2014, Chakkalakal et al., 2012, Cosgrove et al., 2014 and Sousa-Victor et al., 2014). This is thought to be partly due to the influence of extrinsic environmental cues (Conboy et al., 2005) that are known to alter the niche. This is further exacerbated by the decrease of satellite cell numbers in both mouse (Brack et al., 2005, Collins et al., 2007 and Conboy et al., 2005) and humans with age (Renault et al., 2002). Importantly, for this current study, this reduction in stem cell number is far less than that observed in other stem cell systems such as the hematopoietic system.

Exposure of old mouse muscle satellite cells to a youthful environment or growth factors has been shown to partly restore the proliferation and differentiation properties of the stem cells (Brack et al., 2007, Collins et al., 2007 and Conboy et al., 2005). In addition, calorie restriction has been shown to improve the functionality of satellite cell and muscle regeneration of both young and old animals (Cerletti et al., 2012). The plasticity of these changes suggests that the functional decline seen for aged satellite cells could at least be partially a result of changes to their epigenome. Interestingly, DNA methylation is widely increased in old human myoblasts, which notably suppress

Sprouty1 (SPRY1) expression, a crucial factor for ensuring re-entry into quiescence and self-renewal following activation during homeostasis (Bigot et al., 2015).

The muscle satellite cell population is itself phenotypically and functionally heterogenous, both during homeostasis and also upon regeneration (Chakkalakal et al., 2012, Conboy et al., 2007, Kuang et al., 2007 and Rocheteau et al., 2012). Many different methodologies have been used to identify muscle satellite cell subpopulations and to characterize their properties. This has resulted in an understanding that subsets of muscle satellite cells exhibit different proliferation histories and a range of self-renewal, differentiation and regeneration capacities (Zammit et al., 2004). The extent to which this heterogeneity is important/required for tissue homeostasis and regeneration is at present unknown. In addition, it remains unclear how this heterogeneity progresses during ageing. Nevertheless, it has been suggested that the satellite cell population becomes more homogenous with age, with a reduced proportion of cells having a high regeneration potential and increased fraction of cells having low proliferative capacity (Brack and Muñoz-Cánoves, 2016 and Chakkalakal et al., 2012). Therefore, one major unanswered question in the field is whether age-associated decline in satellite cell function is associated with a global functional drift of the population, or with the selection of a few unfit clones. As such, to reduce the functional heterogeneity within our present study, I decided to focus on a subpopulation of functionally homogeneous, quiescent muscle satellite cells from the *tibialis anterior* (TA) skeletal muscle. These cells were defined as being the most highly expressing of all Pax7 positive muscle satellite cells (MuSC) and are thought to undergo asymmetric cell division. These Pax7-high stem cells will hereto be referred to as TA-Hi MuSCs. These cells represent not only a powerful model system but they are also hugely important in the study of ageing owing to the burden of muscle frailty in the aged population.

To characterize how the process of ageing is defined in a predominantly quiescent stem cell population, I created a combined epigenome and transcriptome single cell dataset from young (approximately 3 months of age) and old (more than 24 months of age) TA-Hi MuSCs. This dataset represents the first combined epigenomic and transcriptomic single cell dataset investigating adult cell populations and provides us with an unparalleled resource to address this exciting question.

4.2 Results

4.2.1 Description of the samples collected

TA-Hi MuSCs were collected from male mice containing a GFP reporter under the expression of a Pax7 promoter (Tg:Pax7-nGFP (Sambasivan et al., 2009), on a B6D2F1/JRj background (mixed Black6/DBA2)). Subsequent to euthanisation, the TA muscle was carefully dissected from these mice (Figure 4.2A). Next, the satellite cells were dissociated from the muscle before fluorescence-activated cell sorting (FACS) for GFP (Figure 4.2B). The FACS was performed twice to ensure that the cells being collected were single cells. The TA-Hi MuSCs were defined through gating of the satellite cell population and defined as the top 10% of Pax7 expressing cells (Figure 4.2C). A more detailed description of the isolation of these cells is provided in Materials and Methods section 2.4.2.

The TA-Hi MuSCs were collected into lysis buffer and the single cell combined protocol was performed (Materials and Methods section 2.4.3). Due to the fact that these cells are highly quiescent and thus contained less RNA than cycling cells, 24 cycles of amplification were required to generate a library from the polyA-containing RNA. This is in contrast to the typical 16-18 cycles commonly used, for instance for mESCs. Approximately, 96 cells were collected per mouse.

4.2.2 Quality-control of the transcriptome data

Transcriptome information was assessed from five young and three old individuals. Following alignment of the transcriptome to mm10, gene expression counts were quantified (Materials and Methods sections 2.4.4 and 2.4.4) and the number of transcripts for a given cell were plotted to assess the quality of the libraries.

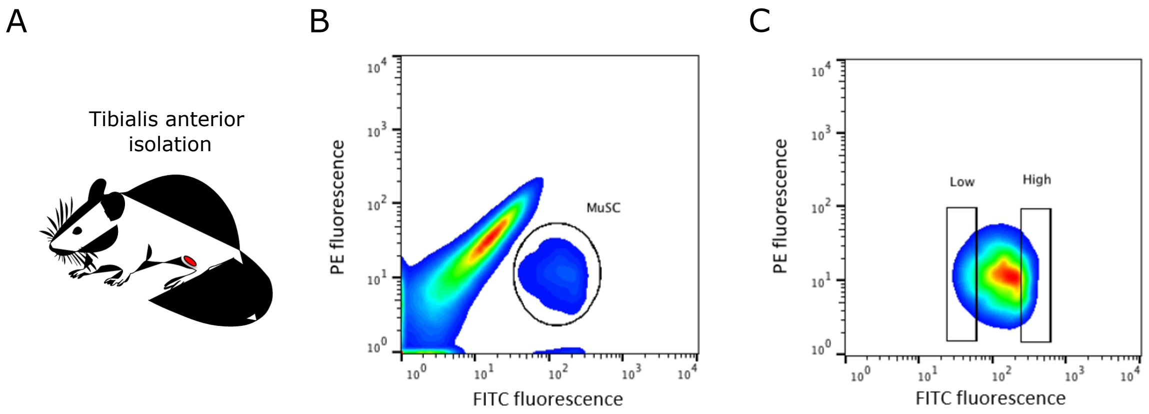


Figure 4.2: Flow sorting of TA Hi MuSCs: (A) Schematic of the location of the TA muscle that was isolated to extract MuSCs. (B) FACS plot of the gating used to sort the TA-Hi MuSCs from the background. (C) FACS plot of the gating used to define the TA-Hi MuSC population from the rest of the TA MuSCs. The channel on the y-axis is staining for Phycoerythrin (PE) and the FITC channel on the x-axis is for GFP. Images courtesy of Dr Brendan Evano.

From this preliminary analysis, I observed that in contrast to previous libraries generated from mESCs, there were, on average, fewer genes expressed for a given cell (>1,000 genes here, compared with >4,000 for mESCs; Figure 4.3). This is likely due to these cells being quiescent and as such requiring little transcription to remain functional. Importantly, the majority of cells were expressing a consistent number of transcripts and the expression profile across the transcriptome was consistent with what would be expected from the distribution of counts. However, it can be seen that there are some cells that express very few transcripts (Figure 4.3). As such, we decided to set an arbitrary quality threshold on the number of genes expressed in a given cell. We set this cut-off at 1,000 genes. The remaining number of cells that passed this quality cut-off and the individuals that were chosen for subsequent analysis are highlighted on an individual basis in Table 4.1.

4.2.3 Assessment of cell cycle from the transcriptome

One of the largest sources of variation in single cell transcriptome analysis is that of cell cycle. As such, although the cells I selected were quiescent, I wanted to ensure that any results I obtained were not a result of variation in proportions of cells in

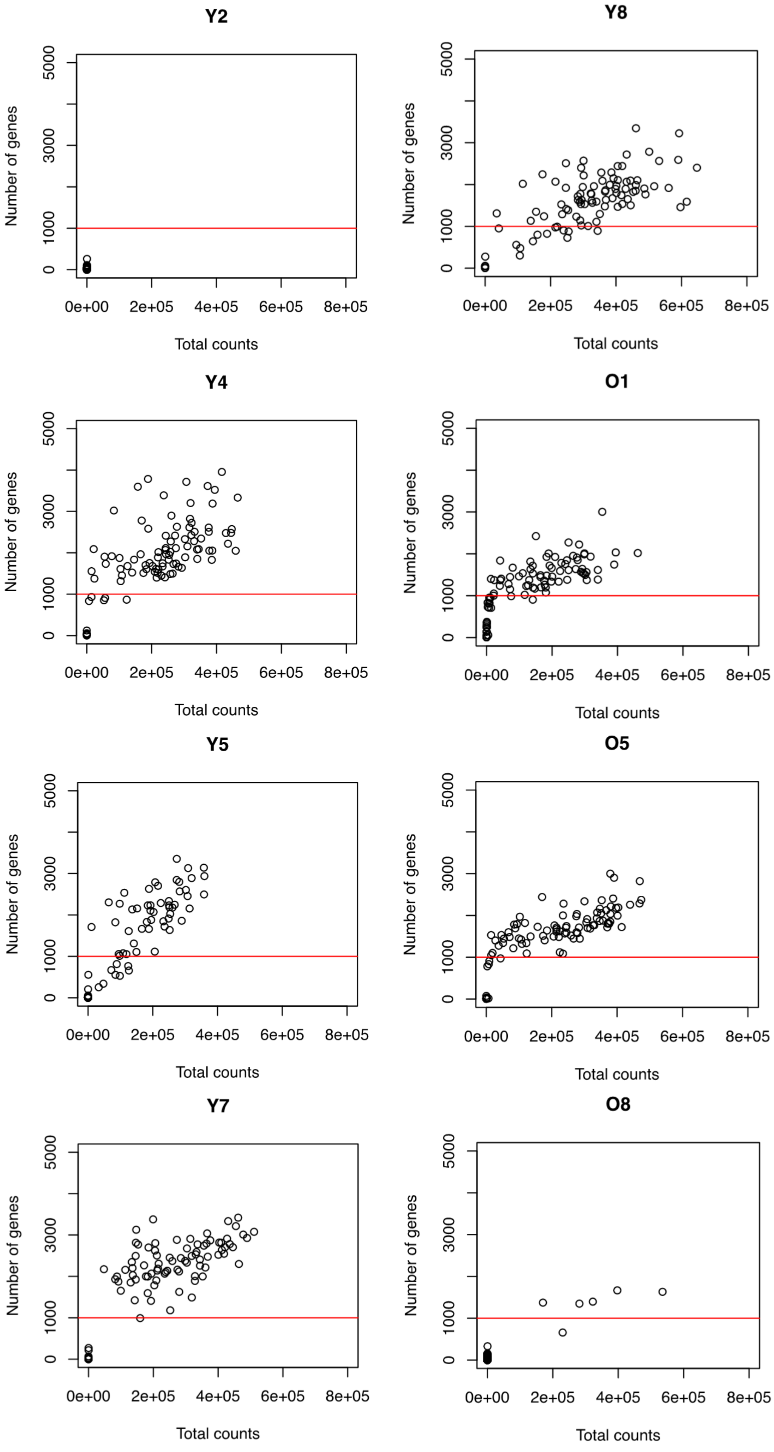


Figure 4.3: Gene count QC of cells: Scatterplots to depict the number of genes per cell. Each point represents a single cell. Each scatterplot represents an individual. Young individuals are depicted with a "Y" and old individuals with an "O". The red line is the cut-off threshold that was used (set to 1,000 genes per cell).

Individual	**Total cells**	**Pass**	**Chosen**
Y2	96	0	No
Y4	96	84	Yes
Y5	96	50	No
Y7	96	79	Yes
Y8	96	77	Yes
O1	96	69	Yes
O5	96	83	Yes
O8	96	5	No

Table 4.1: QC of single cell transcriptomes

different stages of the cell cycle. The reason I decided to assess TA-Hi MuSCs was in part due to their predominantly quiescent nature, which I hoped would minimize this potential issue. However, as a means of performing an additional quality control step, I assessed stages of the cell cycle across all cells that passed the initial quality control. This assessment of cell cycle was performed using a previously described classification algorithm (Scialdone et al., 2015) which assigns a stage in the cell cycle (G1, S or G2/M) to each cell. Since these cells are predominantly thought to be in G0, I would expect these to be classified as G1. This expectation was because G0 cells exit the cell cycle through G1 and these cells still reflect G1 cells in terms of DNA content and many other cellular attributes.

This analysis, highlighted that the majority of cells were in fact classified as G1, as was expected from what was known about these cells being highly quiescent (Figure 4.4). However, there were a number of cells that were classified in separate cell cycle stages. This could reflect the fact that within any homeostatic muscle there are constant, though minimal, requirements for tissue regeneration. The number of cells that seem to be classified outside of G1 appear to be variable between individuals. This further suggests that this could be related to homeostasis upon minor damage that is individual specific. Interestingly, there seem to be a slightly larger number of cells that were not in G0 in the younger individuals. This could be due to the increased activity and/or exercise exhibited in younger mice that results in more regeneration being required to maintain homeostasis. Whatever the reason, in order to ensure that I was studying the effects of ageing and not the potential effects of cell cycle, I decided to exclude from

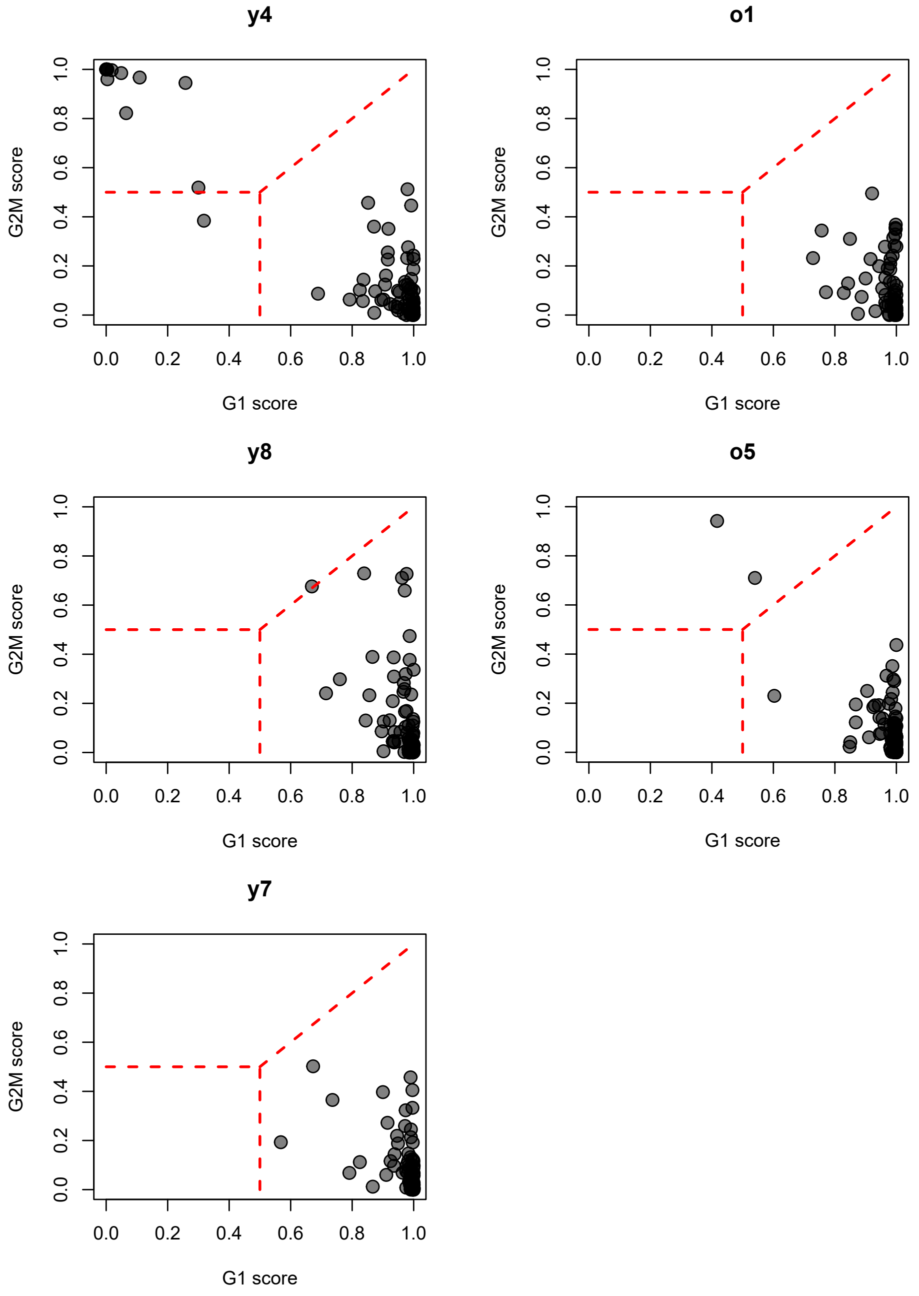


Figure 4.4: Cell cycle QC of cells: Scatterplots to display the cell cycle classifications defined using Scialdone et al., 2015. The red dotted lines depict the boundaries of the three states (G1 G2/M and S). Each individual is depicted in a separate scatterplot, where "Y" individuals are young and "O" individuals are old.

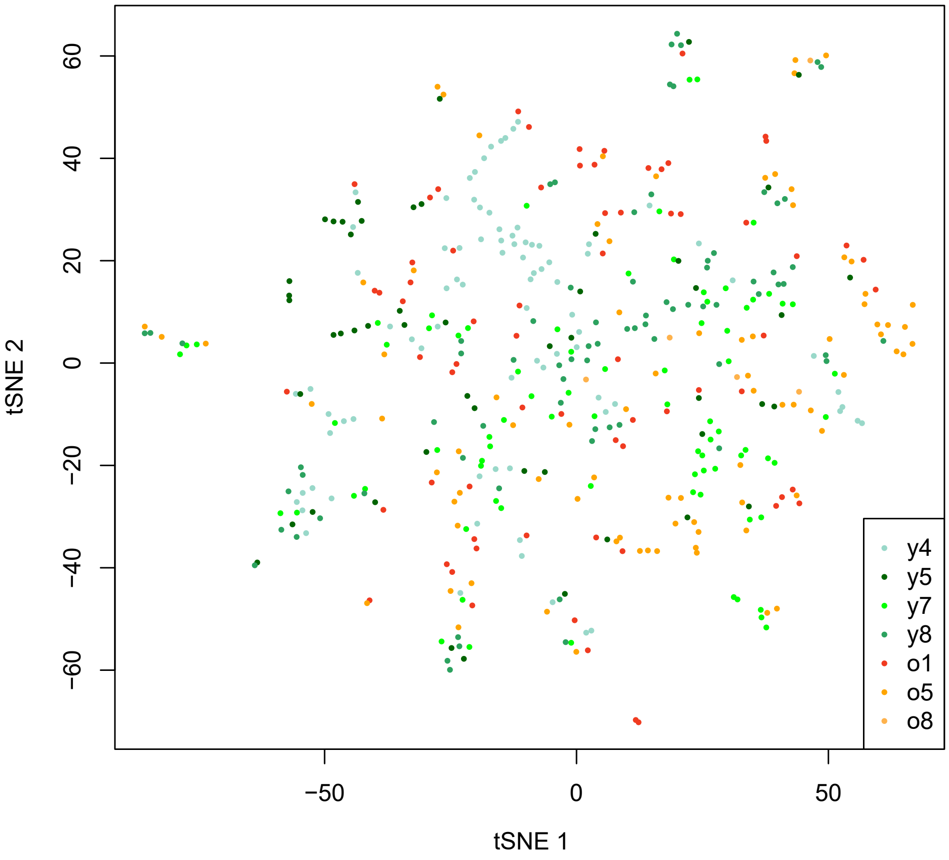


Figure 4.5: tSNE plot of TA−Hi MuSC: A scatterplot of the first two dimensions of the tSNE. Each individual is depicted with a colour, identified in the legend. Young individuals are depicted with a "Y" and old individuals with an "O". There was no idenitifable substructure.

any downstream analysis cells that failed to be classified as G1.

4.2.4 Comparison of transcriptomes between old and young

Having determined which transcriptomes met the quality criteria, I wanted to determine whether it would be possible to separate the old and young TA-Hi MuSCs from one another based solely on their transcriptome profiles. To assess this a tSNE dimensionality reduction-based approach was utilised (Materials and Methods section 2.4.7).

From this analysis it was not possible to identify any discrete substructure (subpopulations) in the dataset (Figure 4.5). Other clustering methods such as PCA and hierarchical clustering were also tested with similar results. This highlights that there were no major batch effects in our dataset between individuals. Whilst I had aimed to minimize any confounding batch effects by conducting young and old library preparation for any one individual in parallel, it was important to verify that there were no batch effects. In addition, it suggested that in this subpopulation of muscle satellite cells, any age-related differences in the functionality of these cells could not be detected purely from an expression level perspective. This suggests that the age-related changes that have been observed for this sub-population are not the result of a consistent age-associated drift in expression profiles, as has been previously postulated (Liu et al., 2013 and Sousa-Victor et al., 2014).

4.2.5 Comparison of transcriptome variability between old and young

Having determined that the cells appear to be homogeneous with regards to their expression profiles, I next assessed whether I could detect differences in the variability of the transcriptomes between young and old. To assess this, the distance between the squared coefficient of variation for each gene and the rolling median of the squared coefficient of variation for a given mean expression level were computed (Figure 4.6). Lowly expressed genes were removed from the analysis, due to them having large numbers of drop-outs. A more detailed explanation of the analysis can be found in the Material and Methods. The most highly variable genes were then analysed.

This analysis highlighted that there was an elevated average distance from the median for highly variable genes in the old cells relative to the young (Figure 4.7). In other words there was an increase in variability seen in the old cells compared to the young (Figure 4.7). This finding held for each individual within the study. In addition, this finding was independent of the exact number of highly variable genes that were assessed (top 100, 300 and 500 were all tested and showed the same behaviour). This is in keeping with the hypothesis that with age, satellite cells become misregulated resulting in their functional decline. Interestingly, there is limited overlap between the most highly variable genes from any one individual. This could partly be a result of the technical

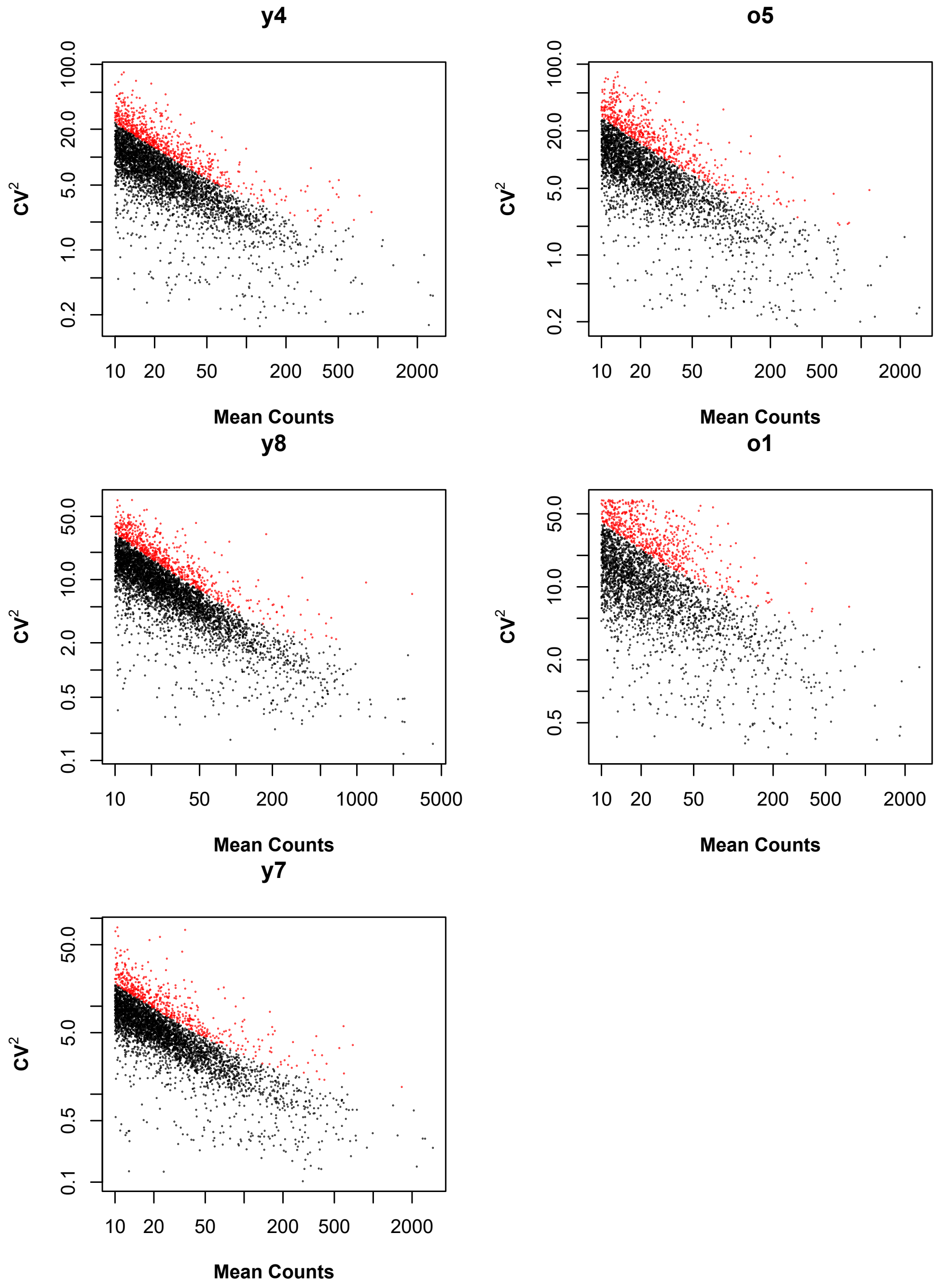


Figure 4.6: Squared coefficient of variance by individual: Scatterplots are shown for the squared coefficient of variance as a function of the mean count per gene. Highlighted in red are the top 500 most variable genes. These are calculated for each individual separately, as the 500 most distant genes from the rolling median. Young individuals are depicted with a “Y” and old individuals with an “O”.

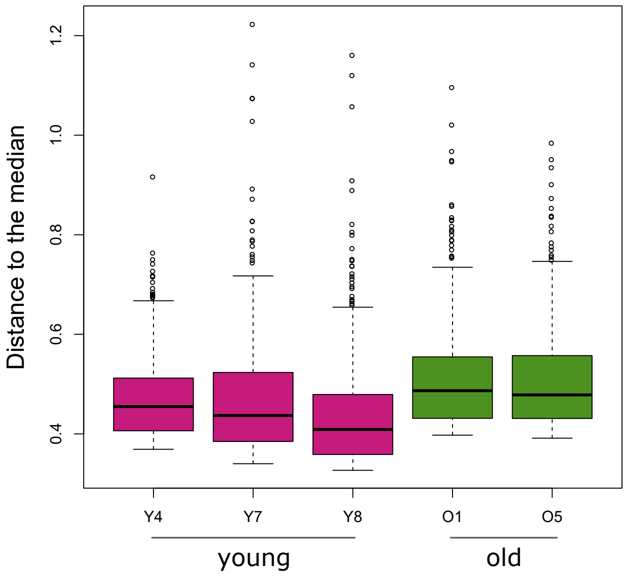


Figure 4.7: Transcriptional variability defined by distance: Boxplot of the distance to median of the 500 most variable genes per individual. Shown in pink are the young individuals and in green the old individuals.

limitations of the single cell approaches and the inherently low expression levels exhibited by these cells, but could also reflect an underlying lack of homogeneity in these cells. This increased variability seems unlikely to be due to sample preparation, since all libraries were amplified the same number of cycles and the library quantification from the bioanalyser did not highlight an inherently lower or higher expression level in the old TA-Hi MuSCs. Since I could not determine any inherent direct overlap of the most highly variable genes, I decided to assess whether the genes that were most or least variable in our dataset were enriched for any functional category. As such, I performed a gene ontology (GO) analysis of the most and least highly variable genes for each individual and for the old and young as groups. For this analysis, the background was defined as the genes expressed in each individual that were considered in the variability analysis.

This analysis highlighted that the least variable genes for both young and old individuals were enriched in gene ontology terms associated with basic cellular processes such as “ribosome” and “structural molecule activity” (Figure 4.8). This is not unsurprising owing to the importance of these processes. More curiously, I found that this was not the case for the highly variable genes in old cells when compared to young cells. I found

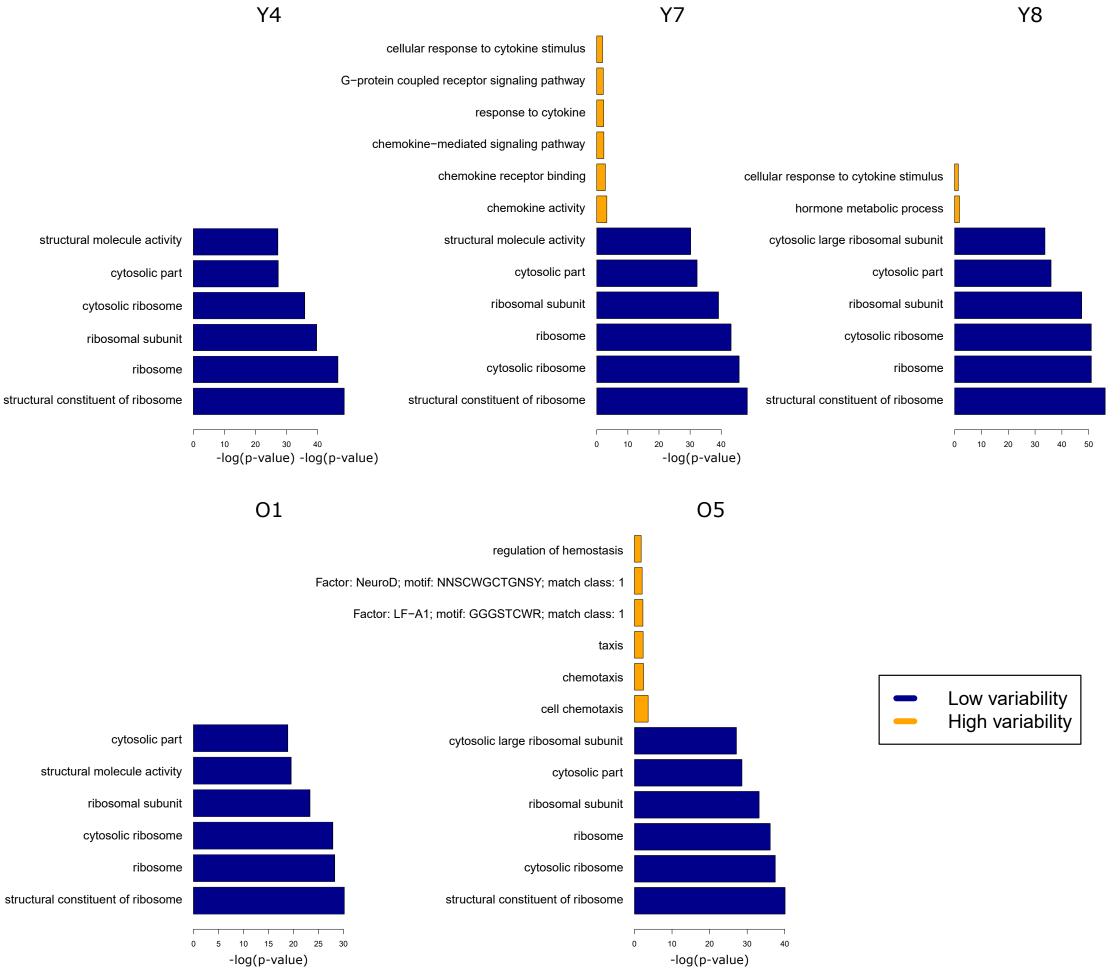


Figure 4.8: GO analysis of variable genes by individual: A barplot depicting the top 5 most significant GO terms for the most and least variable genes. Where there were 5 significant GO terms or less, all significant GO terms were visualised. GO terms assoicated with lowly variable genes are depiected in blue. GO terms associated with highly variable genes were visualised with yellow. Young individuals are depicted with a “Y” and old individuals with an “O”.

that highly variable genes in young cells were associated with terms such as “response to cytokine stimulus” and “chemokine receptor binding” (Figure 4.8). In contrast, I find that highly variable genes in old cells were associated with terms associated with “chemotaxis” (Figure 4.8). Interestingly, in aged individual O5, I found that there was a significant enrichment for the genes bound by the transcription factor NeuroD (Figure 4.8). This is interesting because NeuroD has an overlapping position weight matrix (PWM) with MyoD. Crucially, the activity of NeuroD is frequently lineage restricted owing to chromatin accessibility (Fong et al., 2012). As such, this finding suggests the tantalizing possibility that such regulation is eroded with age.

To assess cell-to-cell variability of these highly variable genes, a pairwise correlation analysis of these genes was performed between cells of a given individual (Materials and Methods section 2.4.9). These pairwise correlations were visualized as correlation heatmaps, one per individual (Figure 4.9) and summarized in a violin plot (Figure 4.10).

This analysis led to the interesting finding that the highly variable genes contained within old TA-Hi MuSCs are less correlated to one another than is seen for young TA-Hi MuSCs (Figure 4.10). This suggests that not only do old TA-Hi MuSCs have more variable genes *per se* but that the expression of these highly variable genes, are in of themselves more random in nature. In other words, with age there appears to be a loosening of the transcriptional network. Interestingly, such low correlations are not even seen in the characteristically heterogeneous environment of the mESC culture or the early embryo (Mohammed et al., 2017). However, this comparison should be taken with some caution, because it could be biased by the difference in number of transcripts that are expressed in any one cell being far lower in the TA-Hi MuSCs than the mESCs or early embryo cells, although this in itself could be considered biological.

Overall, these results show that the MuSC transcriptome is relatively stable, in terms of average expression levels of any given gene, with age. However, the variability of given genes and between cells within an individual is significantly increased upon ageing. There are many reasons why this could be the case, for instance, one could imagine that it is related to the number of cell divisions that have occurred resulting

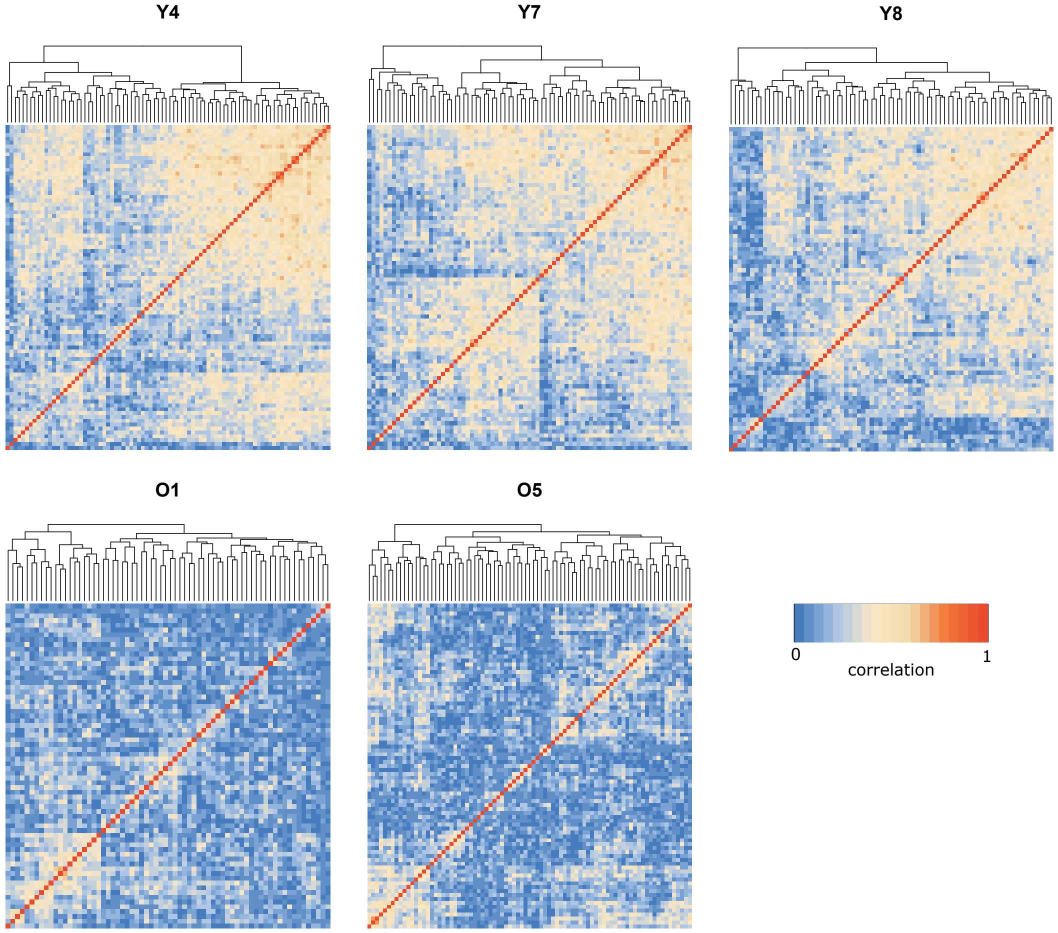


Figure 4.9: Transcriptional correlations between cells: A heatmap is shown per individual. Each position within each heatmap depicts the pairwise Spearman correlation coefficient for the 500 most variable genes between two cells. The correlation values range from 0 (blue) to 1 (red). Young individuals are depicted with a "Y" and old individuals with an "O".

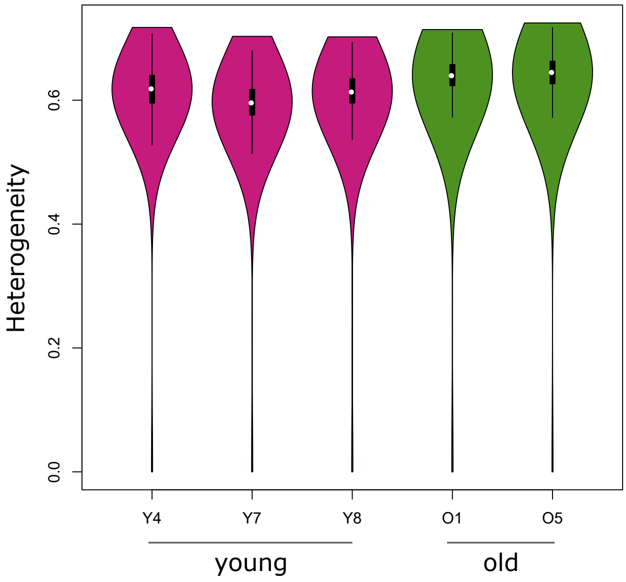


Figure 4.10: Violin plot of transcriptional heterogeneity: A violin plot of the transcriptional heterogeneity within each individual. This is computed as previously described in Mohammed et al., 2017 from the Spearman correlation coefficients. Shown in pink are the young individuals and in green the old individuals.

in the cells in of themselves being more distantly related or due to a loosening of the transcriptional network. Interestingly, it also alludes at a potential mechanism by which these cells have reduced functionality with age owing to a loss of coherent transcriptional regulation.

4.2.6 Quality-control of the methylation data

Next, I wanted to assess the methylation data that I had generated for these cells. Before any analysis could be conducted, I first had to check the quality of the cells. In terms of single cell methylation data, this predominantly means ensuring that there is a sufficient number of reads present within the library for a given cell, that the proportion of reads that align uniquely is not too low, and that the bisulfite conversion had been successful. In addition, I checked that the negative controls had remained uncontaminated throughout the protocol.

In contrast to the sequencing of the transcriptome where I aimed to sequence approximately ninety cells, in the case of the methylome sequencing I aimed to sequence

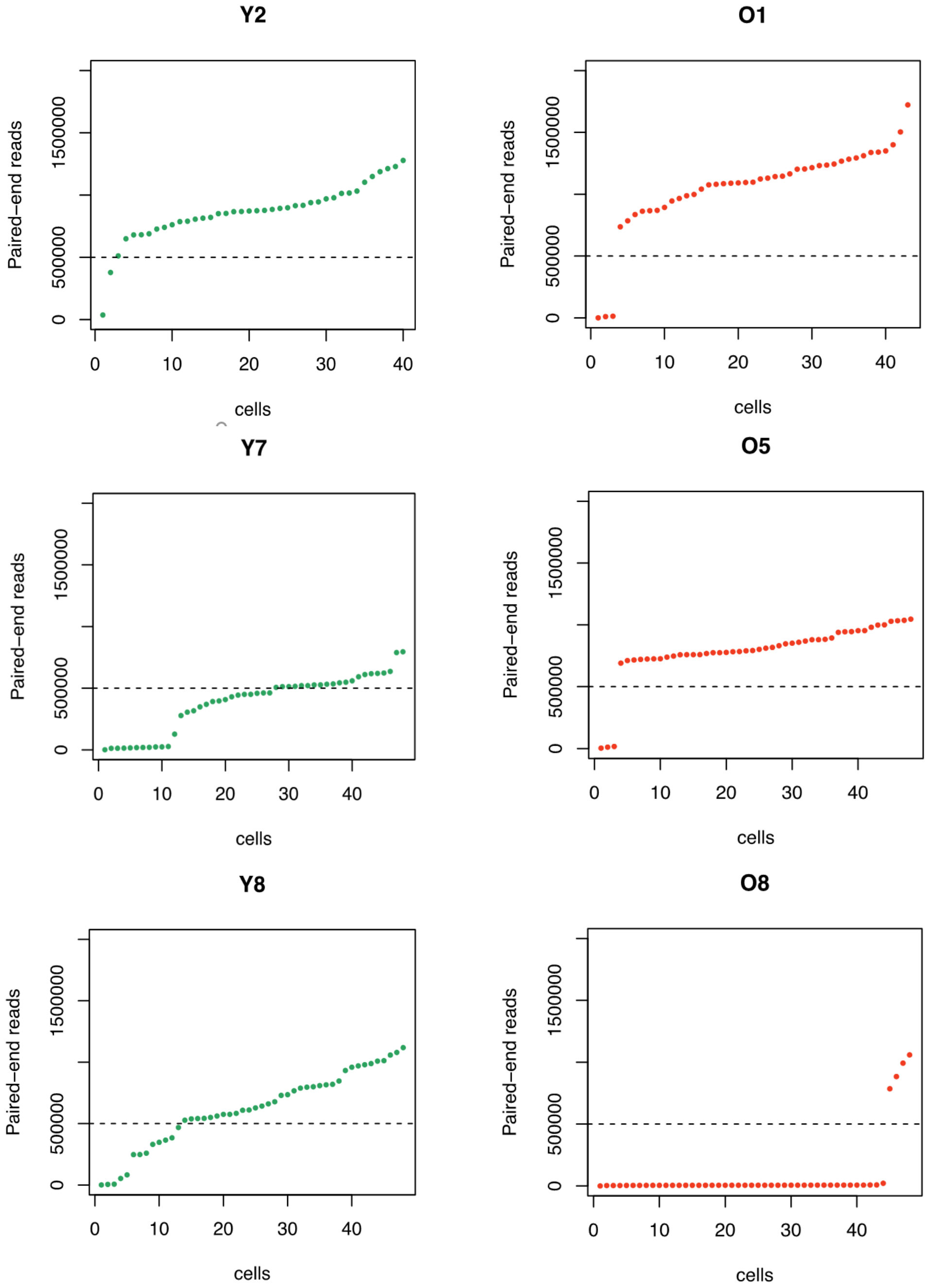


Figure 4.11: Paired end read count QC for the single cell methylomes: Scatterplots to depict the number of uniquely mapping paired-end reads that were achieved from each single cell (each dot). One scatterplot is shown for each cell. Young cells are depicted in green and old cells are depicted in red. The cut-off used for downstream analysis was 500,000 paired end reads, visualised here as a dotted black line.

approximately thirty cells. The reason for this owes to the present cost of sequencing single cell bisulfite libraries. The increased cost relative to the transcriptome is due to a number of factors. For instance, in bisulfite sequencing a greater proportion of the genome is being assayed. In addition, in single cell bisulfite sequencing a far greater fraction of the sequencing reads fail to map than is the case for single cell transcriptomics. Single cell methylomes are typically sequenced in a long paired-end read mode, compared to a single cell transcriptome that would be sequenced using a short read single end mode, exacerbating this cost inequality. The reason for this being that in the case of transcriptomics one simply wants to map the reads and there are four intact bases to map with. In contrast, for DNA methylomes, one is not merely interested in mapping the reads but also in specific information that is contained at specific positions within each read, namely the methylation status of cytosines, predominantly in a CG context.

From the quality analysis, I found that the majority of cells were successfully bisulfite converted. This is defined as having CHH and CHG methylation of less than 5% globally, which would likely be due to poor bisulfite conversion. In addition, I found that the majority of cells had more than 500,000 uniquely mapped and deduplicated reads. As such, I defined this as an arbitrary coverage threshold under which cells would not be considered for future analysis (Figure 4.11). The number of cells that remained following this cut-off is detailed in Table 4.2, sorted by individual. Due to a large number of cells being removed from individual Y7, this indvidual was removed from further analysis. In addition, I found that our negative controls had remained uncontaminated throughout the process, due to them having poor mapping quality (<10% of reads mapping). This gave us confidence that the single cell data I have generated is valid (data not shown).

However, this analysis of the quality of the libraries highlighted that there was a bias in the sequencing depth with the older samples on average having higher coverage than the younger samples (Figure 4.12). From an initial analysis of the samples (not shown), I found that such a bias was detrimental to the integrity of the downstream analyses. As such, the .bam files of all cells were randomly down-sampled such that they had a

Individual	Total cells	Pass	Chosen
Y2	40	38	Yes
Y7	48	20	Yes
Y8	48	35	Yes
O1	43	40	Yes
O5	48	45	Yes
O8	48	4	No

Table 4.2: QC of single cell methylomes

similar representation of reads before continuing with the analysis.

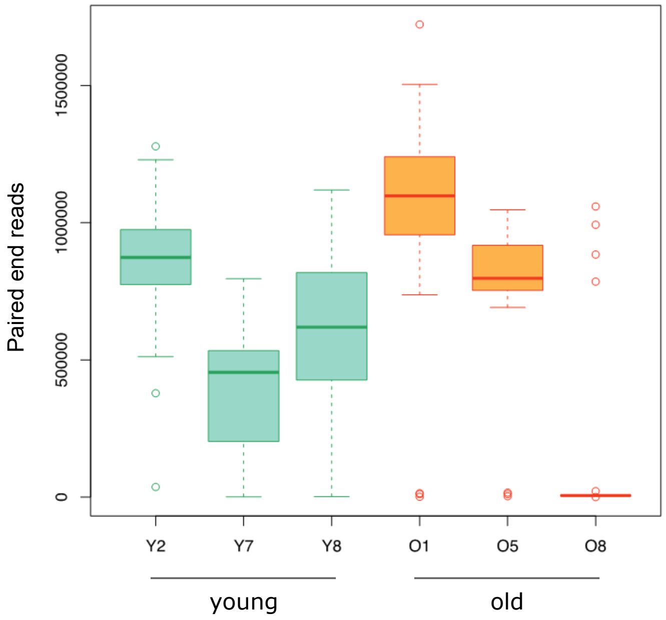


Figure 4.12: Sequencing depth bias in methylomes: Boxplot of the number of paired end reads. Each boxplot represents an individual. Young individuals are depicted in green and old individuals are depicted in orange.

4.2.7 Description of the methylation landscape of these cells

TA-Hi MuSCs have not previously had their DNA methylomes interrogated at a global level. Therefore, I wanted to first assess the DNA methylation profiles of the cells that I was using as my model system to study ageing. The first analysis that I conducted on the methylomes from the cells was to compute the mean methylation, on a cell by cell basis.

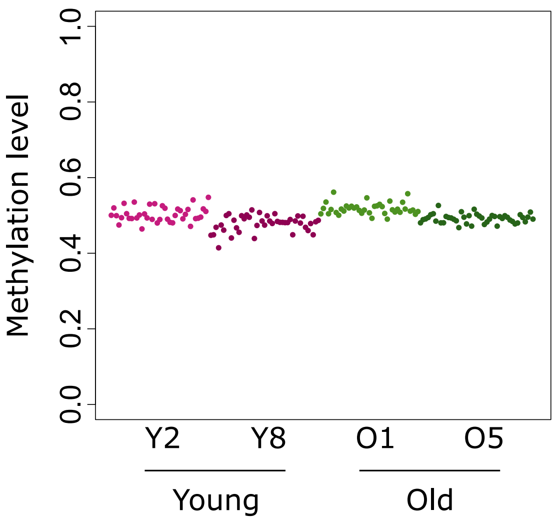


Figure 4.13: Global methylation levels are low in TA-Hi MuSCs: Plot of mean methylation levels for each cell. Young individuals are depicted in two shades of purple and old individuals are depicted in two shades of green. There is an apparent batch difference observed globally between Y2 and O1, and Y8 and O5.

This analysis highlighted that from a global DNA methylation perspective, these cells were unusual for a somatic cell, with global CpG methylation levels of approximately 45% (Figure 4.13). Typically, a somatic cell has approximately 70-80% CG methylation and indeed for many cell types and tissues this has been shown to be the case (Ehrlich et al., 1982). This includes other quiescent stem cell populations for which methylomes are available (Sun et al., 2014). This observation is interesting because it is typically assumed that a high level of DNA methylation is defined during early development and that this level of methylation is maintained at a relatively constant level across the lifespan of an organism (Hackett and Surani, 2013). This suggests that these cells may lose methylation after their specification. In addition, it has been shown for human and sheep differentiated muscle methylomes that the global levels of methylation are as expected (70-80%), suggesting that this methylation must be put back upon or soon after differentiation. One important point to bear in mind with this measurement of global methylation from single cell data, is that it will be slightly lower than obtained from a population-based sequencing experiment. This is likely due to the nature of the polymerases involved and the large number of amplification cycles required. However, even with this considered, the level of methylation that I observed for these cells is far lower than expected. Interestingly, I found that there was no large deviation with

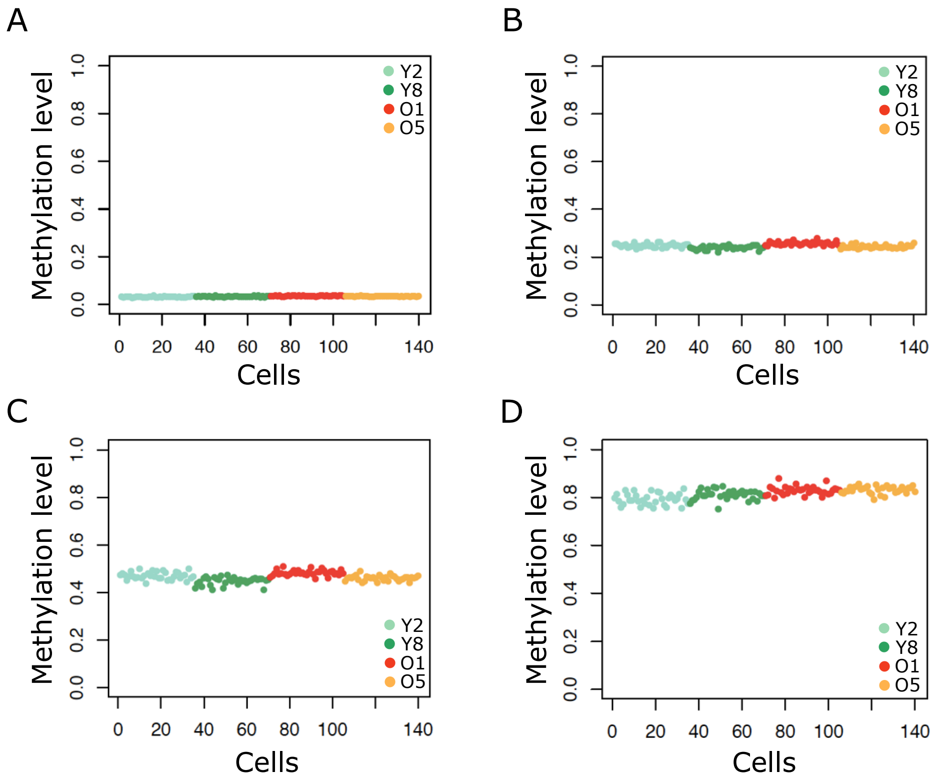


Figure 4.14: DNA methylation assessed for broad feature categories: (A-D) Scatterplots depicting levels of DNA methylation across broad feature classes. (A) Assessment of CGIs (A), exons (B), introns (including repeats; C) and Intra-cisternal A-type particle (IAP) elements (D) was performed. Each dot represents one cell. Each cell is ordered and coloured by individual. Green dots represent young individuals and orange/red dots represent old individuals.

regards to global level of DNA methylation across cells within an individual, nor across individuals nor age. This highlights the robustness of the single cell bisulfite sequencing protocol that I am using, but also the high homogeneity of our cells when assessed in this manner.

Next, I decided to assess whether this global reduction in DNA methylation relative to other somatic cell types was associated with a specific genomic feature or whether it was truly a genome wide phenomenon (Figure 4.14). This is of interest because it is commonly thought that the methylation of repetitive elements in the genome is responsible for ensuring genome stability in the majority of cell types, with a few notable exceptions during early development.

From this analysis I found that, as is typically seen for almost all cell types, DNA methylation was high in repetitive regions of the genome such as IAP elements and

LINE elements. In addition, I found that archetypally lowly methylated regions of the genome such as CpG island promoters (CGI-promoters) were lowly methylated. Surprisingly, I found that introns and exons appeared to be the regions of the genome that were behaving abnormally, and contained far lower levels of DNA methylation than is commonly observed in other cell types. This is exciting, because to my knowledge it is the first cell type to be assessed that exhibits this feature-specific patterning of DNA methylation. Typically, the level of methylation across exons and introns is correlated with the level of methylation over repetitive regions of the genome. For instance, in lowly methylated cell types such as 2i/LIF mESCs, low levels of methylation in exons and introns is observed alongside lowly methylated repetitive portions of the genome (Ficz et al., 2013). One exciting possibility for this low level of exon and intron methylation could be the highly quiescent nature of the stem cells that I am studying. I suggest this because high levels of DNA methylation over gene bodies (exons and introns inclusive) is commonly correlated in bulk datasets with increased expression of the gene under study (Kulis et al., 2012, Maunakea et al., 2010 and Varley et al., 2013). Furthermore, I found that independent of feature there was a subtle increase in DNA methylation with age. The implication of this being that lowly methylated regions gain entropy with age, whereas highly methylated regions become more ordered with age.

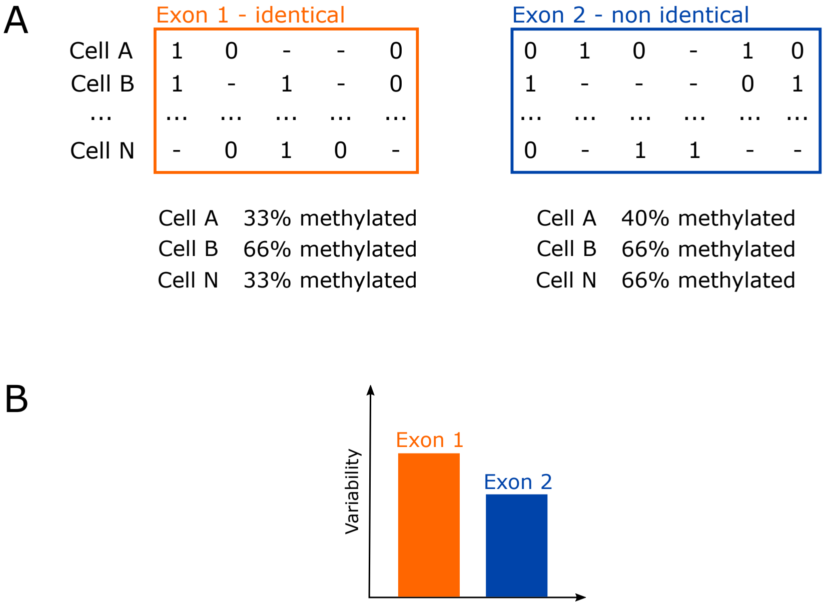
4.2.8 Calculation of inter-individual variability from single cell methylomes

Having assessed the nature of the unusual methylation landscape of these TA-Hi MuSCs, I next assessed whether there were differences in the methylome variability of different regions of the genome with age.

Methylation variability, in a similar manner to transcriptomic variability, is to a large extent influenced by the level of the metric across the population (i.e. transcriptional variability tends to increase with increased levels of expression). In the case of DNA methylation, this metric is the methylation of the population and variability can largely be defined as a function of methylation level. Unlike transcriptomics, where variability is increased with increasing mean expression, in the case of DNA methylation, variability is maximal at 50%. This is because methylation levels are derived from methylated and unmethylated bits of information. When assessing variability in DNA methylation at the single cell level, it is therefore important to ensure the population level of DNA methylation that is being measured is properly controlled. In addition, owing to the sparsity of single cell DNA methylation data, DNA methylation in a single cell context is commonly studied over windows. These windows are commonly defined as regions of the genome containing a certain fraction of the genome, whether defined in base pairs or number of underlying CG sites. This, however, introduces another complication in the study of single cell DNA methylation variation that is frequently overlooked: that the exact positions that are covered within a given window across a given number of cells will more often than not, not overlap equivalent positions. This means that attempts to utilize metrics such as standard deviation or variance of a mean level of methylation across a feature will potentially be misleading. This in turn can generate false impressions of the nature of variability across a feature or between cells. This is highlighted diagrammatically in Figure 4.15.

To address these inadequacies in the analysis of DNA methylation variability, I developed a computational approach based on Hamming Distance (HD; Materials and Methods section 2.4.12). Briefly, the mean methylation and HD from a given pairwise comparison is calculated for a given feature between cells within a given individual. These pairwise comparisons ensure that I only compare sites that are present in both cells such that there is no missingness. Importantly, each pairwise comparison can differ with respect to the exact positions that are utilized when comparing any two cells. This is important as it maximizes the number of events that are being studied within any particular comparison, whilst at the same time eliminating the missingness problem detailed earlier. The exact positions that are utilized in the pairwise comparisons are pre-defined by the features that are chosen for a given analysis. For instance, it would be possible to compute pairwise comparisons for CGI-promoters.

In order to correct for the variability-relation to mean methylation of each cell within



Exon 1 - identical

Cell A	1	0	-	-	0
Cell B	1	-	1	-	0
...	...	...	...	...	...
Cell N	-	0	1	0	-

Exon 2 - non identical

Cell A	0	1	0	-	1	0
Cell B	1	-	-	-	0	1
...	...	...	...	...	...	...
Cell N	0	-	1	1	-	-

Cell A	33% methylated
Cell B	66% methylated
Cell N	33% methylated

Cell A	40% methylated
Cell B	66% methylated
Cell N	66% methylated

Figure 4.15: The implications of missingness for variance calculations: (A) Schematic representation of two exons that are present to differing extents in cells A to N. Exon 1 depicted here is theoretically identical in its distribution of methylation across cells. In contrast exon 2 is variable across the cells under study. Methylation status of cytosines within a CG-context are displayed as 0 (unmethylated) or 1 (methylated). (B) A barplot of the variability determined from these two exons as assessed using the standard approach. Briefly, the standard deviation is computed from the average methylation of each cell for a given feature.

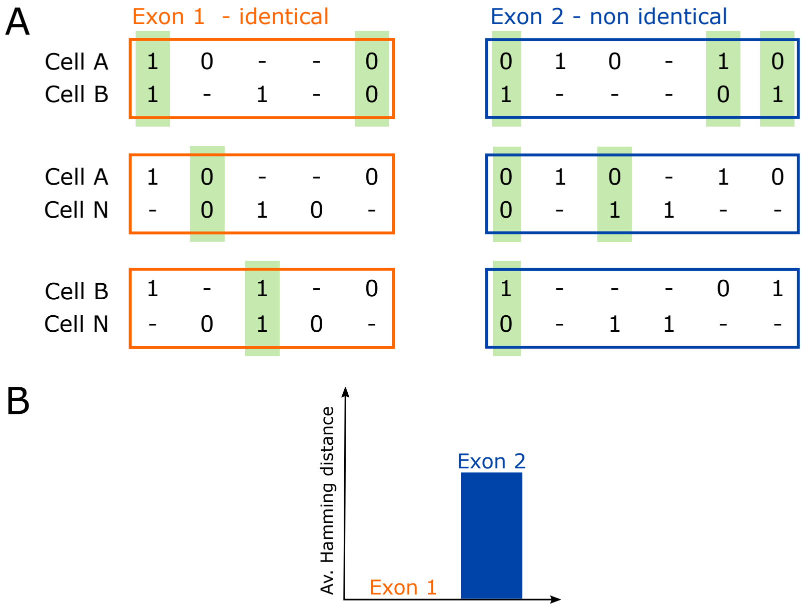

A
Exon 1 - identical
Exon 2 - non identical
Cell A 1 0 - - 0 | 0 1 0 - 1 0
Cell B 1 - 1 - 0 | 1 - - - 0 1
Cell A 1 0 - - 0 | 0 1 0 - 1 0
Cell N - 0 1 0 - | 0 - 1 1 - -
Cell B 1 - 1 - 0 | 1 - - - 0 1
Cell N - 0 1 0 - | 0 - 1 1 - -
B
Av. Hamming distance
Exon 2
Exon 1

Figure 4.16: Solution to the missingness for variability calculations: (A) Schematic representation of two exons that are present to differing extents in cells A to N. Exon 1 depicted here is theoretically identical in its distribution of methylation across cells. In contrast exon 2 is variable across the cells under study. Methylation status of cytosines within a CG-context are displayed as 0 (unmethylated) or 1 (methylated). Highlighted are cytosine positions that are present in both cells in each pairwise comparison. (B) A barplot of the variability determined from these two exons as assessed using the Hamming distance approach.

a pairwise comparison, I compute randomized pairwise comparisons (Materials and Methods section 2.4.12). With these metrics, it is possible to randomly assign zeros and ones to two independent vectors equal to the size of the comparison being made and such that they result in exact matches to the methylation states observed. These vectors can then be utilized to compute the HD. This can be done thousands of times and the resulting distribution can be compared to the true pairwise comparison to assess whether there is more or less variability than would be expected by chance. This is schematically described in Figure 4.16.

The results from this analysis showed that there was a huge difference between the expected heterogeneity for a given feature and the randomized pairwise comparisons (Figure 4.17). This was particularly the case for genomic features that had intermediate levels of methylation. This suggested that these features were perhaps not wholly unbiased in the distribution of their methylation states, and as such, perhaps I was artificially deflating the variability seen for a given feature. For instance, a region with half of sites having 100% methylation and the other half of sites having 0% methylation will be very different in terms of variability from a region where all sites are 50% methylated. However, although this was likely an issue for any independent comparison to random, these results did reveal differences between young and old in terms of variability.

Considering this issue from the previous approach to the analysis of variability, I next decided, prior to computing the HD for our given features, to first sub-divide the features based upon our knowledge of the state of methylation at each single cytosine position in a CG context in the genome (Materials and Methods section 2.4.13). To do this I computed the methylation status at every single CG site in the genome, from an *in silico* combined dataset containing all the single cells. I then filtered for sites that contained more than 10 reads and segmented the genome in to methylation bins of 10%. These bins could then be intersected with the feature annotations that I wished to compute over, and the metrics and randomized comparisons could then be calculated again. This time with a reduced likelihood of the results for the random analysis being confounded by the variable methylation levels that are present within a given feature,

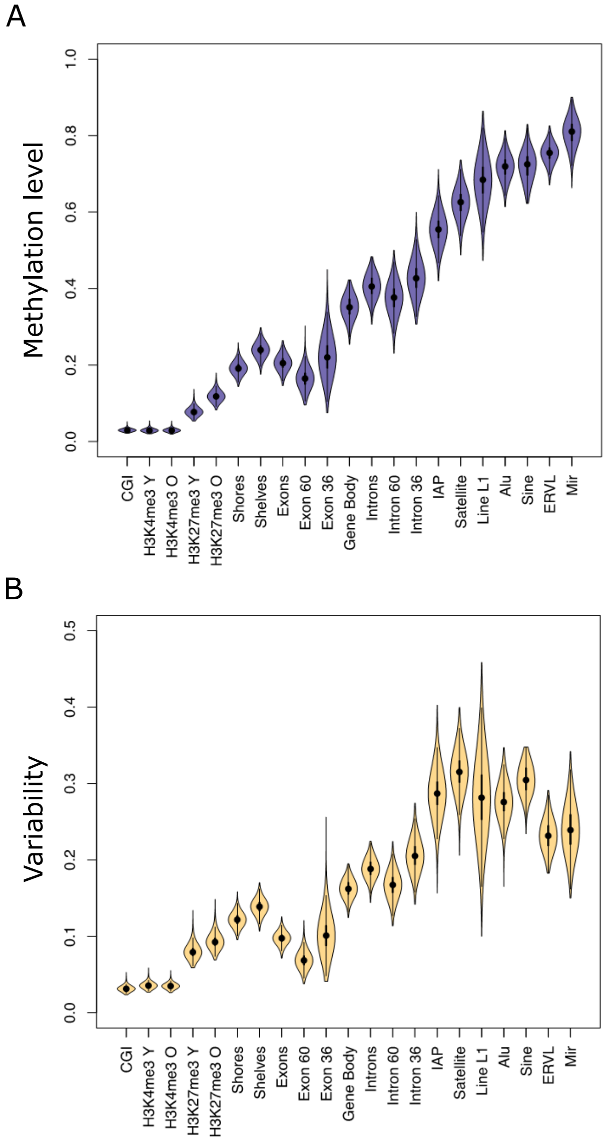


Figure 4.17: Variability in DNA methylation by feature: (A) Violin plot of average pairwise methylation for all individual-contained pairwise comparisons. Broad genome features were assessed (CGIs, CGI shores and shelves, exons, introns and various repeat classes. In addition, a number of muscle-specific features were assessed. Histone mark features studied were identified from MuSC ChIP data (Y=young cells and O=old cells). Exons and introns from genes upregulated upon regeneration after 36 and 60 hours were assessed (exon/intron 36 = 36 hours and exon/intron 60 = 60 hours). (B) Violin plot of Hamming distance of each individual-contained pairwise comparison. Features as defined for (A).

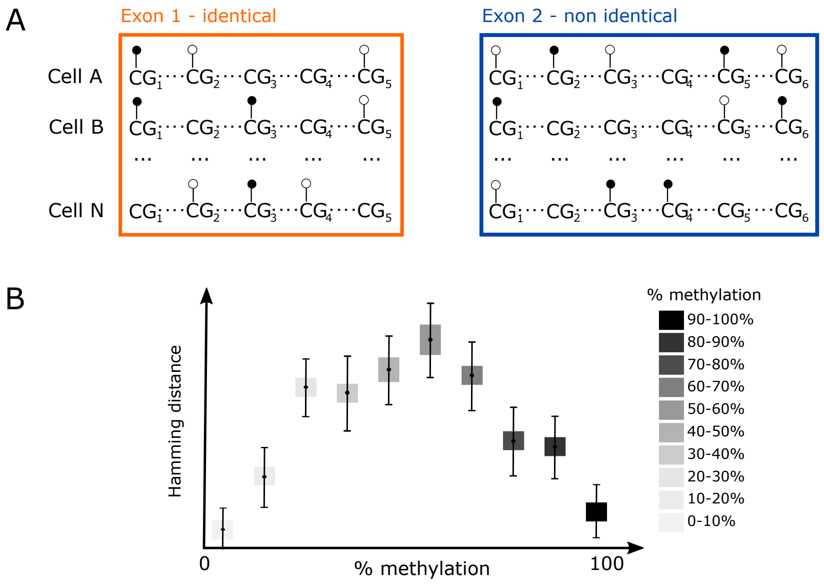


Figure 4.18: Variability with methylation level accounted for: (A) Schematic representation of two exons that are present to differing extents in cells A to N. Exon 1 depicted here is theoretically identical in its distribution of methylation across cells. In contrast exon 2 is variable across the cells under study. Methylation status of cytosines within a CG-context are displayed as empty (white) lollipops (unmethylated) or filled (black) lollipops (methylated), positions for which there is no information do not contain lollipops. The shading depicts the average methylation level of each cytosine across the whole population of cells (white=0% methylated and black=100% methylated). (B) Graphical depiction of the results for variability as a function of methylation, once the features have been split by population level methylation bins. The blue shaded line represents what would be expected by a wholly random process.

such as exons (Figure 4.18).

4.2.9 Comparison of methylome variability between old and young

Having defined this method for assessing methylome variability, I then wanted to assess whether there were regions in the genome that were more or less variable than would be expected by chance and how these compared between young and old. The first feature that I assessed was CGIs (Figure 4.19).

The results from this CGI analysis highlighted the improvement in this randomization approach now that I am considering the “bulk” methylation for each site that I am assaying, with the majority of sites exhibiting the same level of heterogeneity as ex-

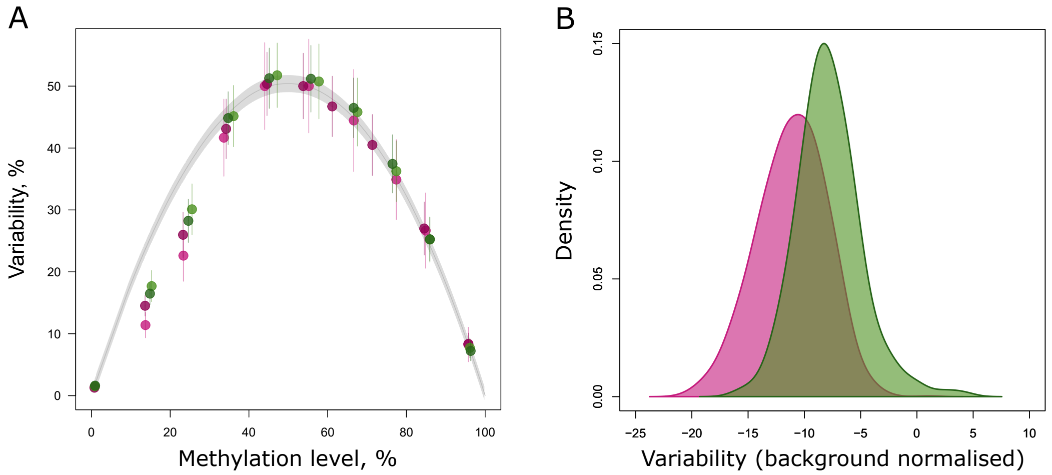


Figure 4.19: Lowly methylated CGIs are homogeneous: (A) Plot of variability in CGIs as a function of methylation for each individual. The two young individuals are shown in differing shades of purple and the two old individuals in differing shades of green. The grey curve represents a smoothed fit of the random trials from each individual combined. Error bars for each individual are described by the median absolute distance (MAD). (B) Density plot of the CGIs that are 20-30% methylated. Background has been subtracted from each individual pairwise comparison. The young individuals are represented in pink and the old individuals in green.

pected by chance. Excitingly, I observed that with age, sites that are lowly methylated gain heterogeneity and sites that are highly methylated lose heterogeneity in CGIs. Strikingly I found a strong CGI-specific homogeneity in both young and old cells for sites that are 10-30% methylated in bulk. Although the young cells were more homogeneous than the old cells. Importantly, this is above what would be expected by chance. These results suggest that lowly methylated CGIs are specifically maintained at a higher degree of homogeneity than the rest of the genome and that with age this maintenance is eroded.

I next decided to investigate a broad range of feature annotations including; exons, introns, repeat elements and published histone marks for these cells (H3K27me3, H3K4me3, H3K36me3). CGIs were assessed again too for comparison. The calculated difference in heterogeneity between young and old for these features is shown in Figure 4.20.

The results from this analysis showed that sites in the genome that were lowly methylated tended to gain heterogeneity with age. In contrast highly methylated sites either did not change substanially or became more homogeneous with age. Perhaps unsurpris-

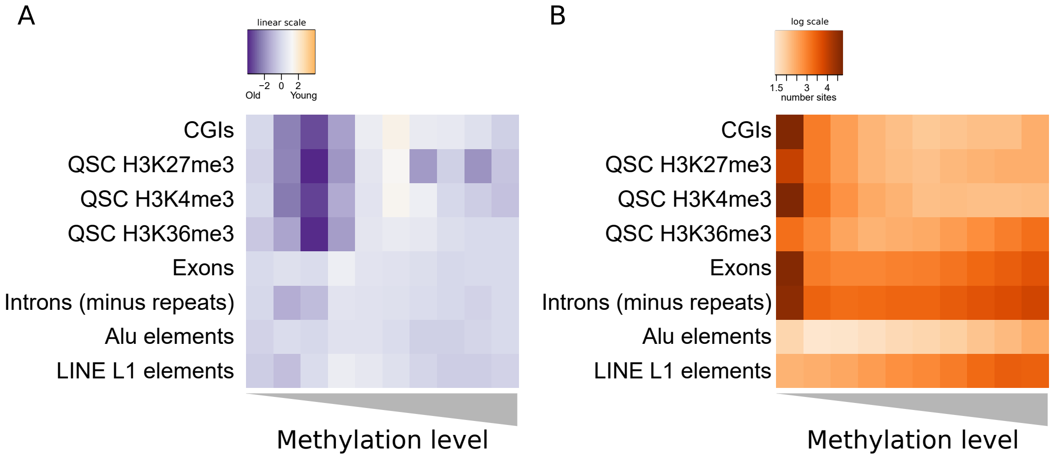


Figure 4.20: Methylation variability changes with age are feature dependent: (A) Heatmap of the difference in variability between old and young for a number of features. The histone modifications are labelled as being from quiescent satellite cells (QSCs) as defined in the original paper (Liu et al., 2013). Differences are represented on a linear colour scale, where blue is more variable in old, white is no change in variability and orange is more variable in young. Methylation levels were segregated into bins of 10%. (B) Heatmap of the average number of sites utilised for each variability comparison. Colours are shown on a log scale, where the darker the shade of orange the more sites were utilised.

ingly, histone marks that are associated with promoters showed strikingly increased heterogeneity similar to CGIs (of which many fall within promoter regions; Figure 4.20A). The number of sites in each comparison is shown here to evaluate whether this could be biasing the results (Figure 4.20A). This does not seem to be the case as there is no apparent relation between the number of sites assayed and the heterogeneity metric computed.

Having assessed these broad annotated feature classes, I next computed HD using this method taking bins of CpG density as my feature classes. This was of particular interest because we know from biochemical studies that the different enzymes involved in the process of adding and removing methylation have differing processivities. As such, it could be imagined that regions of the genome will differ in terms of variability as a function of CG density, since this will affect processivity. To assess this, I defined overlapping regions of even CG coverage throughout the genome, with the caveat that a CG site had to be observed in at least one of our samples (Materials and Methods section 2.4.14). These regions were then split into bins of CG density. To ensure that the bins of CG density were not biased by the presence of Ns in the genome scaffold,

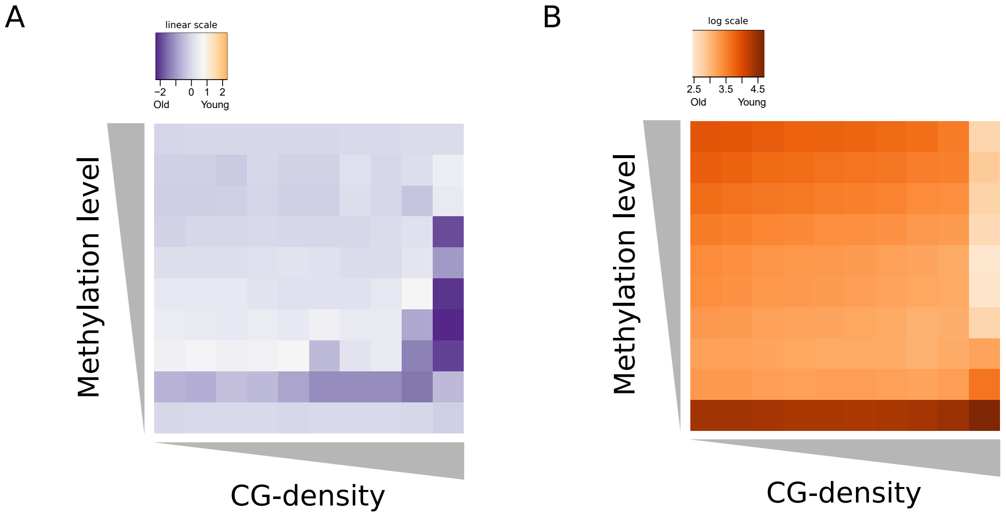


Figure 4.21: Age−dependent changes in methylation variability are CG−density dependent: (A) Heatmap of age-dependent changes in DNA methylation as a function of CG-density. Differences are represented on a linear colour scale, where blue is more variable in old, white is no change in variability and orange is more variable in young. CG-density was segregated into 10 bins. These bins were defined on a logarithmic scale. Methylation levels were segregated into bins of 10%. (B) Heatmap of the average number of sites utilised for each variability comparison. Colours are shown on a log scale, where the darker the shade of orange the more sites were utilised.

CG densities were recalculated from the genome fasta files (Materials and Methods section 2.4.14).

This analysis showed that changes in DNA methylation variability with age were different at differing CG densities (Figure 4.21A). The lowest levels of variability, relative to expectation, were seen for the highest levels of CG density. In addition, at these high CG dense regions I found that the level of variability was dependent on the level of DNA methylation and that variability was lowest for 10-30% methylation. This is perhaps not surprising since I previously saw this phenomenon for CGIs with 10-30% methylation. However, there are far more positions being considered in this CG-density analysis. This suggests that this could be a genome-wide phenomenon relating to CG-density and methylation level and not merely a CGI phenomenon. The number of sites were again computed to assess whether these could be biasing the results 4.21B). This did not appear to be the case.

Lastly, I wanted to assess whether the epigenetic clock sites that I defined previously (Stubbs et al., 2017) or regions surrounding these sites would become more or less

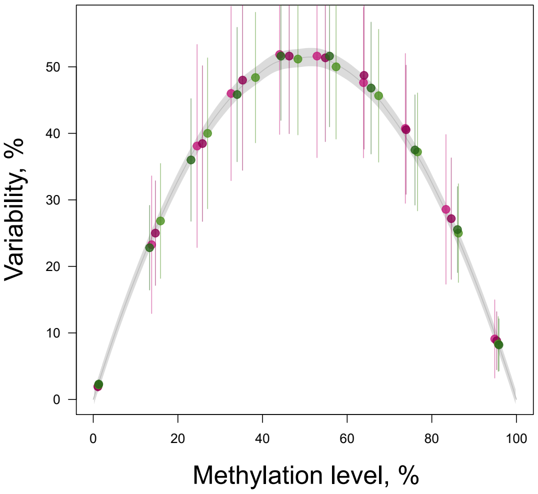


Figure 4.22: Clock−containing regions are as heterogeneous as the background: Plot of variability of clock-containing regions as a function of methylation for each individual. The two young individuals are shown in differing shades of purple and the two old individuals in differing shades of green. The grey curve represents a smoothed fit of the random trials from each individual combined. Error bars for each individual are described by the MAD.

heterogeneous with age. As such I defined fixed distance (5kb) regions around the positions defined in my elastic-net regression model and computed HD for young and old.

From this analysis I found that the regions containing the clock sites in TA-Hi MuSCs mimicked the changes observed for the rest of the genome. Namely, that all regions on average gained methylation, resulting in sites with methylation below 50% increasing in variability and sites with methylation above 50% decreasing in variability. This finding was independent of the directionality of the clock weight (data not shown). In addition, in contrast to what was observed for 10-30% methylated CGIs, the clock regions were not more homogeneous than expected by chance. As such, these changes in variability can be predominantly explained by changes to the DNA methylation levels of the regions themselves. These findings are interesting because they do not align perfectly with what is expected of these sites in bulk tissues, namely that they would all tend to increase in entropy. This could be a reflection of the unusual nature of the DNA methylation in these cells and the fact that 5kb regions around these sites are being observed, not simply the sites themselves. Alternatively it could be related to the stem cell nature of these cells, since no epigenetic predictor to date has assessed predictor site changes in adult stem cells with age.

Together these findings support the clock sites being a representation of the failures of an epigenetic maintenance system and not a "special" set of sites or positions that behave counter to the rest of the genome.

4.2.10 Calculation of methylation patterning from single cell methylomes

Next, I wanted to develop a method that I could use to assess the patterning of methylation in the genome from single cell data. In essence, I wanted to assess whether the proximity to a neighboring cytosine in a CG context was capable of influencing the methylation status of that site. As such, I developed a computational approach to assess this (Materials and Methods section 2.4.15). A schematic of this is shown in Figure 4.23.

Briefly, from each individual cell I determined the cytosines in CG context for which there is coverage. I then intersected these positions with our features of interest. For each cytosine in a CG context that falls within a given feature I then computed the distance to every other cytosine in a CG context within that given feature. I then assessed for each cytosine-to-cytosine distance whether the two cytosine bases share the same methylation status. In other words, were they both methylated, unmethylated or one methylated and the other unmethylated. I computed this for each individual feature within our list of features for each individual cell. In addition, I subset methylated and unmethylated reference cytosines when computing this similarity metric. This separation was conducted so that I could assess whether there was any difference between similarity score based upon whether it was being computed for methylated reference cytosines or unmethylated reference cytosines. For instance, it could be hypothesized that methylated sites have a stronger similarity score than unmethylated states owing to the processivity of the DNMT enzymes. These similarity scores for each distance from a given cytosine can then be averaged across the various features, such that a relatively continuous similarity metric can be defined.

Similar to the variability metrics, much of this similarity score was defined by the in-

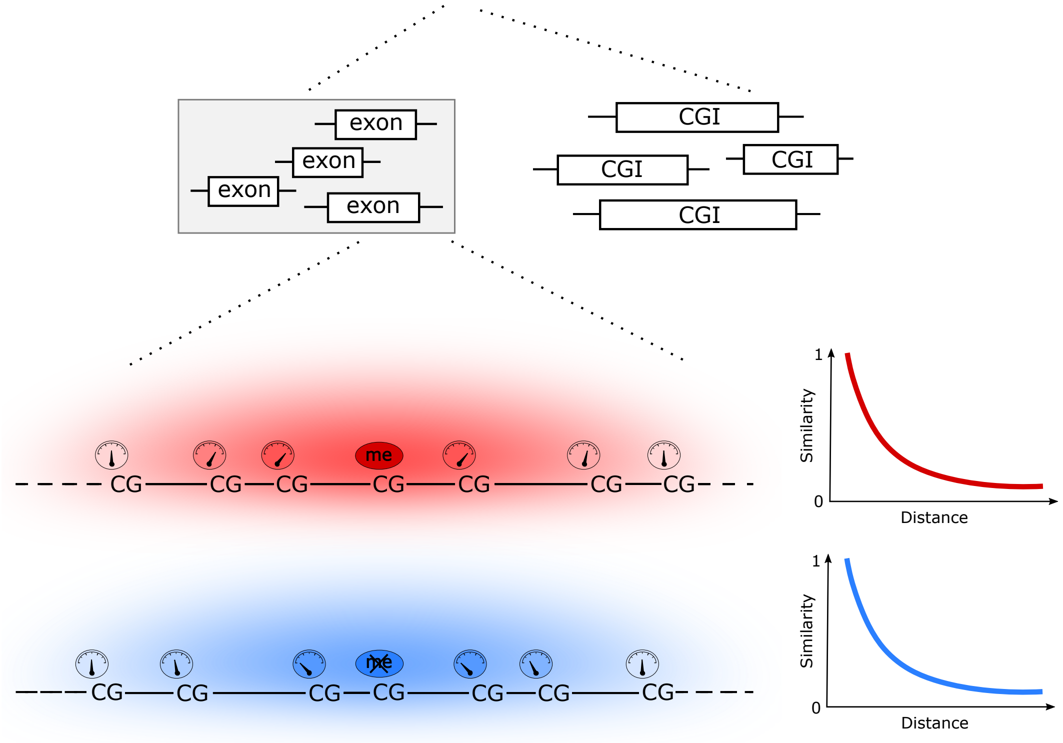


Figure 4.23: Schematic of methylation patterning within single cells: Schematic of the analysis of methylation appterning. In the diagram an example single cell is shown, from which multiple features can be assessed for patterning (shown here are CGIs and exons. The essence of the computational assay is to assess the influence that a reference methylated cytosine (red) or unmethylated cytosine (blue) can have on the methylation status of neighbouring cytosines.

herent methylation status for each feature. As such, similar to the variability metric I defined a randomized control set. In brief, I indexed each cytosine position within each given feature of interest. These indexed positions were then randomly shuffled within each given feature for each cell independently. This enabled a set of randomly defined similarity scores to be computed, enabling me to determine whether what I was observing was more or less than would be expected by chance. Unfortunately, the randomized metric is relatively expensive to compute 1000s of times for each individual cell within our study. Therefore, to assess whether the method was performing as anticipated, I assessed CGIs in one cell derived from a young individual (Figure 4.24).

The results from this analysis showed that I was able to compute similarity scores for both methylated and unmethylated reference cytosines in CGIs. Interestingly, I saw that similarity score decays with increasing distance for both the random and the actual values for the methylated reference but not the unmethylated reference. However, I saw that the slope of the decay with distance is greater for the actual values than the random values. As such, I found that at shorter distances (less than 100 bp), the similarity score for the actual values is greater than that for the random values. This suggests that the knowledge of a neighboring methylated or unmethylated cytosine at a given position increases the predictive power of determining the status of a given cytosine under study over and above what would be possible from methylation status of the feature alone. This is unsurprising for CGIs and is the reason that many DNA methylation imputation-based approaches utilize neighboring site information (Angermueller et al., 2017 and Zhang et al., 2015). As expected, I found that there is a difference between the similarity score for methylated and unmethylated reference positions, with unmethylated sites being far more likely to be neighboring unmethylated sites than methylated sites. This is unsurprising owing to the relatively low levels of methylation found in CGI features throughout the genome, and more specifically in our dataset. Interestingly, when I subtracted the random from the actual data, in an attempt to normalize to the background expectation, I found that the methylated reference sites exhibited an increased similarity over and above background relative to

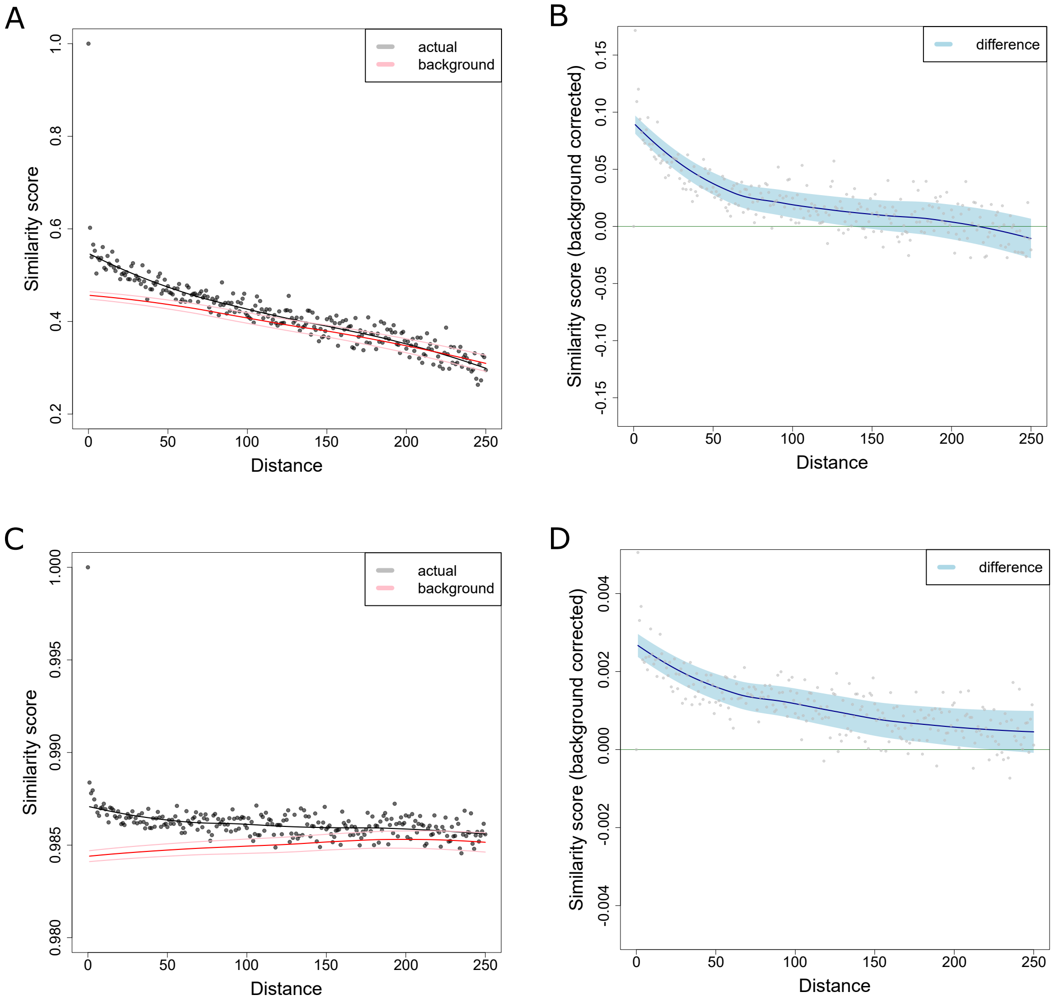


Figure 4.24: Example patterning of a cell in CGIs: (A and C) Similarity score of cytosines in CGIs in an example young cell. The similarity score displayed is that derived when the reference cytosine is methylated (A) or unmethylated (C). Shown in black are the real values for the similarity and shown in red is the background comparison (standard deviation displayed as two pale red lines either side of the smoothed fit). (B and D) Background normalised similarity plots. The reference cytosine is methylated in B and unmethylated in D. The blue line is a smoothed fit of the data, and the shading represents the standard deviation of the background.

the unmethylated reference sites. This is interesting because it suggests that methylated sites contain more information governing neighborhood methylation status than unmethylated sites. This could, however, also be due to the levels of methylation being assessed. This analysis would need to be conducted on additional features with differing methylation levels in order for this to be assessed more clearly.

Unfortunately, owing to the time taken to derive each random computation for each cell, it was not feasible to run the randomized analysis thousands of times. As such, I defined 10 randomised controls for each list of features for each cell. These were then averaged to compute the difference. The standard deviations for these randomized comparisons were also recorded, such that this information could be used to define confidence intervals around the randomized trials that could be used to assess deviation from the background expectations. In addition, this information would allow comparisons to be drawn from multiple cells and multiple individuals, enabling me to assess age-related differences.

In short, I have developed a method that seems capable of assessing similarity as a function of distance across features within single cells.

4.2.11 Comparison of methylation patterning between old and young

This method has the benefit of averaging across each feature for a given distance, allowing many data points to be considered in any specific analysis and resulting in a finer, more continuous distribution of values. The first comparison that I made using this analysis was of similarity in CGIs between young and old cells (Figure 4.25).

From this analysis, I found that old cells tended to have more similarity between neighbouring cytosines than young cells. Interestingly, I found that this is true for both methylated and unmethylated reference positions. Although the results were more striking for unmethylated positions. This result suggests that with age the information carried by a methylated position is greater than is observed in a young cell. Since this is background normalised to remove any differences that could be due to differences

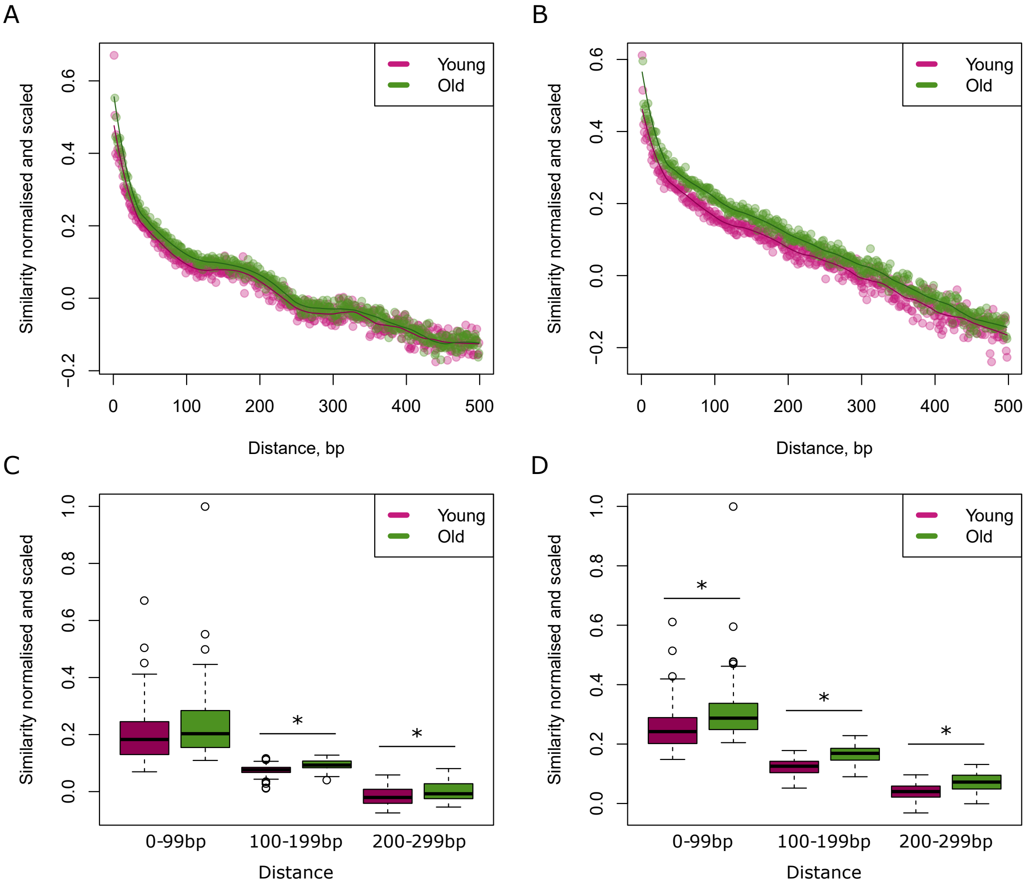


Figure 4.25: Neighbourhood similarity in CGIs: (A and B) Scatterplots showing similarity score as a function of distance (bp) for methylated (A) and unmethylated (B) reference cytosines within CGIs. (C and D) Boxplots of similarity score for windowed distances of 100 bp within CGIs. Methylated reference similarities are shown in C and unmethylated reference similarities in D. Asterisks reflect significant differences between young and old (Mann-Whitney Test; Bonferroni corrected p-value <0.05).

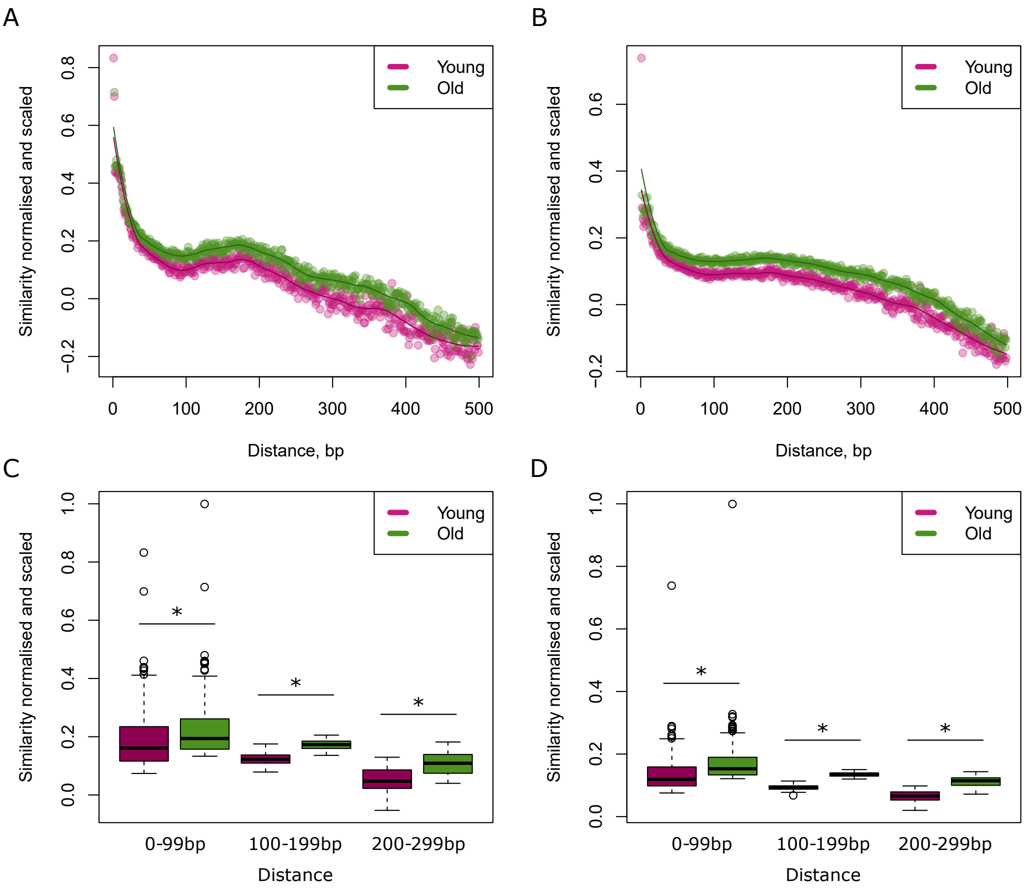


Figure 4.26: Neighbourhood similarity in exons: (A and B) Scatterplots showing similarity score as a function of distance (bp) for methylated (A) and unmethylated (B) reference cytosines within exons. (C and D) Boxplots of similarity score for windowed distances of 100 bp within exons. Methylated reference similarities are shown in C and unmethylated reference similarities in D. Asterisks reflect significant differences between young and old (Mann-Whitney Test; Bonferroni corrected p-value <0.05).

in DNA methylation levels between the various cells, this suggests that the manner in which DNA methylation is being added and removed is resulting in an increase in the order within CGIs over time. One possible explanation for this would be that over time DNA damage results in the removal of more heterogeneous regions. When DNA methylation is re-established within these regions it is re-established in a manner defined by broader methylation levels of the region and not so locally defined.

Next, I decided to assess whether the same behavior could be seen for exons (Figure 4.26). I found that indeed the same pattern of change with age was observed in the case of exons. In fact, I found that this pattern of increasing similarity was more striking than that seen for CGIs. Curiously, the decay with distance is more distorted

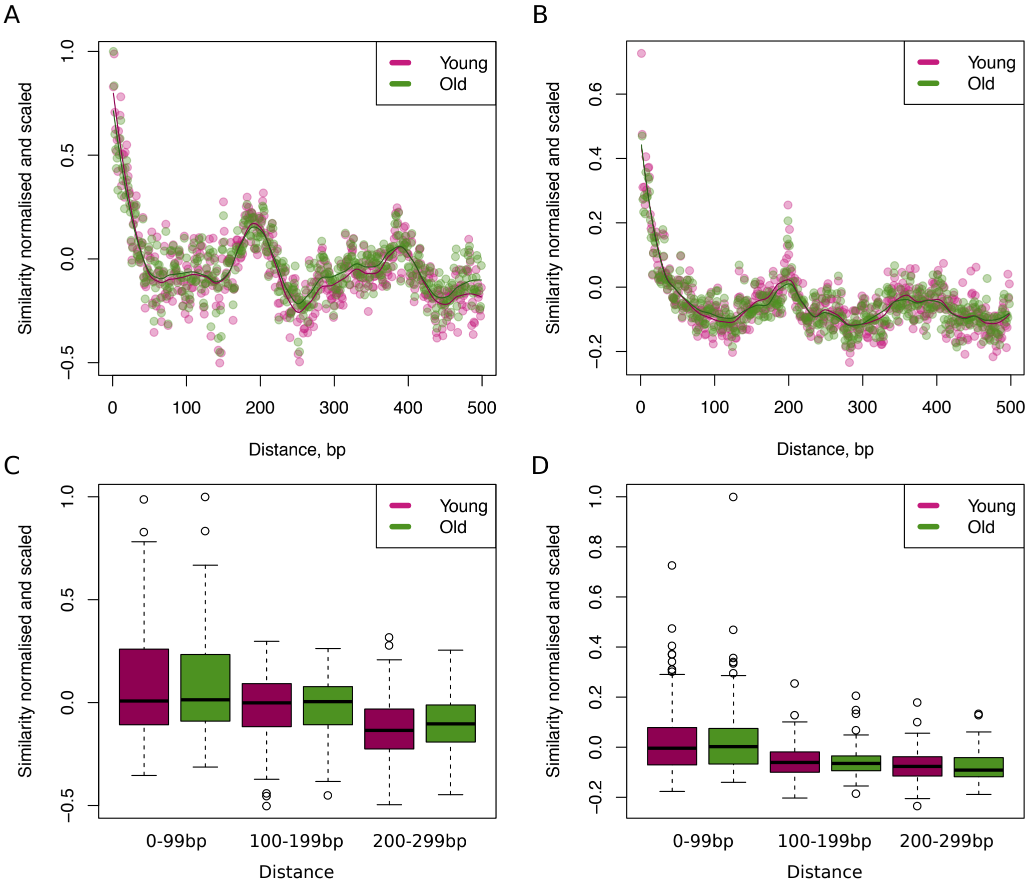


Figure 4.27: Neighbourhood similarity in L1 LINEs: (A and B) Scatterplots showing similarity score as a function of distance (bp) for methylated (A) and unmethylated (B) reference cytosines within L1 LINEs. (C and D) Boxplots of similarity score for windowed distances of 100 bp within L1 LINEs. Methylated reference similarities are shown in C and unmethylated reference similarities in D. There were no significant differences between young and old (Mann-Whitney Test; Bonferroni corrected p-value <0.05).

for exons than was seen for CGIs (Figure 4.25A and B, Figure 4.26A and B). One explanation for this is the larger variation in size seen in the case of exons. These results suggest that the behavior I observed for CGIs, is not specific to that genomic feature but is potentially a more genome wide phenomenon. To assess, this I decided to examine what would happen in the case of long interspersed elements (LINEs) in the genome.

In contrast to exons and CGIs, LINEs are highly methylated in this cell type, and in general appear to become more homogeneous with age. These results showed that there was no difference between young and old cells with regards to similarity (Figure 4.27). This was true for both methylated and unmethylated reference positions. This result was important because it suggests that the gains in similarity observed for other features were not the result of sequencing bias or batch. These results also suggest that although gains in similarity are seen for exons and CGIs, perhaps this phenomenon is not wholly genome-wide.

Lastly, I wanted to assess whether the regions surrounding the epigenetic clock sites that I had identified also followed this behavior (Figure 4.28). This was of particular interest as it could suggest a mechanism behind the changes in DNA methylation that are observed for the clock. This may seem counter intuitive, since one of the key observations from the analysis of the sites of my epigenetic clock and others, is that with age there is an increase not a decrease in entropy. However, it should be noted that we and others have identified epigenetic clock sites to be associated with regions containing very defined boundaries in terms of methylation status, such as CpG shores. As such, one could imagine, owing to randomly occurring events of DNA damage and increasing order within individual cells in repairing this damage, that on a bulk level this would result in increased entropy or variability.

From this analysis, I saw that with age there was an increasing amount of predictive power from neighboring sites within these regions. This is in keeping with the rest of the genome. Excitingly, this was the largest increase in similarity seen of any of the features assessed, this was particularly true for unmethylated reference cytosines. This suggests that these regions, although in keeping with the rest of the genome are exaggerated in this regard and could be one of the reasons they are identified as predictive in my multi-tissue predictor.

In summary, I find that with age there is an increasing similarity within a given cell between neighboring positions. I observe that this is in contrast to the heterogeneous nature of the changes in variability which seem to be feature and methylation state dependent.

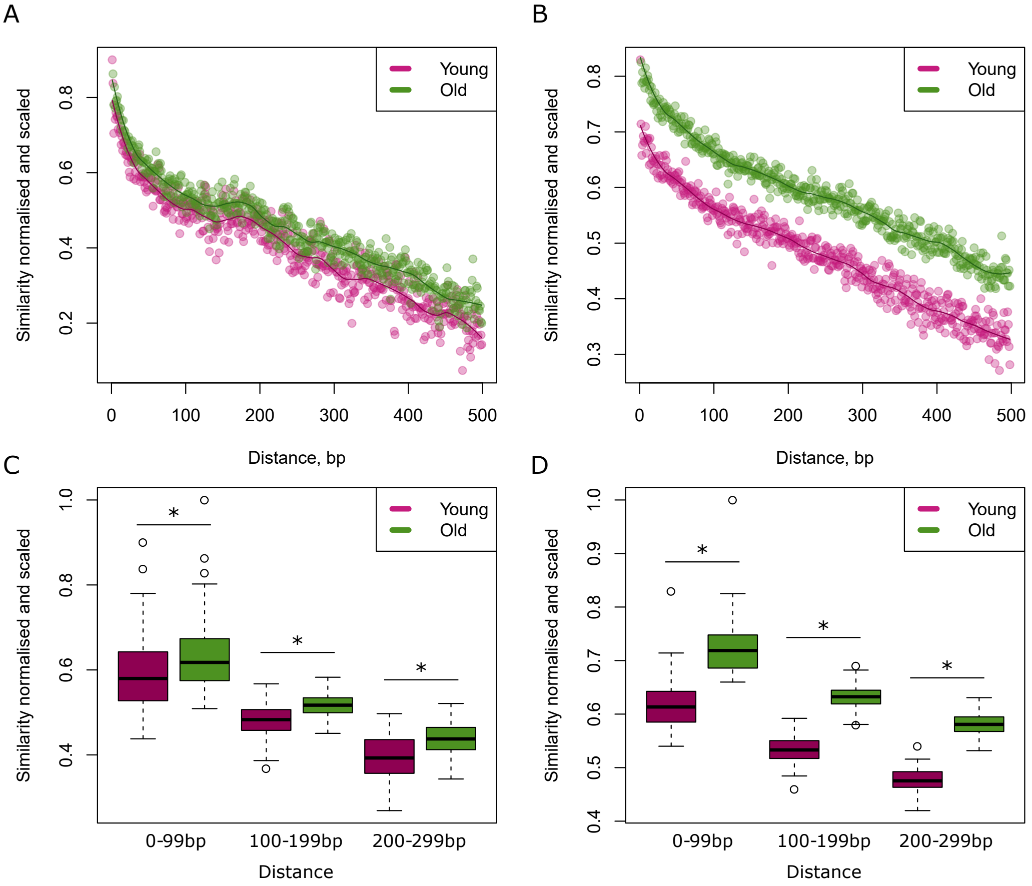


Figure 4.28: Neighbourhood similarity in clock regions: (A and B) Scatterplots showing similarity score as a function of distance (bp) for methylated (A) and unmethylated (B) reference cytosines within 5 kb of the mouse epigenetic predictor sites. (C and D) Boxplots of similarity score for windowed distances of 100 bp within 5 kb of the mouse epigenetic predictor sites. Methylated reference similarities are shown in C and unmethylated reference similarities in D. Asterisks reflect significant differences between young and old (Mann-Whitney Test; Bonferroni corrected p-value <0.05).

4.2.12 Association of transcriptome and methylome

Having observed that with age there is an increase in transcriptional variability alongside an increase in DNA methylation variability, I wanted to assess whether the two types of variability were correlated with one another. This was of particular interest owing to previous work conducted in mESCs where it has been shown that when methylation levels are low, for instance in 2i/LIF-cultured mESCs there is no or little correlation between methylation and transcription (Angermueller et al., 2016). This was hypothesized at the time to be due either to a technical lack of cell numbers or due to the early developmental stage that these cells are emulating, and not due to the low levels of DNA methylation *per se*. Until now, there has not be an alternate cell type that has been assayed at the single cell level using the combined method that also exhibited decidedly low levels of DNA methylation.

I derived correlations between DNA methylation and transcription for all available comparisons using a similar method to the previously published work in mESCs (Angermueller et al., 2016). The number of available comparisons is slightly more limited than the number of genes and promoters covered individually, owing to the criteria of requiring coverage of the methylome and the transcriptome from within the same cell (Materials and Methods section 2.4.16). I calculated Pearson and Spearman correlations assessing associations between expression of a given transcript and promoter or gene body methylation of said transcript.

From this analysis, I observed that in our TA-Hi MuSCs there are very few significantly-associated correlations that are either negative or positive (Figure 4.29). In fact, I observed that methylation and transcription were not significantly correlated for any of the genes analyzed when Spearman correlation was performed (Figure 4.30). In the case of Pearson correlations, there were a small number of significantly-associated correlations (Figure 4.29). Interestingly, there was a positive skew in the distribution of correlation values. In other words, with increasing promoter DNA methylation, there was a concomitant increase in expression. This is counter to what was expected from analyses in bulk settings and also counter to what has been observed previously for

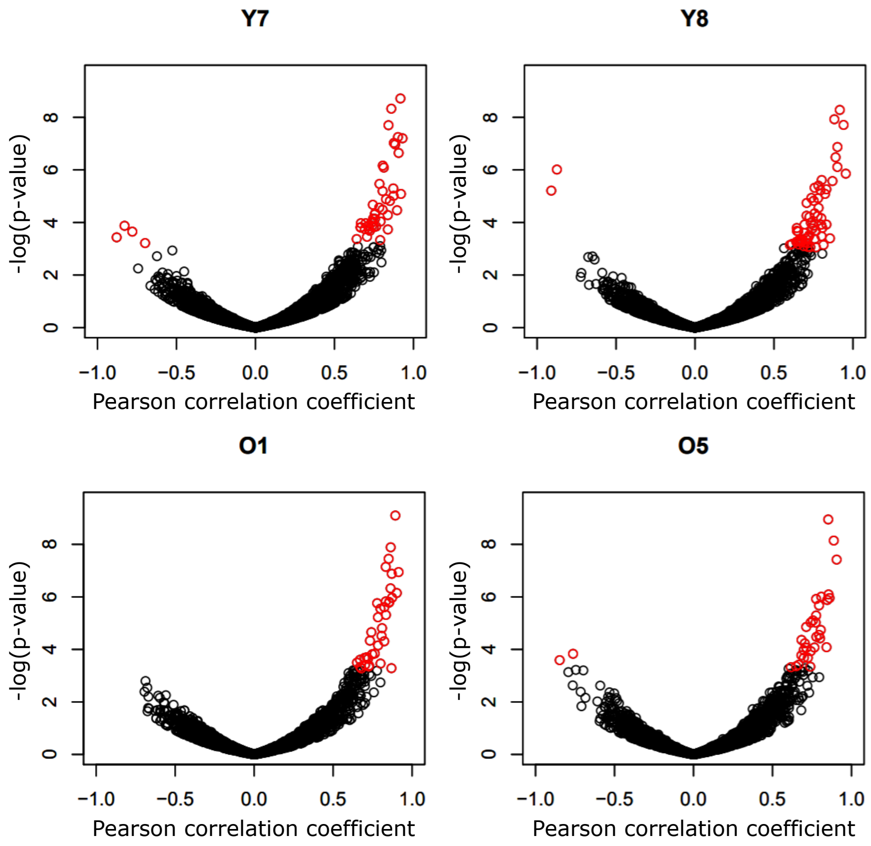


Figure 4.29: Pearson correlations of promoter methylation with transcription: Volcano plots of correlation coefficient against -log(p-value) for Pearson correlations between promoter methylation levels and transcription. Red points are multiple-testing significant using a false discovery rate (FDR) of 10%. There is a volcano plot for each individual. Young individuals are depicted with a “Y” and old individuals with an “O”.

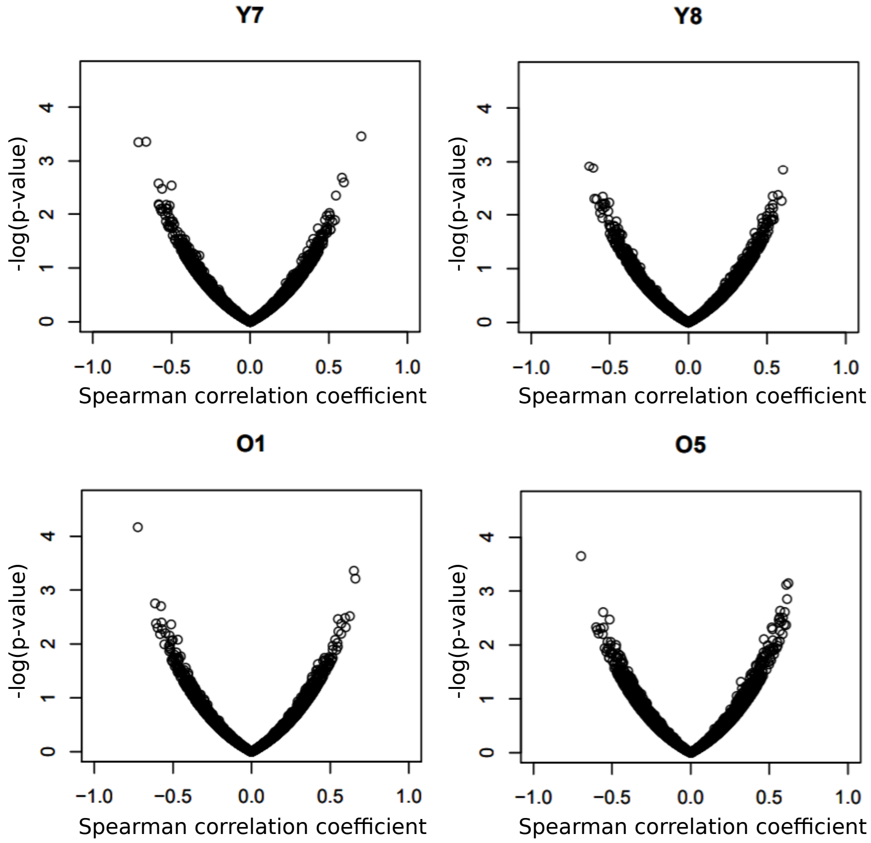


Figure 4.30: Spearman correlations of promoter methylation with transcription: Volcano plots of correlation coefficient against -log(p-value) for Spearman correlations between promoter methylation levels and transcription. There were no significant correlations found using a FDR of 10%. There is a volcano plot for each individual. Young individuals are depicted with a "Y" and old individuals with an "O".

serum-grown mESCs from a single cell combined dataset (Angermueller et al., 2016). Interestingly, this positive skew is exactly what is seen for 2i/LIF-cultured mESCs. This suggests that perhaps this skew in the distribution is not so much of a developmental stage phenomenon, or even a cell number phenomenon as was previously thought but instead relates to the actual low levels of DNA methylation. For instance, it might perhaps be the case that at low levels, DNA methylation is unable to act in its usual repressive fashion and perhaps its presence more closely resembles gene body methylation.

Having assessed promoter methylation correlations with transcription, I next wanted to investigate how gene body methylation correlated with transcription. From this analysis, I found that there were no significant correlations for the Spearman analysis and there was a positive skew in correlation in the case of Pearson correlation, similar to what I saw for promoter methylation (data not shown). This was not too surprising, as it has been shown numerous times that gene body methylation is correlated with active transcription (Kulis et al., 2012, Maunakea et al., 2010 and Varley et al., 2013).

4.3 Discussion

I have generated the first *in vivo* single cell combined dataset allowing the process of ageing to be assessed in unprecedented detail. This dataset was derived from a homogeneous population of TA-Hi MuSCs. The reason for choosing a homogeneous population of cells was to remove confounding variables in my analysis of changes occurring with age, enabling the focus to be solely on cell intrinsic changes.

From the transcriptome, I found that with age there was an increase in variability. I found that this increase in variability was not restricted to a subset of genes but could be found in hundreds of genes. I identified that in young individuals the highly variable genes were associated with GO processes such as “response to cytokine stimulus” and “chemokine receptor binding”. In contrast, I found that highly variable genes in old cells were associated with GO processes such as “chemotaxis” and “cell chemotaxis”. This change in processes that are being defined as highly variable could highlight the increased proportion of cells in old individuals that are now being relied upon to maintain homeostasis of the muscle. Alternatively, it could suggest that these old cells are more likely to become mobile or activated. In addition, I found that there was an enrichment in one of the old individuals for processes associated with the neural lineage-specific transcription factor NeuroD, this is interesting because NeuroD is another basic-helix-loop-helix protein with an incredibly similar motif (position weight matrix (PWM)) to MyoD, the transcription factor responsible for downstream differentiation of muscle satellite cells into myoblasts (Fong et al., 2012). MyoD and NeuroD are known to be regulated in their functionality through the restriction of their epitopes in lineages that they should not be functioning on (Fong et al., 2012). As such, this suggests that there is a loss of regulation of the muscle lineage in old TA-Hi MuSCs with age, and that potentially this could be interfering with the ability of these cells to function properly upon activation. In addition, I assessed cell-to-cell variability in young and old TA-Hi MuSCs, finding that with age there was an increase in cell-to-cell transcriptional variability. Curiously, I saw that this variability was over and above what is seen for the early embryo dataset I compared it with (Mohammed et al., 2017). This is interesting, because it is commonly thought that with a reduction in differentiation potential there would be a concomitant reduction in cell-to-cell variability. This suggests that although TA-Hi MuSCs are an incredibly homogeneous population of cells, and have a very well defined differentiation potential, when similarity in variable genes is assessed across cells, there is less concerted expression than for embryonic stem cells. This could be perhaps due to technical considerations owing to the quiescent nature of these cells, and low levels of transcription, resulting in apparently less coordinated transcriptional regulation owing to the drop-out rate in the single cell transcriptomic method. However, this seems unlikely, since I filter out lowly expressed transcripts (Materials and Methods section 2.4.9). Alternatively, it could be due to the fact that in embryonic stem cells the cell-to-cell variability in variable genes is more coordinated and regulated in perhaps some sort of cyclical process, that is perhaps important for ensuring lineage proportions are defined properly. It will be interesting to assess the levels of cell-to-cell variability across cells as more cell types become available for interrogation.

I defined the first ever TA-Hi MuSC whole genome methylome data. From this methylome data I found that TA-Hi MuSCs had an unusual methylation composition. This was associated with a globally lower level of methylation than is commonly seen for somatic cells. Interestingly, I found that this global reduction in methylation was not genome wide, as is seen in the case of for instance 2i/LIF-cultured mESCs but was in fact localized to specific features. These features included genic regions, for instance exons and introns. This pattern of methylation is interesting, because it is very rare to find a cell type that contains highly methylated repetitive regions of the genome while simultaneously having lowly methylated exons and introns. As such, it could be an incredibly useful model system for future studies interested in understanding the placement of DNA methylation. It will also be fascinating to see whether other highly quiescent cellular populations exhibit similar patterns of methylation deposition. This would be interesting, because one hypothesis for the lack of methylation found in gene -bodies would be the low levels of transcription in these cells. I found that with age there was an increase in methylation across a number of features including repetitive elements such as LINEs. In addition, I defined novel computational approaches for the robust assessment of DNA methylation variability between cells and the patterning of methylation within cells. These methods enabled me to identify regions of the genome that are less variable than would be expected by chance, such as in CGIs that are 10-30% methylated. In addition, these approaches allowed changes in methylation variability with age to be assessed, in addition to whether they are independent of changes in DNA methylation levels or not. I found that with age there was a tendency for regions of the genome that were lowly methylated to gain methylation and with it heterogeneity, and for regions of the genome that were highly methylated to further gain methylation and lose heterogeneity. In addition, I found that there were certain regions of the genome such as CGIs, that gained heterogeneity, with a minimal change in methylation level. From my patterning analysis, I found that with age there was an increase in the predictive power of neighboring positions in the genome, suggesting a reduction in complexity of the methylome in exons and CGIs amongst other features.

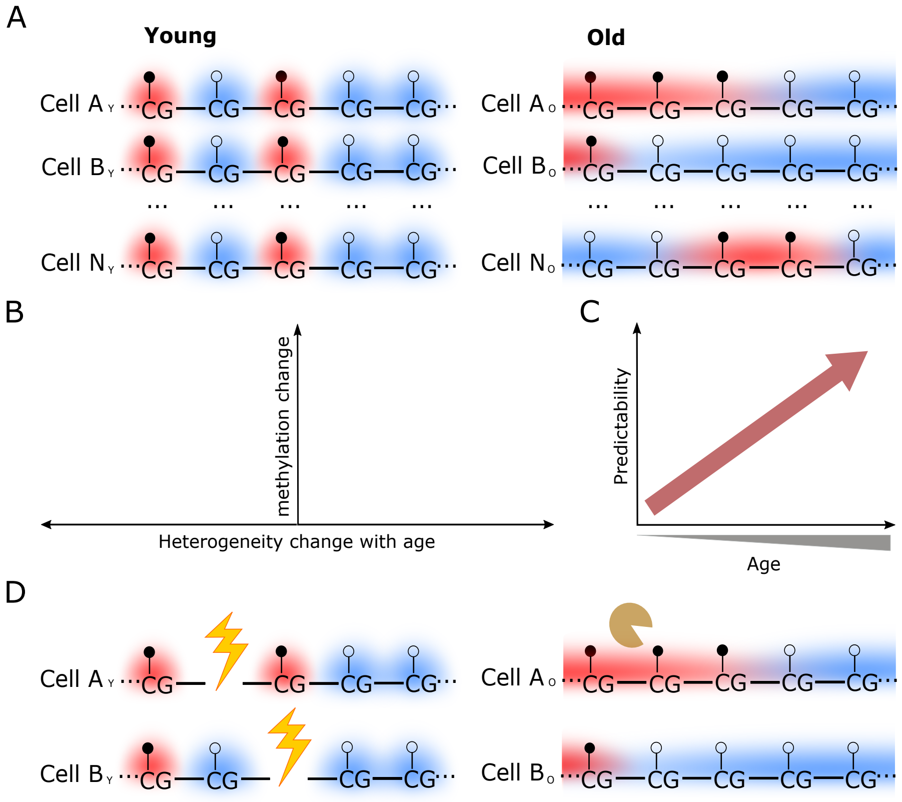


Figure 4.31: Model of epigenetic changes in TA Hi MuSCs with age: (A) Depiction of DNA methylation heterogeneity in young and old cells. The reason shown was lowly methylated in young, hence gain in heterogeneity. Methylated cytosines are shown using filled lollipops and red shading. Unmethylated cytosines are shown using empty lollipops and blue shading. (B) Visualisation of methylation changes with age and the implication for heterogeneity. Lowly methylated regions in the young will tend to gain heterogeneity and highly methylated regions will tend to lose heterogeneity. (C) Visualisation of the incrceasing similarity between neighbouring cytosine positions in the genome with age. (D) Model of how increasing similarity could result in increased variability. Lightning bolts represent DNA damage and the pacman represents a DNA methyltransferase.

This is exciting because it hints at a potential mechanism for the age-associated changes in DNA methylation that are perhaps paradoxically often associated with gains in entropy. The mechanism that I would like to propose for this is that random DNA damage results in DNA repair, that erases DNA methylation from the genome. At the same time, DNA methylation is being added to the genome by enzymes that act in a biased or more processive manner, resulting in cell intrinsic homogeneity or loss of switching. This I suggest is partially CG-density dependent, suggesting the importance of processive enzymatic processes in this enhancement of neighborhood similarity. Owing to this process being, to a certain extent random, when this process occurs across many cells in concert it results in changes that could result in gains or losses in heterogeneity. Interestingly, one would hypothesise that losses in heterogeneity would be associated with regions of the genome that were more prone to DNA damage than the genome as a whole, resulting in more homogeneous methylation patterning across cells. It will be exciting to test this hypothesis further with DNA damage induction and in other cell types, to assess the generalizability of this phenomenon.

When comparing variability in transcription and methylation, I found very little correlation between the two. This suggests that the sources of variability may be different. In the future, to improve such analysis, it would be interesting to assess this relationship in an allele specific fashion, rather than making the assumption of a 1n genome. In this present study, this was not performed due to the cross being derived from one heterozygous (containing the pax7-GFP reporter) and one homozygous parent. This was done to ensure that I could define our homogeneous cell population accurately. It would also be exciting to assess whether this coupling between transcription and DNA methylation increases upon differentiation of these cells.

In summary, I have found that increases in variability that are seen with age can be the result of cell intrinsic gains in homogeneity. This is an exciting proposition that will be possible to test in the future.

Chapter 5

Live cell imaging of DNA methylation in single cells

5.1 Introduction

DNA methylation dynamics are hugely important in a number of settings, including but not limited to development, cancer and ageing. In addition, with the advent of single cell methods to study DNA methylation, it is now becoming increasingly apparent that heterogeneity between cells within a given population is also of great importance in our understanding of these processes and others.

At present, whole genome bisulfite sequencing-based approaches are the gold standard in studies on DNA methylation, since they allow interrogation of this mark at single base-resolution. Although the cost of sequencing is reducing at pace, these sequencing-based approaches are still expensive owing to the size of the genome under study (Lander et al., 2001, Sierro et al., 2014 and Waterston and Pachter, 2002). This is particularly true for DNA methylation, where one is studying a continuum of values between a fully methylated site and a fully unmethylated state. This is in contrast to most genomics studies where typically there are three states to consider (homozygous reference, heterozygous, and homozygous alternate). This necessitates many multiple copies of the same position in the genome being sequenced in order to observe small yet meaningful changes in DNA methylation. This is especially the case for single cell epigenetic approaches that typically require as few as 20 cells to be sequenced on a single lane (Angermueller et al., 2016). As such, for larger scale studies of epigenetic heterogeneity, additional methodologies are required to sample a given population of cells. These cost issues are trying to be dealt with presently. For instance, there are now pooling-followed-by-regional-pull-down-based approaches that are attempting to enrich for genomic regions of interest while reducing the cost per cell (Mulqueen et al., 2017). These approaches, however, are still expensive/ineffective owing to them still being in their infancy and at present are little better than conventional single cell whole genome bisulfite sequencing (Angermueller et al., 2016, Clark et al., 2017 and Smallwood et al., 2014). Unfortunately, none of these sequencing methods are currently applicable to the direct observation of DNA methylation dynamics. This has meant that at present it has only ever been possible to infer dynamic behaviour from static snapshots. This remains a large technical limitation.

These limitations are not sequencing-specific, and apply also to all other methods currently available for the study of DNA methylation, including, for instance, mass spectrometry and immunofluorescence (IF) measurements. The destructive nature of the sample preparation makes these techniques incompatible with continuous observation of the same cell or cells over multiple time points.

A few studies have been conducted to qualitatively assess the dynamics of the distribution of DNA methylation in living cells (Ingouff et al., 2017, Kimura et al., 2010,Ueda et al., 2014, Yamagata, 2010, Yamazaki et al., 2007 and Zhang et al., 2017). These studies have utilised various methyl-binding domain proteins as a proxy to visualise DNA methylation dynamics. These studies have shown that it is possible to visualise DNA methylation in living cells in several different contexts, such as DNA methylation in CpG and CHH contexts (Ingouff et al., 2017). At present, these papers have solely assessed the distribution and nature of DNA methylation in a qualitative fashion. For instance, assessing how the dynamics of the distribution of DNA methylation changes during early mouse development (Yamazaki et al., 2007). However, these studies have not attempted to utilise the fluorophore fused to the MBD to derive a quantitative assay of global DNA methylation dynamics.

The aim of this work was to develop a system for real-time quantitative measurement of DNA methylation in live single-cells. To achieve this, I utilised a previously defined and validated MBD-NLS-eGFP (hereto referred to as MBD1-eGFP) construct (Yamagata, 2010). In achieving this, I wanted to develop a method that could be extended to assess the dynamics of large populations of cells cheaply and effectively and would enable epigenomics to be studied at the single cell level at a far greater throughput than is currently attainable.

5.2 Results

5.2.1 Method design

The most commonly used protein for visualising CpG methylation in the field has been the methyl binding domain (MBD) of methyl binding domain 1 protein (MBD1). MBD1 is a member of a family of proteins that bind to single symmetrically methylated CpG dinucleotides. All MBD family members share a common MBD consensus sequence (Zou et al., 2011; Figure 5.1).

The MBD of MBD1 was chosen for live cell imaging applications over the other family members due to it previously being observed to have the highest sensitivity to methylation changes (Baubec et al., 2013, Ohki et al., 2001 and Zou et al., 2011). This MBD domain is known to have an approximately 10-fold increased affinity for CpG sites that are fully methylated than those that are unmethylated (Baubec et al., 2013, Ohki et al., 2001 and Zou et al., 2011). For visualisation, this protein domain is fused at its C-terminus to a fluorescent protein such as green/red fluorescent protein (GFP/RFP) spaced by a nuclear localisation signal (NLS) to ensure correct folding, flexibility of the MBD itself and efficient nuclear localisation of the folded reporter protein. The potential utility of the MBD of MBD1 for the study of DNA methylation has now been validated using several approaches. Briefly, pull-down experiments for DNA methylation and RFP (linked to the MBD) have been used to show that both the methyl mark

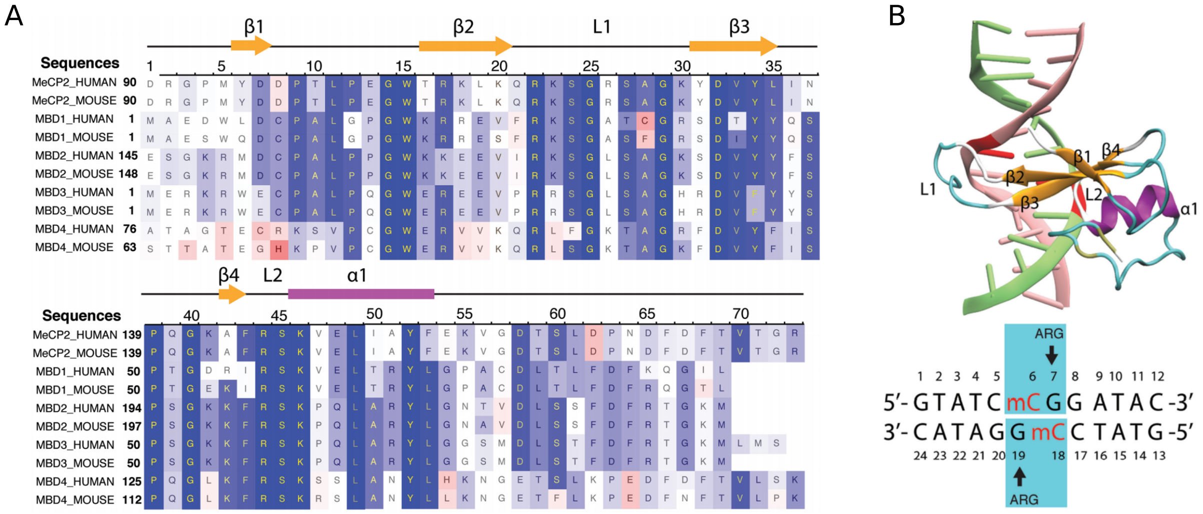


Figure 5.1: MBD alignment across MBD-containing family members: (A) Alignment of the MBD domain for both human and mouse is displayed. Secondary stucture is displayed above the alignment. Sequence conservation is shown on a blue, white, red scale; where dark blue is highly conserved. (B) The solution nuclear magnetic resonance (NMR) structure of the MBD of human MBD1 is shown. In the structure the domain is bound to a CpG methylated dsDNA. Figure adapted from (Zou et al., 2011)

and the RFP-linked MBD occupy similar regions of the genome (correlation coefficient of >0.597 for regions larger than 5kB; Ueda et al., 2014). In addition, comparison of the distribution and localisation of the MBD domain with 5'methyl-cytosine has been conducted using IF: both are found to stain heterochromatic regions of the genome containing predominantly major and minor satellites (Ueda et al., 2014). Excitingly, recent studies have also shown that it is possible to incorporate these reporters into live animals, enabling the assaying of DNA methylation distribution from the early embryo to the adult in both mice and zebrafish (Zhang et al., 2017).

To ensure the widest applicability of a quantitative live single cell approach it needed to be minimally damaging to the cell to enable as many time points to be assayed as possible. In addition, it needed to allow time points to be collected from single cells across a broad range of intervals, from seconds to weeks so as to maximise its utility. Lastly, I required that the method utilised no more than one fluorophore, so as to minimise any bias associated with cooperative effects that would result from the necessity of proximity within the nucleus (as seen for fluorescence resonance energy transfer (FRET)) or derived from the DNA sequence itself (as seen for cooperative antibody binding). The method that I decided to implement is called Differential

Dynamic Microscopy (DDM). DDM is an image analysis method that uses Fourier transform techniques to determine ensemble dynamics of isotropic processes, a more detailed description of the approach is provided in Materials and Methods section 2.5.13 and Results section 5.2.11. This method has been previously used to study the motion of particles in solutions (Cerbino and Trappe, 2008). In addition, this approach has also been used to assess the dynamics of ensemble bacterial motion, using a modified version of the method known as confocal-DDM (Lu et al., 2012 and Wilson et al., 2011). However, for this current study, there are two hurdles that will need to be overcome before this method can be made applicable. Namely, that DDM has not been applied in an intracellular manner, nor has it been applied in a fluorescence manner to assess a biological system.

In addition, to the development of this DDM-approach, I developed an alternative fluorescence recovery after photo-bleaching (FRAP) approach. FRAP is a fluorescence-based imaging method that determines diffusion kinetics from the observation of the recovery rate of a sample region following photo-bleaching, a more detailed description of the approach is provided in Materials and Methods section 2.5.11 and Results section 5.2.3. This FRAP approach was used to determine the feasibility of the novel DDM approach for quantifying DNA methylation dynamics. In addition, since FRAP is itself, an inherently live cell measure of DNA methylation, it represents an interesting alternative approach to explore for certain applications. However, there are several drawbacks to FRAP reducing the potential utility of this approach over DDM for future applications. For instance, the destructive nature of the imaging process limits the number of time points that could be captured for any single cell, and results in increased confounding factors such as cell damage and death. In addition, the time resolution of FRAP is limited by the time required to achieve recovery following photo-bleaching. This in turn limits the minimal time interval to observe DNA methylation dynamics to minutes.

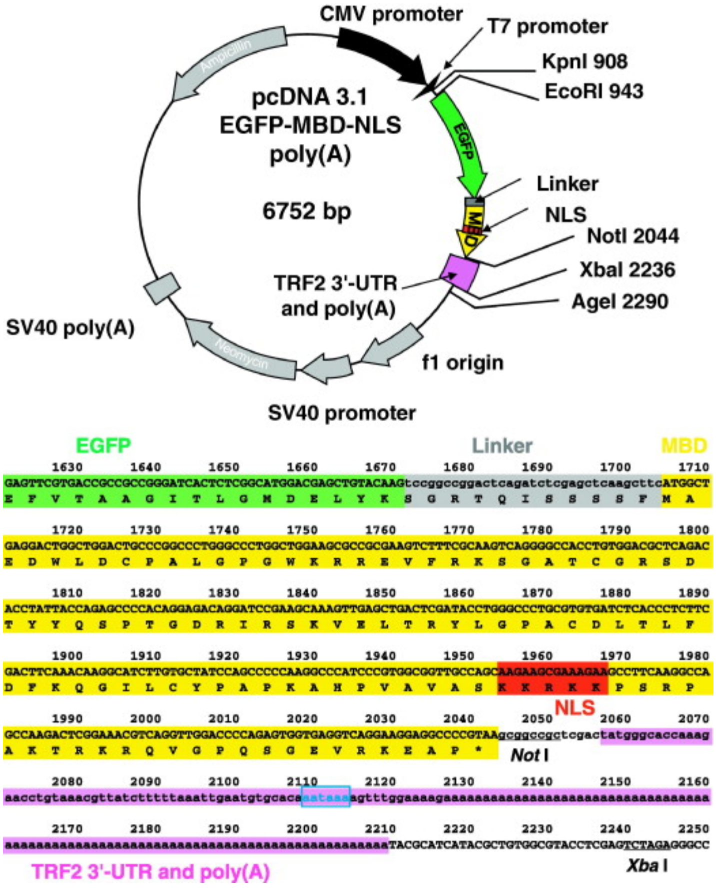

CMV promoter
T7 promoter
KpnI 908
EcoRI 943
Ampicillin
pcDNA 3.1
EGFP-MBD-NLS
poly(A)
6752 bp
EGFP
Linker
NLS
MBD
NotI 2044
XbaI 2236
AgeI 2290
TRF2 3'-UTR
and poly(A)
SV40 poly(A)
Neomycin
f1 origin
SV40 promoter
EGFP
Linker
MBD
1630 1640 1650 1660 1670 1680 1690 1700 1710
GAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGtccggccggactcagatctcgagctcaagcttcATGGCT
E F V T A A G I T L G M D E L Y K S G R T Q I S S S S F M A
1720 1730 1740 1750 1760 1770 1780 1790 1800
GAGGACTGGCTGGACTGCCCGGCCCTGGGCCCTGGCTGGAAGCGCCGCGAAGTCTTTCGCAAGTCAGGGGCCACCTGTGGACGCTCAGAC
E D W L D C P A L G P G W K R R E V F R K S G A T C G R S D
1810 1820 1830 1840 1850 1860 1870 1880 1890
ACCTATTACCAGAGCCCCACAGGAGACAGGATCCGAAGCAAAGTTGAGCTGACTCGATACCTGGGCCCTGCGTGTGATCTCACCCTCTTC
T Y Y Q S P T G D R I R S K V E L T R Y L G P A C D L T L F
1900 1910 1920 1930 1940 1950 1960 1970 1980
GACTTCAAACAAGGCATCTTGTGCTATCCAGCCCCCAAGGCCCATCCCGTGGCGGTTGCCAGCAAGAAGCGAAAGAAGCCTTCAAGGCCA
D F K Q G I L C Y P A P K A H P V A V A S K K R K K P S R P
NLS
1990 2000 2010 2020 2030 2040 2050 2060 2070
GCCAAGACTCGGAAACGTCAGGTTGGACCCCAGAGTGGTGAGGTCAGGAAGGAGGCCCCGTAAgcggccgctcgactatgggcaccaaag
A K T R K R Q V G P Q S G E V R K E A P *
Not I
2080 2090 2100 2110 2120 2130 2140 2150 2160
aacctgtaaacgttatctttttaaattgaatgtgcacaaataaaagtttggaaaagaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
2170 2180 2190 2200 2210 2220 2230 2240 2250
aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaTACGCATCATACGCTGTGGCGTACCTCGAGTCTAGAGGGCC
TRF2 3'-UTR and poly(A)
Xba I

Figure 5.2: Construct map of the MBD1-eGFP construct: A schematic of the plasmid is shown. Highlighted are integral plasmid features. Underneath the plasmid schematic, a sequencing-based annotation for the MBD domain is displayed. This sequence annotation highlights the location of the eGFP and NLS with respect to the MBD domain. Figure adapted from (Yamagata, 2010).

5.2.2 MBD1-eGFP localisation recapitulates 5mC staining

Before attempting to quantify DNA methylation in living cells, I first wanted to validate that I could reproduce the expected patterns characteristic of DNA methylation in the nucleus (Yamagata, 2010). A graphical representation of the construct that was utilised in my experiments is shown in Figure 5.2.

To assess the distribution of the protein derived from the MBD1-eGFP construct, I performed transfections in serum-grown mouse embryonic stem cells (mESCs), before fixing the cells and imaging them on a confocal microscope. These images, shown in Figure 5.3 highlight that the same pattern of localisation can be achieved for the construct as has been previously defined for DNA methylation in a CpG context (Ueda

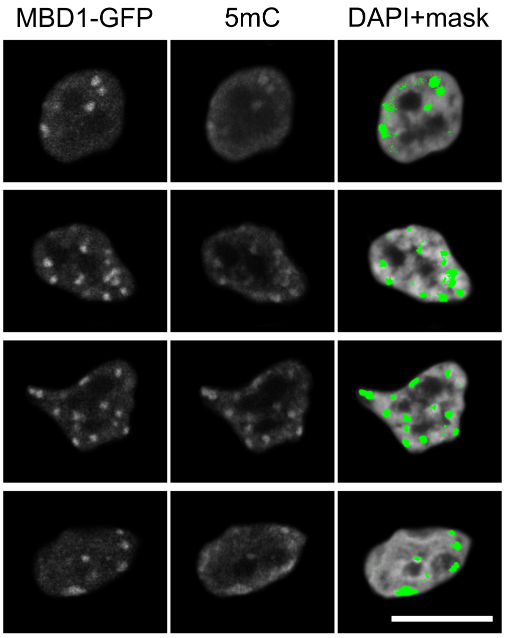


Figure 5.3: Co−localisation of DNA methylation and MBD1-eGFP protein in serum mESCs: Exemplar fixed serum mESCs are shown. The first column displays the localisation of the MBD1-eGFP protein as measured from the eGFP. The second column displays the 5'-methylcytosine for the same cells, as measured by IF. The third column displays the DAPI stain of the nucleus overlayed by a colocalisation mask of the MBD1-eGFP and 5'-methylcytosine. The scale bar represents 10 μm.

et al., 2014). Namely, that the construct is visibly associating with heterochromatic foci within these mESCs.

5.2.3 Description of the FRAP method used

Having seen that the construct is behaving as was expected from the previous publications that have utilised it, I next decided to perform FRAP experiments to determine the feasibility of using this construct for the quantification of global DNA methylation levels in living cells.

The concept of FRAP is to identify a small region of interest containing a number of fluorophores that are in an equilibrium state and using either laser excitation or LED excitation to photo-excite all the fluorophores in the region of interest until they are no longer able to emit de-excitation photons, this area is referred to as being

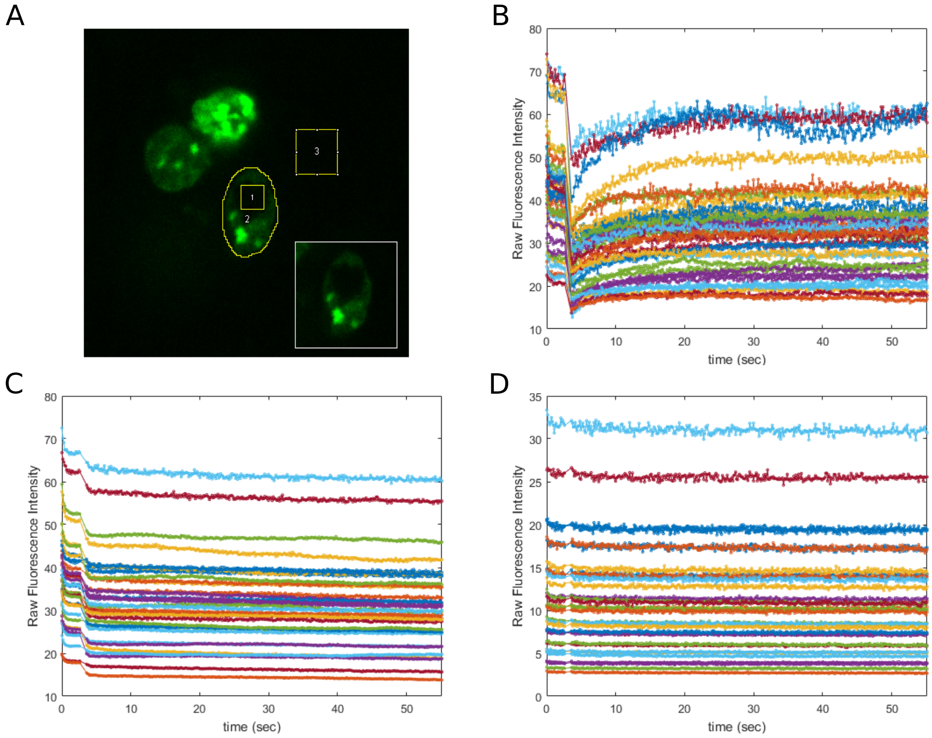


Figure 5.4: FRAP recoveries explained: (A) An example FRAP image. Regions of interest (ROI) are highlighted. ROI1 is the ROI in the nucleus that is photo-bleached. ROI2 contains the nucleus of the cell from which measurments are being taken. ROI3 is a background region of the image that does not contain a cell. (B-D) Example raw recovery curves (pre-normalisation) for ROI1 (B), ROI2 (C) and ROI3 (D).

photo-bleached. For a static system e.g. a fixed cell, the area that is photo-bleached will never recover to become fluorescent again simply because the fluorophores are immobile. Likewise, in a diffusive system e.g. a fluorescent dye suspended in water the region of interest will eventually recover to the original level of fluorescence after a period of time (Figure 5.4).

The time for full recovery is dependent on many factors including: the size of the photo-bleached region and the size/hydrodynamic radius of the fluorophore under study. Lastly, it should be taken into consideration that in non-diffusion mediated systems e.g. driven/trafficked motion, such as is found for lipid rafts (Day et al., 2012), such recovery would be dependent on the efficiency and type of mediated motion that is involved.

The resolution limits of a FRAP system are measurements of the intensity levels of the fluorophore and that the recovery time of the photo-bleached area is much greater than integration time of the instrument, particularly the initial frame subsequent to photo-bleaching. A general FRAP recovery curve will be able to be fit to the form (Klein and Waharte, 2010).

$$f(t) = A(1 - e^{(-t \cdot \tau)}) \tag{5.1}$$

where A is the amplitude of the signal, t is the time of measurement and is the recovery rate. The recovery rate can be converted into a recovery half-life with,

$$t_{1/2} = \frac{ln(0.5)}{\tau} \tag{5.2}$$

This recovery half-life ($t_{1/2}$) was used to compare across conditions and experiments. The recovery half-life in the region of interest, w^2, is also related to the diffusion of the fluorophore, assuming the photobleached area is spherical and gaussian,

$$D = \frac{w^2}{4t_{1/2}} \tag{5.3}$$

The amplitude A in equation 1 can be determined as the long-time asymptote value and has no bearing on the argument of either equation.

$$\begin{aligned} f(t) \lim_{(t \to \infty)} &= A(1 - e^{-(\infty \cdot \tau)}) \\ &= A(1 - e^{-\infty}) \\ &= A(1 - 0) \\ &= A \end{aligned} \tag{5.4}$$

This is because the amplitude of the FRAP curve can be normalized, leaving only the argument of the equation as important for determining the diffusion or type of anomalous diffusion that is present in the experiment.

The measurements in this particular FRAP setup were made using a square photo-bleached area. For mathematical ease of deriving the recovery dynamics, the square bleached area can be approximated as a circle of radius half the side of the square.

$$\begin{aligned} ratio &= \frac{\pi r^2}{2r^2} \\ &= \frac{\pi}{4} \\ &= 0.785 \end{aligned} \tag{5.5}$$

The area difference will result in a systematic bias that at most will correspond to 10-12% measurement error. The rationale for approximating the area as a circle is due to the derivation of the mathematics used to establish equation 5.1.

In these experiments a full normalisation was performed, briefly this involved the scaling of the fluorescence at t=0 to 0 and the fluorescence upon full recovery to 1. Mathematically, this is represented by equation 5.6.

$$Full\ Normalisation = \frac{f(t) - f(t_{21})}{f(t_{asym}) - f(t_{21})} \tag{5.6}$$

Where:

$$f(t_{21}) = Fluorescence\ upon\ bleaching$$

This normalization results in an amplitude term that in some cases would be not equal to unity due to the fluctuations on the source data, however it can subsequently be used as a filter on the quality of the data sets. If a particular fit did not return a value of A ∼1 then these data could be excluded from further analysis as they are not representative of a standard FRAP curve due to noise levels.

All samples were normalised and fit using equation 5.1 to obtain measures of recovery half-life.

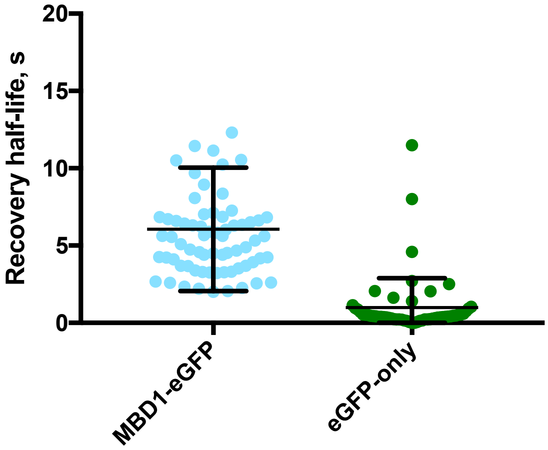


Figure 5.5: MBD1-eGFP has a longer recovery half-life than eGFP-only: Plot showing the recovery half-lives of MBD1-eGFP and eGFP-only serum-grown mESCs. Significance was assessed using a two-tailed *t*-test, the difference between the two conditions was highly significant, p-value <0.001.

5.2.4 MBD1-eGFP diffusion can be quantified in live cells

To assess the feasibility of quantifying global levels of DNA methylation in living cells, I performed some preliminary FRAP experiments using serum-grown mESCs as my model system.

The first experiment that was performed was to compare the MBD1-eGFP construct to an eGFP-only construct. The FRAP imaging and primary analysis was conducted as detailed in the Materials and Methods section 2.5.11. The resulting quantification was then fully normalised before curve fitting of equation 5.1 was performed using the *nls* function in R (R Core Team, 2017; Section 5.2.3).

The results from this experiment in serum-grown mESCs showed that as expected the recovery half-lives for the MBD1-eGFP mESCs were significantly increased relative to the eGFP-only construct (Figure 5.5). This showed for the first time that it was possible to measure the relative difference in recovery half-life between a eGFP-only (mean $t_{1/2} = 0.6$ seconds) construct and one that contains an MBD protein (mean $t_{1/2} = 6.2$ seconds). This experiment suggested to me that it may be possible to measure DNA methylation dynamics in living cells, owing to the magnitude of the difference between these two constructs and the precision of the recovery half-life between replicates.

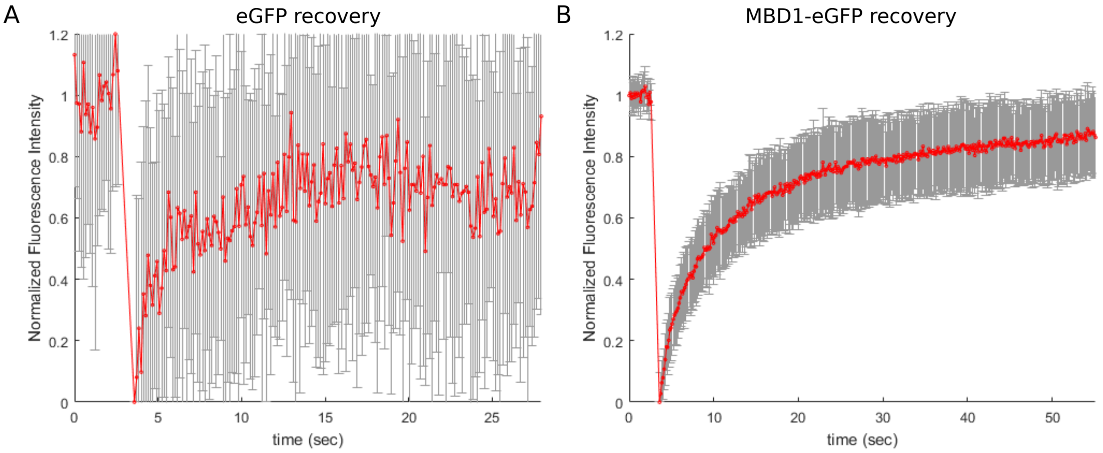


Figure 5.6: Recovery speed of the eGFP-only mESCs precludes accurate measurement: (A and B) Normalised and cell averaged recovery plots for eGFP-only mESCs (A) and MBD1-eGFP mESCs (B). Error bars represent standard deviation in the measurement and far larger in the case of eGFP-only.

Unfortunately, the speed with which the eGFP recovered was too fast to enable accurate measurement of the recovery half-life (Figure 5.6). This led to difficulty in assessing the extent of the photo-bleaching and also greatly reduced the number of time points with which it was feasible to assess the recovery over. This in turn resulted in increased noise within the fitting and in many cases difficulty in converging on a fit. As an alternative control for the eGFP-only cells to ensure that they were actually being photo-bleached to a similar extent as the MBD1-eGFP cells, I fixed cells before performing the FRAP (using the same settings as previously). As expected, in this instance the bleaching was visible and matched in magnitude to the MBD1-eGFP-containing cells (Figure 5.7). As such I concluded that the MBD1-eGFP protein is at least 10-fold slower in serum mESCs than the eGFP-only control, and that this likely reflects the MBD1-eGFP binding to DNA.

5.2.5 Maximising the signal from the FRAP experiment

Having seen that it was possible to measure differences in the recovery half-life of MBD1-eGFP and eGFP in serum mESCs, I next wanted to ensure that for future experiments I was maximising the FRAP signal whilst minimising any experimental confounding factors that I can control. These experiments were conducted in serum-

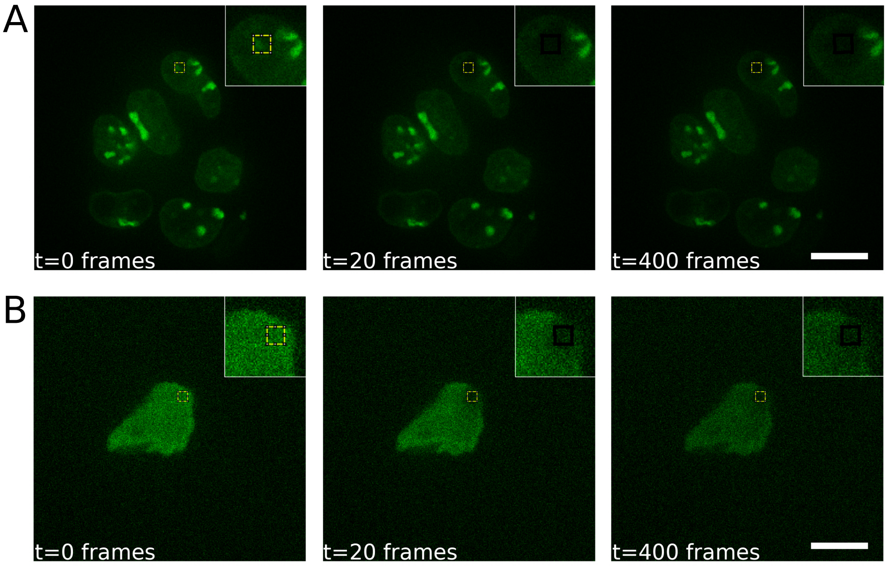


Figure 5.7: eGFP-only serum mESCs were photo-bleached: (A and B) Frames from pre-bleach (t=0 frames), bleach (t=20 frames) and post-bleach (t=400 frames) for MBD1-eGFP serum mESCs (A) and eGFP-only serum mESCs (B). Insets are shown to highlight the region of the nucleus that was photo-bleached. ROI1 is depicted with a yello or black square.

grown mESCs containing the MBD1-eGFP construct. Unfortunately, certain variables although potentially confounding of my results, are beyond the limits of our control, it is not possible for us to achieve a uniform and circular photo-bleach, for instance (Section 5.2.3).

Assessing photo-bleach diameter

The first parameter that I assessed was the diameter of the photo-bleached square. I wanted to ensure that I was maximising the size of the photo-bleached region under study. This is important since the limit of the accuracy in the measurement of the recovery upon photo-bleaching is in part due to the number of pixels that are assessed within the bleached area. As such the more pixels that are visualised within the bleached area the more information there is to average over and determine a recovery curve. There is a linear relationship between the area of the photo-bleached region and the recovery half-life ($t_{1/2}$). In other words, the larger the area of the photo-bleached

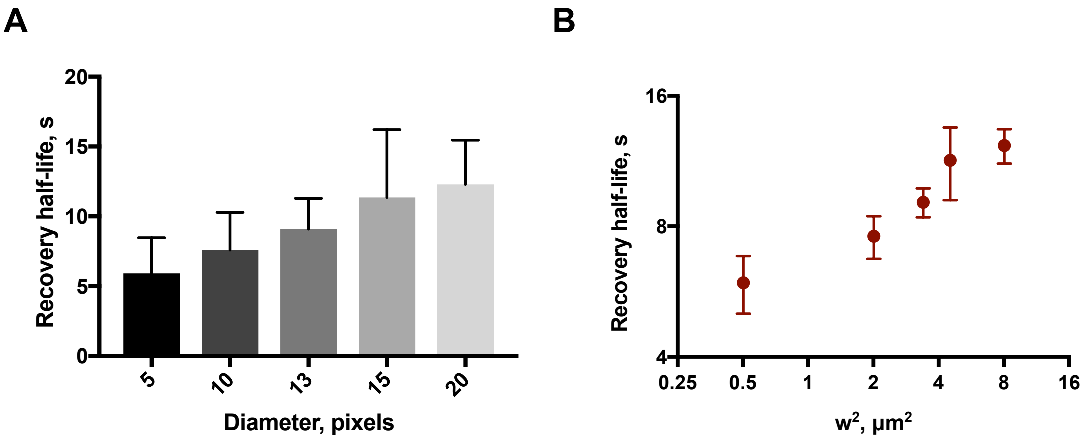


Figure 5.8: Recovery is size invariant over measured bleach diameters: (A) Recovery half-life is displayed as a function of increasing pixel size. Error bars represent standard deviation. (B) Recovery half-life shown as a function of photo-bleaching area. Measurements were made in MBD1-eGFP serum mESCs.

region the longer the recovery takes, since the diffusion coefficient itself should remain constant (equation 5.3). However, there is a limit to this gain in accuracy, and that is due to the constraint of the bounds of the cell. The mathematics for deriving the diffusion coefficient from these FRAP-based measurements, makes the assumption that the space around the photo-bleached area is uniform and infinite. Both of these assumptions are inherently wrong when applied to a cellular context, however, they are true enough that equation 5.1 holds as long as the photo-bleached area is not too large. As such, I decided to perform a control experiment to test the assumption of linearity with increasing bleach area. To assess this, I performed FRAP experiments using different diameters of bleaching area.

This experiment showed that it was possible to derive recovery half-lives across a range of photo-bleached areas (Figure 5.8). In addition, it exemplified the purpose of the experiment, which was to reduce the noise associated with small photo-bleached areas, since at low photo-bleaching areas the standard deviation in the measurement increased. In addition, this experiment showed that the assumption of infinite area outside of the photo-bleached area held for all photo-bleached regions tested (up to 10.24 μm^2). To ensure that I maximised the number of pixels and to reduce the chance of any future issues, where cells with slightly smaller nuclei could result in these boundary effects again, I decided to utilise a photo-bleaching area of 5.76 μm^2 for future experiments.

Assessing the impact of expression differences

The second potentially confounding factor that I wanted to address came from the observation that even if I derived a stable cell line containing the MBD1-eGFP construct, different cells within the population would be expressing differing amounts of the protein (as measured by eGFP levels that are visible from confocal microscopy). One possible explanation for this is that this protein is being expressed in a cell cycle dependent manner. Since I want to be able to quantitatively assess DNA methylation levels and not simply to describe patterns of the modification within cells, I wanted to ensure that the different levels of MBD-eGFP in the cells were not confounding my results. Theoretically, assuming that the substrate that the protein is binding to is not limiting there should be no difference between a cell expressing low or high amounts of the construct in terms of apparent diffusion. Since there are 3x10^9 bp in the genome with approximately 2.5x10^7 CpG dinucleotides in the genome, of which in a serum mESC (assuming 70% CpG methylation) approximately 1.75x10^7 of these CpG dinucleotides will be symmetrically methylated this seemed unlikely to be the case but still warranted testing.

As such I addressed whether there was a relationship between the apparent recovery half-life of any given serum-grown mESC, and the expression level of the MBD1-eGFP protein within a given serum-grown mESC; measured as the average fluorescence intensity of the nucleus.

This experiment showed that there was no correlation between the mean fluorescence intensity of a given cell and the recovery half-life of the FRAP experiment conducted on that cell (Figure 5.9). As such, it was concluded that the expression level of the protein, as far as it was assessed, does not appear to be correlated to the recovery half-life measured. This is perhaps not surprising when the number of CpG dinucleotides in the genome is considered in relation to the likely protein expression level within the cell (Milo, 2013). This result is important because it not only means that the variability in the levels of the construct are not an issue for this approach, but it also means that in the future, these experiments do not require deriving a stable clonal cell line.

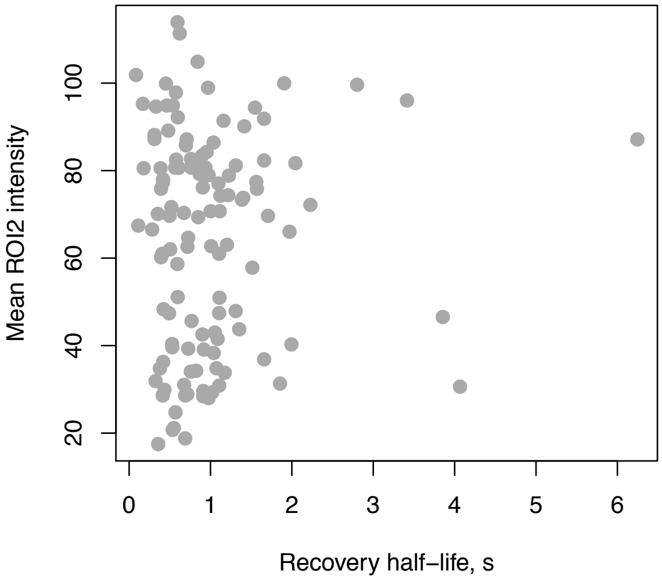


Figure 5.9: Recovery is independent of protein level: Scatterplot comparing mean ROI2 fluorescence intensity and recovery half-life. No significant correlation was found between the two.

5.2.6 Measuring DNA methylation using FRAP

Having determined the optimal photo-bleaching area and having validated that the system functions in a manner that is MBD-eGFP expression independent, I sought to assess whether it would be possible to differentiate between serum-grown mESCs with and without DNA methylation. To conduct this experiment, stable clonal cell lines containing the MBD1-eGFP construct were derived in two genetically-independent DNA methyltransferase triple knock-out (DNMT-TKO) mESC lines and the corresponding control genetically-identical background cell lines (hereto referred to as DNMT-TKO-C). These mESCs were HA36CB1/159-2 (denoted hereafter as 159; Domcke et al., 2015). This comparison was made because DNMT-TKO mESCs are known to have incredibly low/no DNA methylation, whereas in contrast the DNMT-TKO-C mESCs have approximately 75% CpG DNA methylation genome-wide when grown in serum. This meant that I could assess whether it was possible to differentiate between MBD1-eGFP in contexts containing low or high amounts of CpG DNA methylation.

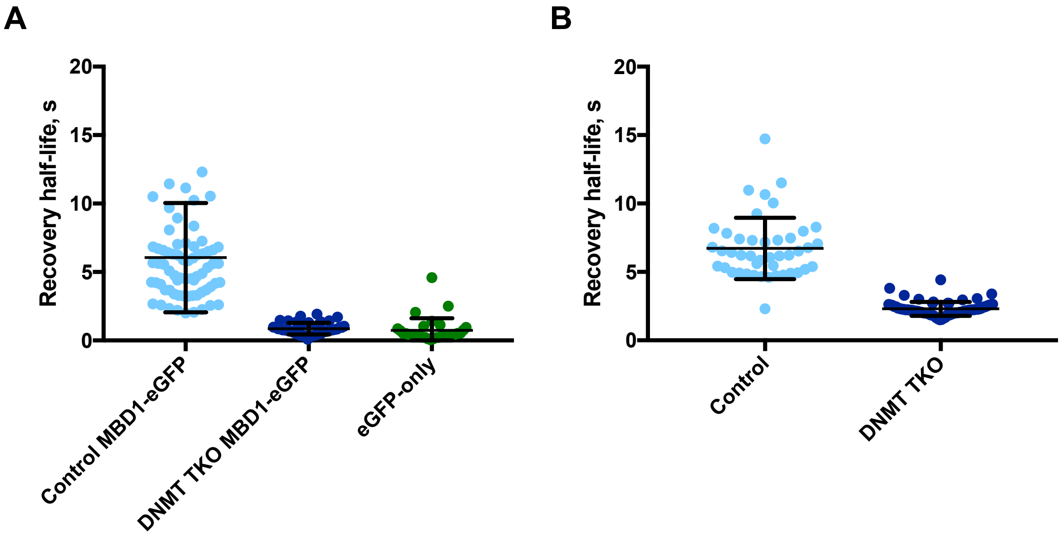


Figure 5.10: Measurement of DNA methylation is possible using FRAP: (A) Comparison of serum DNMT-TKO and control cell lines containing the MBD1-eGFP construct. eGFP-only serum mESCs are shown for comparison. Control MBD1-eGFP is significantly different from both DNMT TKO MBD1-eGFP and eGFP-only (p-value <0.001). DNMT TKO MBD1-eGFP has a significantly slower recovery than eGFP-only (p-value <0.01). (B) Comparison of 159 DNMT TKO serum mESCs and control cell lines containing the MBD1-eGFP construct. The control 159 mESCs are significantly different from the 159 DNMT TKO cells (p-value <0.001).

The FRAP images were processed and primary analysed as in Materials and Methods section 2.5.11. The resulting output was then normalised and fitted using equation 5.1. The results from this experiment showed that the two DNMT-TKO cell lines both had shorter recovery half-lives than the DNMT-TKO-C cell lines (Figure 5.10). This is consistent with them containing less DNA methylation than the control cell lines and hence the MBD1-eGFP construct being less constrained in its motion. In addition, I found that, where compared, the DNMT-TKO cell lines, although faster recovering than the control cell line were still decidedly slower than the eGFP-only serum-grown mESCs. This is consistent with the biochemical observations made previously that this protein is still capable of interacting with unmethylated DNA, albeit with lower binding efficiency than methylated DNA, thus hindering its motion. This experiment, was an important next step in the development of the method because it had previously been shown that, although with lower affinity, the MBD of MBD1 will bind to both CpG contexts in hemi-methylated and unmethylated contexts, alongside binding to DNA in non-CpG contexts altogether (Baubec et al., 2013, Ohki et al., 2001 and Zou et al., 2011). Importantly, the ability to undertake this experiment in two completely independent DNMT-TKO cell lines was important, to ensure that any change I saw was not dependent on the characteristics of any one knock-out. One potential confounder of these experimental systems, is that the shape and structure of the cell/nucleus is very different between serum-grown DNMT-TKO mESCs and serum-grown DNMT-TKO-C mESCs. As such, I could be attributing differences in recovery half-lives to simply differences in cell shape. The next two experimental systems that were studied were conceived to address this question, amongst others.

5.2.7 Measuring different levels of DNA methylation using FRAP

Having seen that it is possible to measure an approximately 70% difference in DNA methylation, I wanted to assess the sensitivity of my FRAP-based quantification of DNA methylation. To assess this, I utilised two experimental systems.

Serum-grown mESC comparison FRAP

The experimental system chosen to assess very subtle changes in DNA methylation was that of serum-grown mESCs derived from different mouse strains. This was chosen since it is known that mESCs from different mouse strains exhibit differences in global levels of CpG methylation in their genomes (<5%). As such, I compared serum-grown 159 and E14 mESCs.

The results of this comparison (Figure 5.11) showed that I could detect population differences between the cells from the different origins that were in keeping with what is known from mass spectrometry and sequencing-based approaches. Unfortunately, it was not possible to isolate the two different populations on a cell-by-cell basis. In other words, it was not possible to predict the population that any one cell came from with any great certainty. This, however, is potentially not a technical limitation of the FRAP methodology but could also be due to the nature of DNA methylation

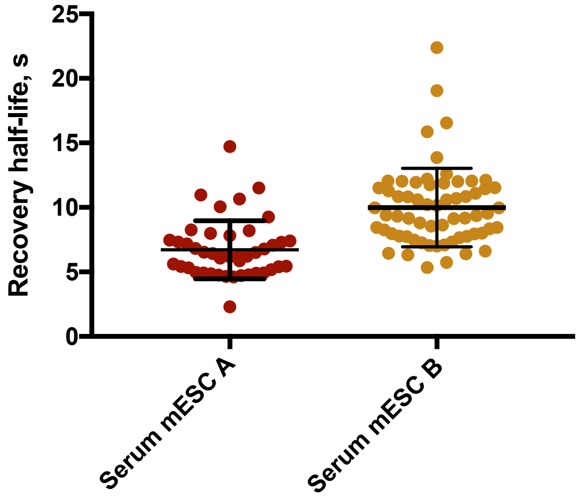


Figure 5.11: Detection of methylation differences between mESC strains: Comparison of serum E14 mESCs containing the MBD1-eGFP construct (serum mESC A) and serum 159 mESCs containing the MBD1-eGFP construct (serum mESC B). A significant difference was observed between these two conditions (p-value<0.01).

heterogeneity itself. It has previously been shown that mESCs exhibit heterogeneous levels of DNA methylation globally, when assessed at the single cell level using the gold standard bisulfite sequencing-based approach (Angermueller et al., 2016, Clark et al., 2017 and Smallwood et al., 2014). I conclude therefore that this FRAP-based method is able to determine very subtle differences in terms of global levels of DNA methylation dynamics, however, it is difficult to determine how much of the differences in recovery half-life measurements are down to technical or biological noise. This is not a simple question to determine an answer to, because the nature of the cell culture-based protocol results in huge variability in the measured amounts of DNA methylation within a population of mESCs. In addition, although serum-grown mESCs are very similar morphologically to one another, there is still the debated question of whether the subtle differences that I observe, although consistent with mass spectrometry and sequencing measurements, are still merely a reflection of differences in, for instance, cell shape. This concern is addressed further in section 5.2.7.

DNMT1-inducible knock-out FRAP

To address this outstanding issue of cell morphology and to assess perhaps more clearly the sensitivity of the FRAP-based methodology for assessing DNA methylation, I conducted FRAP-based experiments utilising the DNMT1 fl/fl doxycycline-inducible knock out mESC line (hereto referred to as DNMT1-iKO; Sharif et al., 2016). Upon induction with doxycycline a region of the DNMT1 gene spanning exons 2-5 is removed from the protein. This results in a misfolded protein that is unable to methylate DNA resulting in a loss of methylation maintenance. As such, it has been shown that upon induction of the deletion in mESCs there is a concomitant passive loss in DNA methylation owing to cellular replication (unpublished Berrens, R. and Sharif et al., 2016). Recently, the dynamics of this process have been assessed using RRBS and WGBS across a number of defined time points (unpublished Berrens, R. and Sharif et al., 2016). These results have shown that within 3 days the genome has lost roughly 35% of its methylation, and by 6 days the global methylation levels drop to 30% globally. It should be noted that, due to the costly nature of single cell methylome studies, the dynamics of this demethylation process have not been studied using single cell bisulfite sequencing. As such the hypothesis of passive demethylation, although likely, has never been directly proven.

To study this demethylation process using this FRAP-based system, I performed an experiment whereby I induced demethylation using doxycycline and performed FRAP on the resulting DNMT1-iKO mESCs at intervals of 24 hrs over four days and compared these results to an uninduced control.

To validate that this experiment had been successful, IF quantitation of DNMT1 protein levels were made. Figure 5.12 shows that as expected the levels of DNMT1 expressed in the cells is reduced as a function of time following doxycycline induction.

This experiment showed that it was possible to discern methylation changes across the whole time course upon induction with doxycycline (Figure 5.13). This is in keeping with our ability to detect even subtle differences in DNA methylation levels using this method. Indeed, from averaging across recovery half-lives of the cells it is possible to obtain a fit of loss of methylation that recapitulates the demethylation dynamics seen from WGBS (Figure 5.14A). Excitingly, I was able to detect subtleties in the dynamics of the loss of DNA methylation in single, living cells. These subtleties have not been

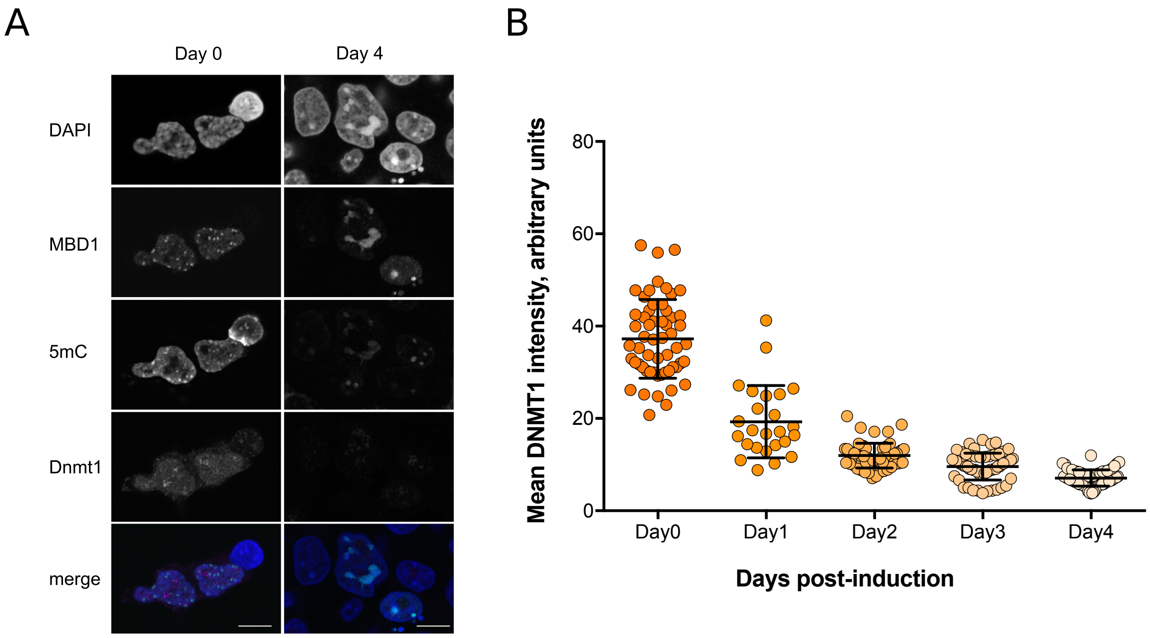


Figure 5.12: DNMT1 protein level declines upon doxycycline induction in DNMT1fl/fl mESCs: (A) Example images from IF and fluorescence measurements on fixed cells. In the first column DNMT1 fl/fl mESCs are shown (pre-induction). In the second column, DNMT1 fl/fl mESCs are shown four days post-induction with doxycycline. (B) A quantitation of DNMT1 protein levels as a function of time post-induction. Error bars represent standard deviation.

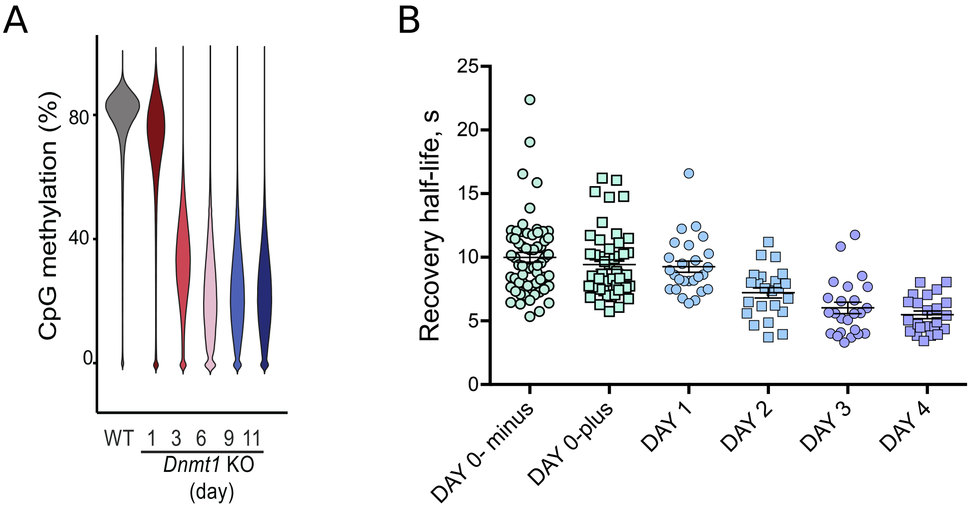


Figure 5.13: FRAP measurement of DNA methylation changes upon DNMT1 deletion: (A) Violin plot of DNA methylation levels following DNMT1 knock-out (unpublished, Berrens, R.). (B) FRAP recovery half-lives following DNMT1 knock-out. Day0-minus (pre-induction) Day0-plus (post-induction).

detected using IF-based measurements, and would be incredibly costly to observe using current scBS-Seq approaches. I was able to detect four sub-populations of cells across the entire time course. These four cell clusters were determined using an expectation-maximisation approach with a Gaussian Mixture Model (GMM) containing four states (Figure 5.14B). Cells were assigned to their maximal states and the dynamics of state transitions visualised across time (Figure 5.14C). This suggested that the spectrum of global methylation levels that were observed in the bulk WGBS dataset, were simply mixtures of varying numbers of cells from these four states. This is not necessarily surprising as it is hypothesised that this process would result in a largely passive, replication-dependent demethylation, but it is exciting to see the single cell behaviour of these dynamics. However, it should be noted that by assigning states I am inherently biasing for their existence.

It is interesting to note that although this is thought to be a passive, replication dependent process, the cellular states do not represent mid-way methylation states, i.e. 70% to 35%. This is likely because the *de novo* DNMTs are still functioning normally in this situation. In addition, I observe that there appears to be no fifth state, suggesting that cells are unable to lose more methylation than the amount present in this state. Also, I find that at day 4 not all cells are in this final state, i.e. the maximal demethylation is yet to have occurred in all cells (Figure 5.14D). It would be interesting in the future to extend the time course to validate that all cells are in state-3 by day 6, as expected from WGBS (Figure 5.13A).

Lastly, I was able to identify two separate clusters of cells within the starting population of mESCs that differed with respect to their recovery rates (Figure 5.15).

This is intriguing because I had not previously seen this behaviour before in my own serum-grown mESC experiments, but is in keeping with what has previously been observed on occasion when comparing more or less pluripotent mESCs. More or less pluripotent mESCs are characterised as having high or low levels of nanog expression. More pluripotent mESCs (nanog high) are associated with reduced levels of global methylation, whereas less pluripotent mESCs (nanog low) are associated with high levels of DNA methylation. (unpublished, Lee, H.; Figure 5.16).

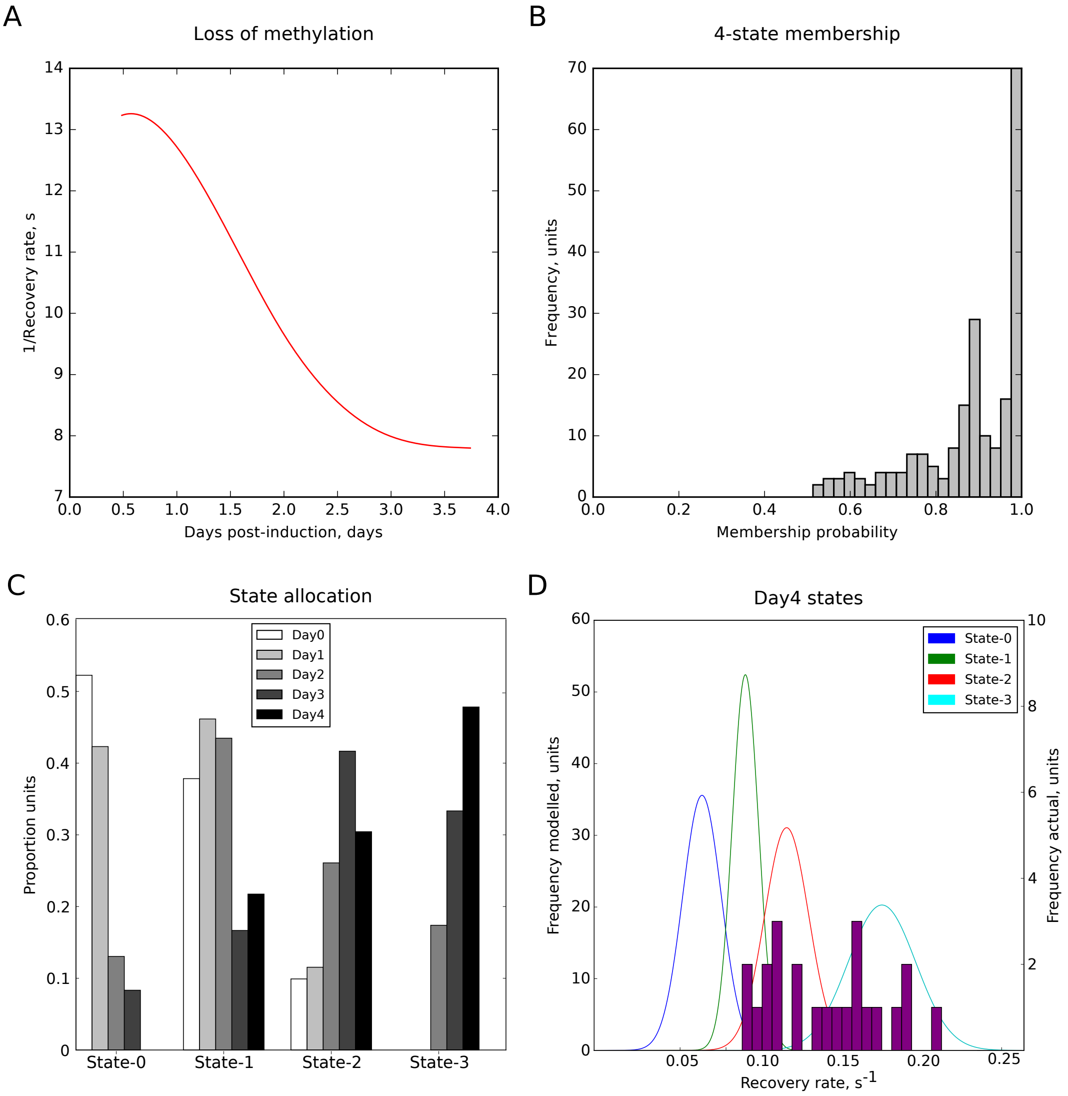


Figure 5.14: Model of DNA methylation loss: (A) Smoothed fit of the reciprocal average recovery rate with days post-induction with doxycycline. (B) Histogram of maximal membership probabilities for the four-states. Each cell was assigned to it's maximal state. (C) Barplot of the proportion of cells in each state at a given time point. (D) The states that cells are occupying on Day 4 of induction. Not all cells are assigned to State-3.

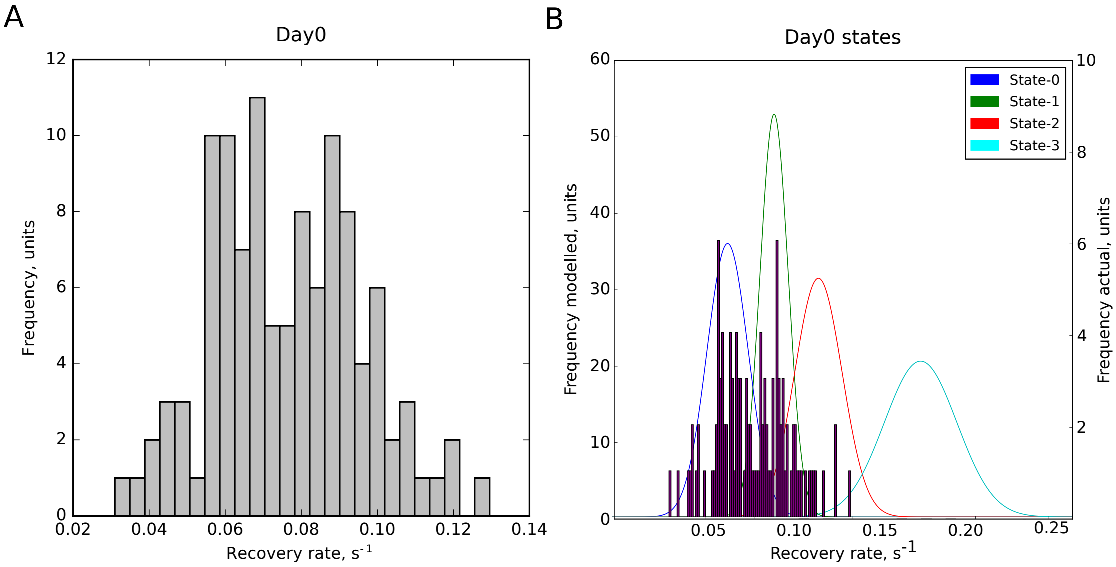


Figure 5.15: Two populations of serum mESCs measured by FRAP: (A) Histogram of recovery rate for DNMT1 fl/fl serum mESCs (uninduced, day 0). (B) The same histogram mapped on to the four states defined from the whole time course. The majority of day 0 cells are in states 0 and 1.

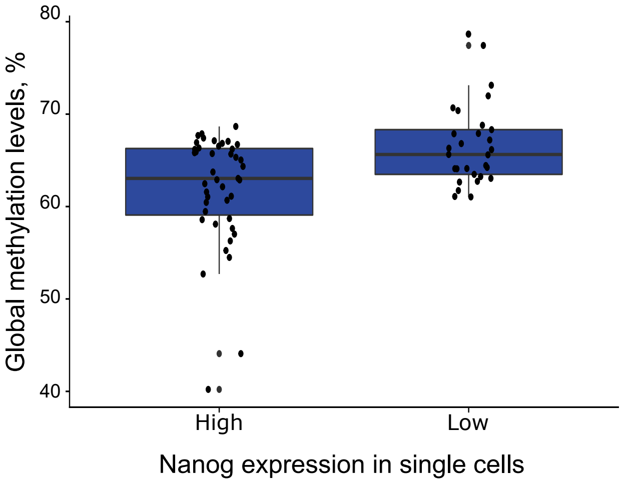


Figure 5.16: Nanog expression is inversley associated with DNA methylation levels: Boxplot to visualise global DNA methylation levels (as measured using scM&T-Seq) for high and low Nanog expression. The expression cut-off between high and low was set at 100 counts per million (CPM).

Comparison to gold-standard methods

Lastly, I wanted to assemble the results from these experiments together to assess whether the relationship between recovery half-life and global DNA methylation levels is strictly linear, or whether a non-linear function is required to explain the relationship. To assess this, I derived global methylation levels from existing sequencing-based datasets. Fitting a linear function using the *lm* function in R (R Core Team, 2017), I found that indeed the relationship between global DNA methylation levels and the recovery half-life appears linear (R^2>0.99). This is exciting because it greatly reduces the complexity of interpreting differences between DNA methylation levels when utilising this sort of approach. However, it is important to note that it is still possible that I am missing a better fit using a more complex function owing to the limited number of DNA methylation points that I am assessing, allthough this seems unlikely. One way this could be addressed in the future is by including additional cell types with differing global methylation levels.

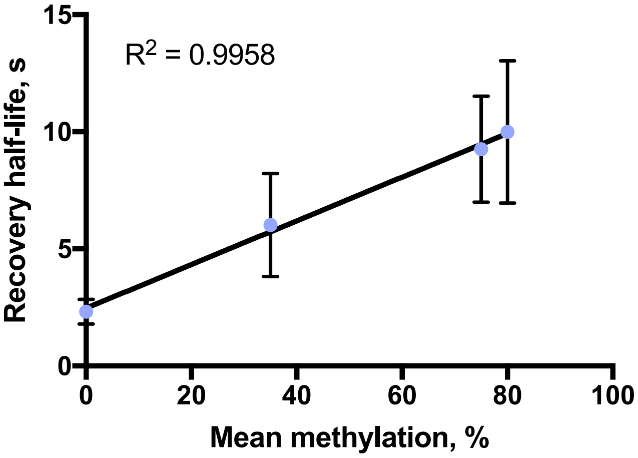


Figure 5.17: FRAP measurements correlate with WGBS: Plot to visualise the strength of correlation between WGBS and recover half-life measurements of global CpG methylation levels (R^2>0.99). Error bars represent standard deviation in the recovery half-life measurement.

5.2.8 Making multiple measurements from the same cell using FRAP

Lastly, I wanted to assess whether in principle I could obtain multiple measurements of global levels of DNA methylation from the same single cell. This was an important experiment for two main reasons. Firstly, I wanted to determine whether the exposure to the FRAP experiment altered the recovery half-life that was being calculated. One

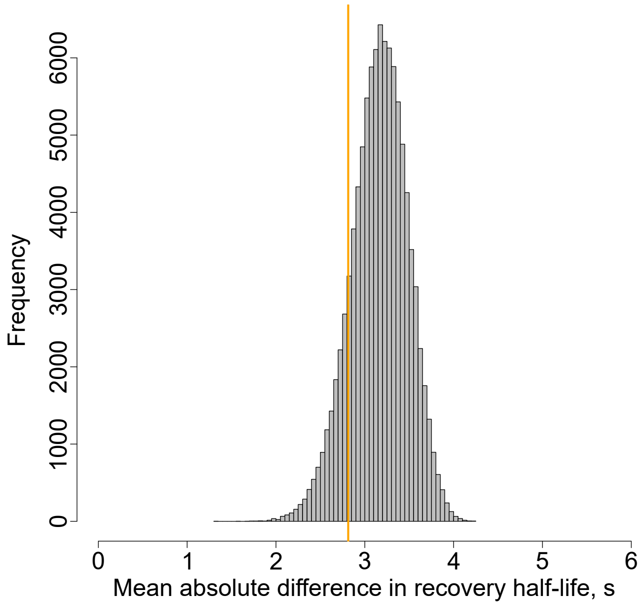


Figure 5.18: QC of multiple measurements from the same cell: A histogram of background pairwise mean absolute differences computed for the population of serum mESCs. Highlighted as an orange line is the mean absolute difference between true replicates from the same cells. There is no statistical difference between the two.

could imagine that for instance the DNA damage induced during the experiment could perhaps alter the viscosity of the cell or perhaps the accessibility of the MBD1-eGFP protein to the DNA itself. Secondly, assuming that this was not the case, I wanted to assess how reproducible any two FRAP measurements were when taken from the same cell. The reason for assessing this question from the same cell is to minimise any biological variability in the system, enabling me to get a stronger grasp on the magnitude of the technical variability. To assess this, I conducted paired FRAP experiments on serum mESCs.

I observed that it is possible to derive multiple measurements from the same cell. In addition, I could show that the variability between the measurements for any paired FRAP experiment was on average lower than you would expect by chance compared to random sampling, although not significantly so (p-value >0.05; Figure 5.18). This suggests that the variability that I am seeing in the population is potentially not solely due to technical variability in the measurement, however this cannot be fully excluded.

5.2.9 Conclusion from the FRAP experiments

These FRAP experiments have demonstrated that it is possible to measure global levels of DNA methylation in a highly reproducible fashion using the recovery half-life of the MBD1-eGFP protein. Importantly, these measurements have been conducted in single, live cells, validating that such measurements are indeed possible. It has been possible to demonstrate that the method can discern differences between global levels of DNA methylation that are on par with those observed for the gold standard single cell bisulfite sequencing approach. I have shown that such live, single cell, global methylation measurements are possible through the use of a single fluorophore, greatly reducing the potential for any confounding effects due to cooperativity. Excitingly, I have shown that this method is able to add interesting insights into a well-established and well-studied DNA methylation model system, that of the DNMT1-iKO mESCs. In addition, I have been able to demonstrate, although only over time intervals of minutes, that it is possible to assess global levels of DNA methylation in a dynamic fashion. Lastly, these experiments have enabled me to define important baseline and parameter relationships that can be utilised in the development of a DDM-based imaging method that is perhaps more amenable to long-term DNA methylation measurements.

5.2.10 Comparison of FRAP and Differential Dynamic Microscopy (DDM)

Having shown that it is possible to measure quantitative methylation levels in living single cells using FRAP, I next wanted to begin to assess whether it would be possible to define an alternate approach that could be utilised to capture many same-cell snapshots. As such I developed a differential dynamic microscopy (DDM) approach. There are a number of imaging and data analysis benefits to this method over FRAP. Firstly, no manual image annotation is required. This is in contrast to FRAP where regions of the image, namely, the photo-bleached region, the region containing the cell and a background control region, have to be defined post-image acquisition. Secondly, the imaging can be conducted on a standard wide-field microscope and as such does not require any of the specialized equipment or software specific to the FRAP-based experiment. Thirdly, the image acquisition time itself is far shorter. In comparison to FRAP, the data acquisition in DDM is more than 10x faster, which enables faster imaging and greater time resolution, whilst minimising damage to the cell. Lastly, cellular damage is also minimized by the absence of any high-intensity laser treatment of the cellular material in DDM, this is in contrast to FRAP, which relies upon laser ablation.

5.2.11 Description of the Differential Dynamic Microscopy approach

Differential dynamic microscopy (DDM) is an image analysis method that applies Fourier transform techniques that can be used to determine ensemble dynamics of isotropic processes (Cerbino and Trappe, 2008, Lu et al., 2012 and Wilson et al., 2011; Figure 5.19).

The principal of DDM involves calculating the mean Fourier transform of the time correlated difference of a series of images, *I(r,t)* (Figure 5.21). Mathematically this is described here, adapted from previous publications (Cerbino and Trappe, 2008, Lu et al., 2012 and Wilson et al., 2011).

$$B(\underline{r},\tau) = I(\underline{r},t+\tau) - I(\underline{r},t) \tag{5.7}$$

where r is the 2-dimensional vectorial position in the image, t is the frame number of the image and τ is the correlation time used to calculate the difference and is less than the total number of frames, T. It should also be noted that to remove all static and non-motile components in the images the initial frame is subtracted from all images so in effect equation 5.7 is,

$$B(\underline{r},\tau) = [I(\underline{r},t+\tau) - I(\underline{r},0)] - [I(\underline{r},t) - I(\underline{r},0)] \tag{5.8}$$

However, for ease of following the mathematics equation 5.7 will be used. The Fourier

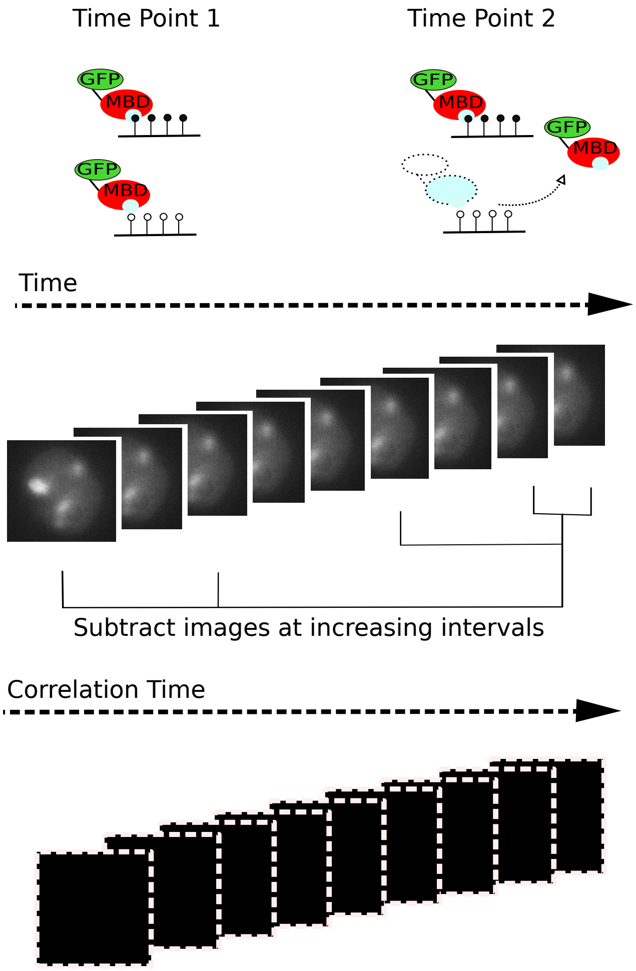


Figure 5.19: Diagram of DDM imaging steps: A visualisation of DDM imaging. A cell is shown at two separate time points to depict the motion of the eGFP within the cell. This motion is captured in an ensemble manner in the images shown below. All images are subtracted from one another to define regions that change in pixel intensity over a given time frame. These subtracted images are then Fourier transformed and averaged for each time difference (correlation time) for downstream processing.

transform of equation 5.7 is calculated for and multiplied by its complex conjugate.

$$F_B(\underline{q},\tau) = \int B(\underline{r},\tau)e^{i\underline{q}.\underline{r}}d\underline{r} \tag{5.9}$$

$$|F_B(\underline{q},\tau)|^2 = F_B(\underline{q},\tau)F*_B(\underline{q},\tau) \tag{5.10}$$

$$\langle|F_B(\underline{q},\tau)|^2\rangle = \langle F_B(\underline{q},\tau)F*_B(\underline{q},\tau)\rangle \tag{5.11}$$

where the parenthesis $\langle$ ... $\rangle$ denotes the mean for all increments of τ in the time series t to T and q is the reciprocal distance (Figure 5.20).

$$\langle|F_B(\underline{q},\tau)|^2\rangle = A(q)[1 - f(q,\tau)] + C(q) \tag{5.12}$$

If there is no variation or correlation in the image the values for $\langle|F_B(\underline{q},\tau)|^2\rangle$ will vanish rapidly to zero. It has been established for an isotropic (frequently occurring) process that equation 5.11, the differential intensity correlation function (DICF) can be expressed as,

$$\langle|F_B(\underline{q},\tau)|^2\rangle = A(q)[1 - f(q,\tau)] + C(q) \tag{5.13}$$

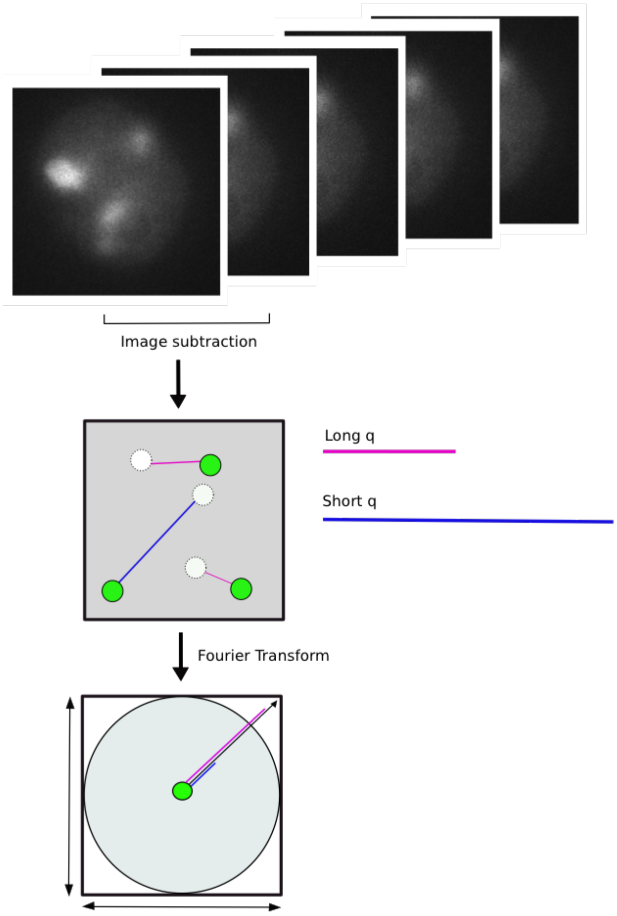


Figure 5.20: Depiction of reciprocal distance, *q*: Example DDM images of a cell are shown. These images are subtracted from one another to identify regions where the pixel intensity has increased (shown in green) or decreased (dotted white circles). A Fourier transform then identifies patterns that can explain these gains and losses. Distance in inverse space is described by reciprocal distance, q (Equation). Shown in the Fourier-space, short distances correspond to long q-values and long distances to short q-values.

5.2.12 Development of the Differential Dynamic Microscopy method

Deriving initial DDM imaging parameters

To ensure that I could capture the dynamics of the MBD1-eGFP using DDM, I first had to optimise the image acquisition for our microscope system (Materials and Methods section 2.5.13). This is important because the shorter the interval between any two frames of the DDM image (of which there are 1000 frames in total), the shorter the correlation time I can resolve. The limit to the frame rate for a given microscope is defined by a number of parameters including: the storage size of any given frame, the magnification, image binning, the shutter-speed, the aperture, and software considerations (such as location determination). In addition, to these imaging parameters, there are constraints placed on this limit owing to the nature of the sample under study. For instance, the amount of incident light required to excite enough eGFP molecules, which in turn will then emit enough photons to enable detection. In addition, an increase in intensity of the incident light will result in an increased quenching burden, which in turn would result in a reduction in image quality across the frames and result in fewer independent snapshots captured from any single cell. From testing a range of these various parameters, I qualitatively determined that the optimal frame acquisition rate was 50Hz (i.e. an image every 0.02 seconds; Materials and Methods section 2.5.13). Since I aim to capture 1000 frames per single cell, this results in an overall image acquisition time of 20 seconds. This was achieved at 100x zoom, with a reduced field of view (Materials and Methods section 2.5.13).

Description of secondary DDM analysis

Having determined the optimal image acquisition parameters for our system, I next assessed whether it was possible to detect differences in diffusion using DDM in an intracellular manner. To do this, DDM images were derived from MBD1-eGFP containing serum-grown mESCs and eGFP-only containing serum-grown mESCs.

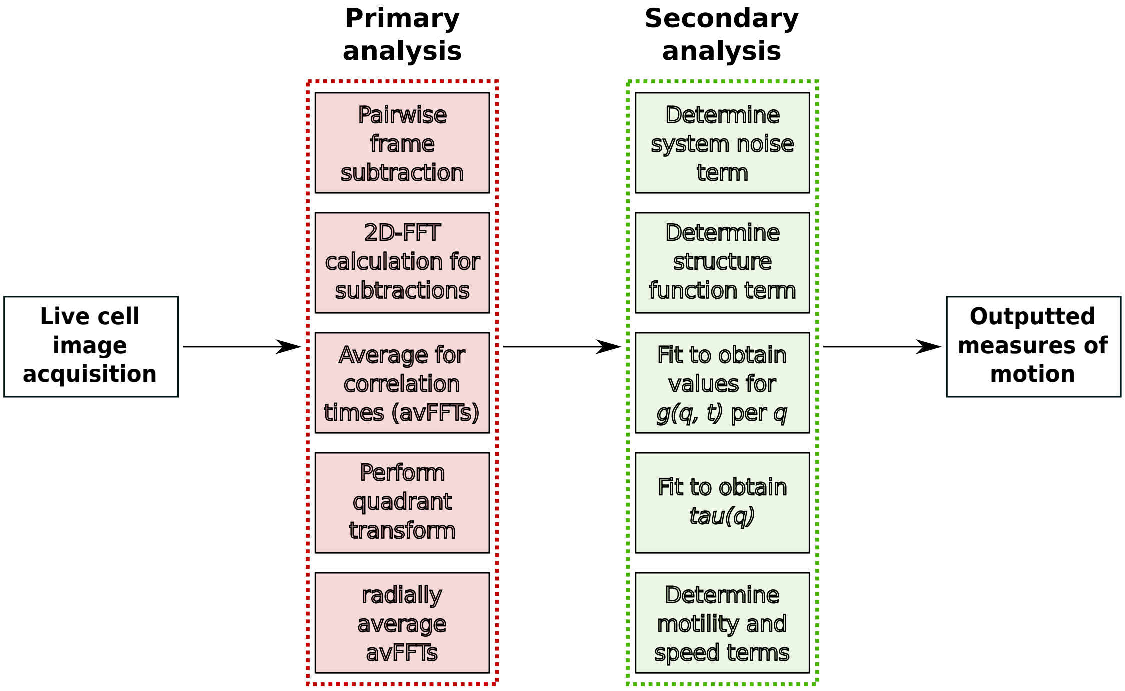

Primary analysis
Secondary analysis
Live cell image acquisition
Pairwise frame subtraction
2D-FFT calculation for subtractions
Average for correlation times (avFFTs)
Perform quadrant transform
radially average avFFTs
Determine system noise term
Determine structure function term
Fit to obtain values for *g(q, t)* per *q*
Fit to obtain *tau(q)*
Determine motility and speed terms
Outputted measures of motion

Figure 5.21: Image analysis workflow for DDM: A flow chart of the image analysis steps in the DDM pipeline. Highlighted in red are the primary analysis steps and in green are the secondary analysis steps.

From these .nd2 files I derived .tiff stacks, using ImageJ, that could be utilised by an initial implementation of the primary DDM image analysis written in matlab (unpublished Harrison, A.). This script derives the quadrant-transformed, averaged fast-Fourier Transform (FFT)-differences for each inverse distance given each correlation time.

This inverse distance averaging, assumes that the averaged 2D FFT is symmetric about the central point (q=0). This is not strictly the case in this setting because the image that I have contains defined boundaries within it. This is due to the field of view not representing a homogenously fluorescent region, but a bounded cell surrounded by background. In addition, there are strong edge effects, this can be seen from the intensity along the y-axis (Figure 5.23). However, since these appear to be systematic effects across all correlation times, I reasoned that the 1D simplification was acceptable.

From this outputted secondary dataset, I then visualised the decay of averaged FFT intensity for every inverse distance (q) as a function of correlation time (Figure 5.23).

To derive the auto-correlation function I first had to remove the system noise term $B(q)$

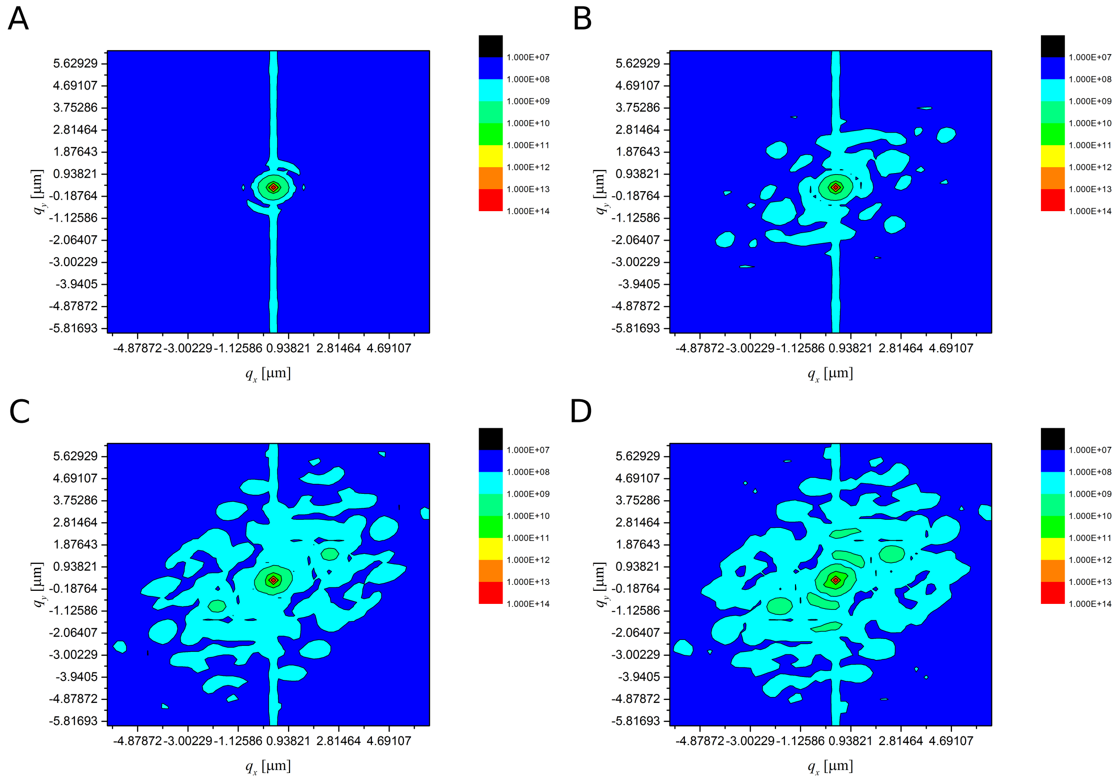


Figure 5.22: Systematic effects found that are independent of correlation time: (A-D) Representations of Fourier-transformed pixel intensities as a function of q. The representations differ with respect to correlation time (1 second (A), 5 seconds (B), 10 seconds (C), 15 seconds (D).

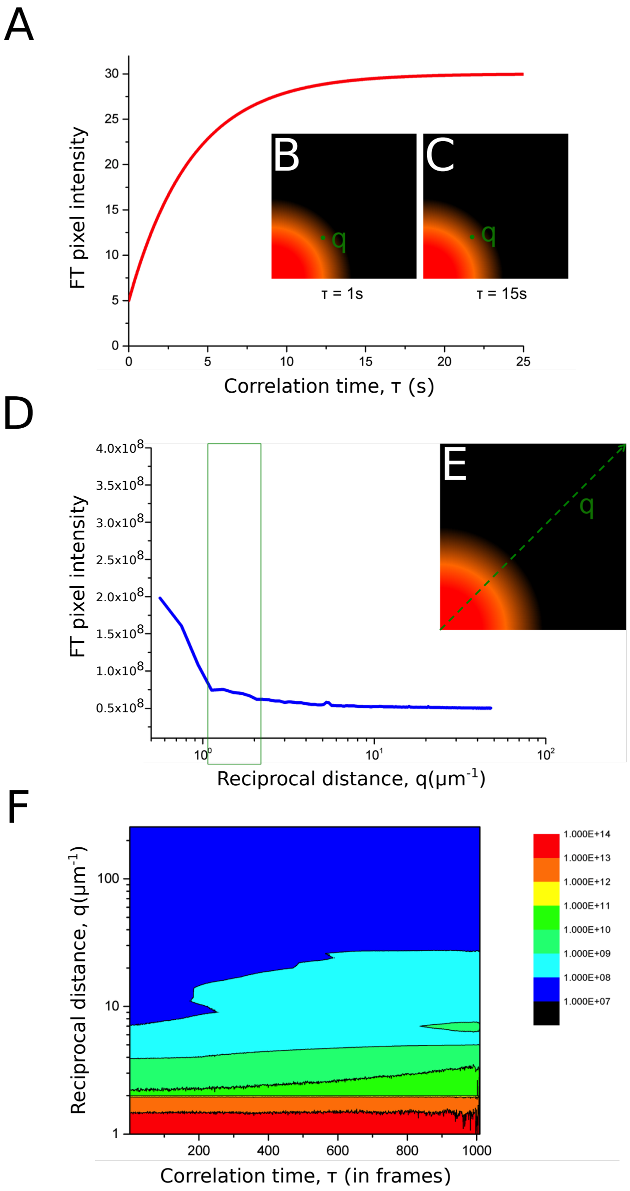


Figure 5.23: Diagramatic representation of the data following primary analysis: (A) Theoretical plot of the 1D Fourier-transformed pixel intensity as a function of correlation time (τ). Insets (B and C) depict the Fourier transformed pixel intensity (pre-radial averaging) for a q-value at two correlation times (B = 1 second and C = 15 seconds). (D) Real data plot of the 1D Fourier-transformed pixel intensity as a function of reciprocal distance (q) for a given correlation time (τ). The inset (E) depicts the Fourier transformed pixel intensity (pre-radial averaging) for all q-values at a given correlation time (τ). (F) Plot to visualise the change in Fourier-transformed pixel-intensity as a function of correlation time and reciprocal distance.

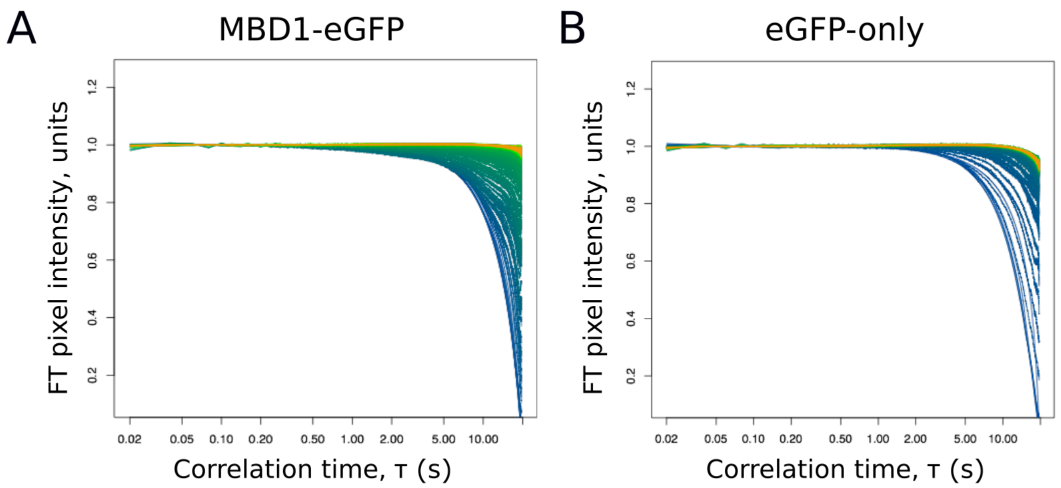


Figure 5.24: Exemplar autocorrelation functions for serum mESCs: (A and B) Plots of auto-correlation function against correlation time for each q-value ((A) is of serum mESCs containing MBD1-eGFP, and (B) is of serum mESCs containing eGFP-only).

and the structure function term $A(q)$ both of which are correlation time (t) independent. Fortunately, $g(q, t)$ in the formalism,

$$\Delta(q,t) = A(q)[1 - g(q,t)] + B(q) \tag{5.14}$$

Is defined by the exponential function, for Brownian diffusion,

$$e^{-t/\tau(q)} \tag{5.15}$$

and as such, at $t = 0$,

$$e^{-t/\tau(q)} = 1 \tag{5.16}$$

and,

$$\Delta(q,0) = B(q) \tag{5.17}$$

Whilst as t tends to infinity,

$$e^{-t/\tau(q)} = 0 \tag{5.18}$$

and,

$$\Delta(q,0) = A(q) \tag{5.19}$$

Having determined $g(q, t)$ as a function of correlation time (Figure 5.24), it is then possible to fit for,

$$g(q, t) = e^{-t/\tau(q)} \tag{5.20}$$

to obtain $\tau(q)$ for each value of q. This $\tau(q)$ corresponds to a lifetime decay, and from this it is possible to derive an effective speed and motility, which gives us our measure for diffusion,

$$\tau(q) = A.q^b \tag{5.21}$$

where A = effective speed, b = motility and where $\tau(q)$ decays with $b = -1$ for speed, $b = -2$ for diffusion and $b < -2$ for sub-diffusive behaviour.

It is important to note that constraints had to be set on this subsequent fitting. This is due to short q's representing distances that are unreasonable to expect to have captured within a single cell. Where,

$$q = \frac{2\pi}{pixelsize \times 256} \tag{5.22}$$

In addition, the signal-to-noise ratio at long q, very short distances, is such that these are also excluded.

From the initial images, it was possible to determine effective speed and motility (Figure 5.25). All values for motility were sub-diffusive, as was hypothesised from the nature of the MBD1-eGFP protein.

In addition, I compared the speed of the MBD1-eGFP construct to that of the eGFP-only mESCs and found that as expected the MBD1-eGFP construct was far slower than that of the eGFP-only. This is encouraging because it reflects the expectation that increased 5'methylcytosine binding will result in retarded diffusion. This experiment enabled the rough set up of the DDM analysis to be defined and highlighted that such an approach could work robustly. In addition, the eGFP-only data shows that DDM is able to reliably detect faster motion than my FRAP system, which was previously

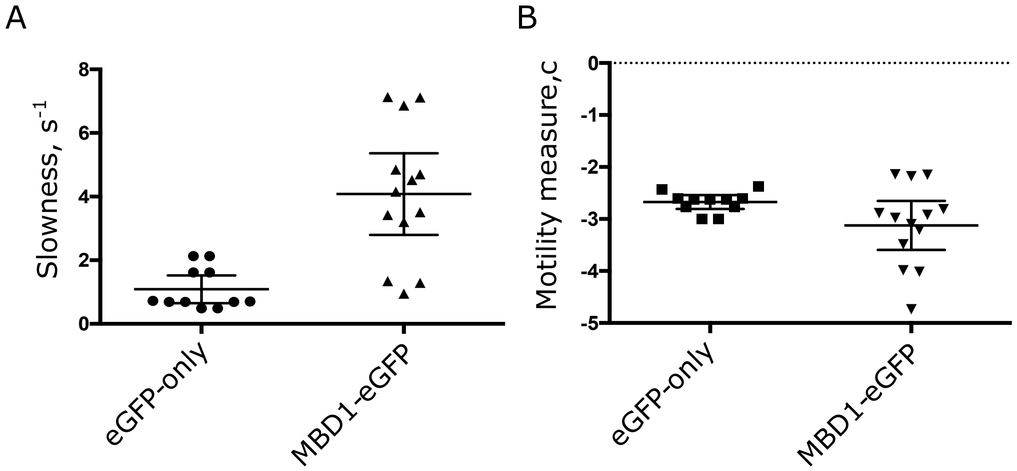


Figure 5.25: Differences in motion between MBD1-eGFP and eGFP-only detected with DDM: (A and B) Plots of slowness (A) and motility (B) for serum mESCs containing either MBD1-eGFP or eGFP only. Slowness is equal to $\frac{1}{A}$ and motility is b (Equation 5.21). Slowness for the two conditions were significantly different (p-value<0.05), motility differences were not significant.

confounded by the speed of recovery from the photo-bleaching and the limit of the speed of the microscope to capture the first frame following the bleach.

5.2.13 Implementing the primary analysis in a cluster-friendly manner

The primary analysis in this protocol is computationally costly, owing to the number of calculations that need to be made. As such, in order to increase the throughput of the approach so that I could analyse hundreds of images, I needed to develop a fast implementation of the software. As such, a primary analysis workflow was defined in java, that could be run on the cluster. Importantly, this software reduced the computational time required by any single image by a factor of 10.

In addition, a java interface was developed that could allow the interrogation of any single intermediary file, pre-radial averaging, to assess the state of the averaged FFT in all three-dimensions (Figure 5.26).

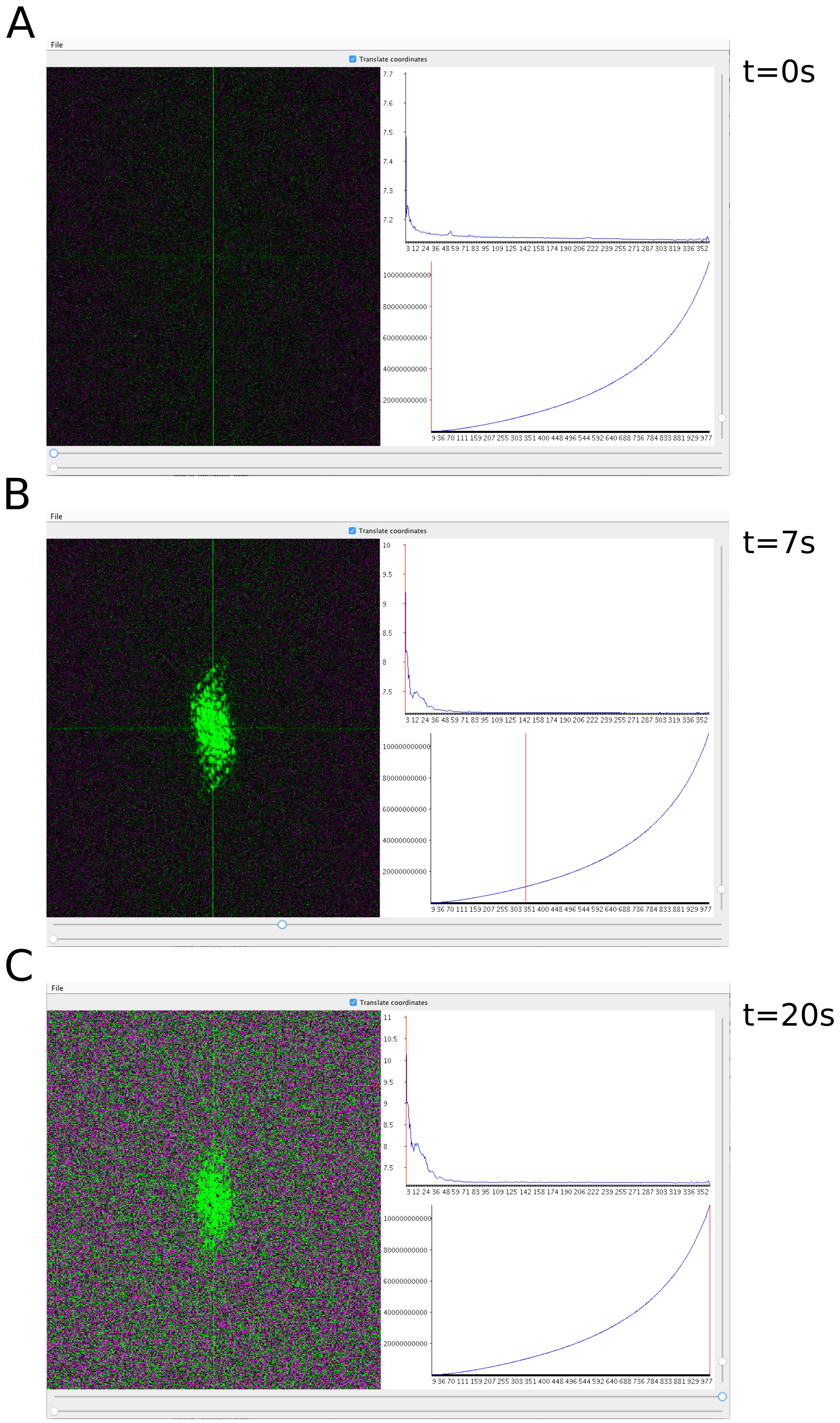


Figure 5.26: Visualisation software of primary analysis: (A-C) Screenshots of the visualisation software at different correlation times ((A) correlation time = 0s, (B) correlation time = 7s, (C) correlation time = 20s).

5.2.14 Attempts to automate the secondary analysis

Having derived a fast, cluster-compatible software pipeline, I next wanted to assess whether it would be possible to automate the fitting steps of the secondary analysis, such that many cells could be processed in batches. This is particularly important for an approach such as DDM, where for a given cell there are hundreds of curves that must be fit. This curve fitting requirement, is a positive aspect of the method in that it reduces the importance of any single fit in the output derived from the analysis. However, from an analysis point-of-view it is incredibly time-consuming.

Many approaches were trialed to fully automate the fitting stages, however, these largely failed. This was due to many factors. Firstly, it is hard to define all the potential ways in which one sample could fail, for instance owing to cell movement during the imaging process. Secondly, it is difficult to define objective criteria for pass and fail for the initial fitting in the secondary analysis, ideally such criteria would be independent of q. Lastly, although there are more, it is difficult to define parameters that you can initialize the nls function that will work in all possible scenarios, downstream of these initial problems. This does not mean that such an approach is not possible but it will require a large amount of work.

However, it was possible to derive a semi-automated secondary analysis that only required the user to determine the suitability of one fit in the second fitting step of the secondary analysis. Although this analysis was far from perfect, it enabled me to make an initial determination of the applicability of this method for measuring global levels of DNA methylation inside single, living cells.

5.2.15 Validation of the DDM method

Having built a semi-automated pipeline for the DDM analysis, I next wanted to validate that what I was measuring matched expectations in a number of control settings.

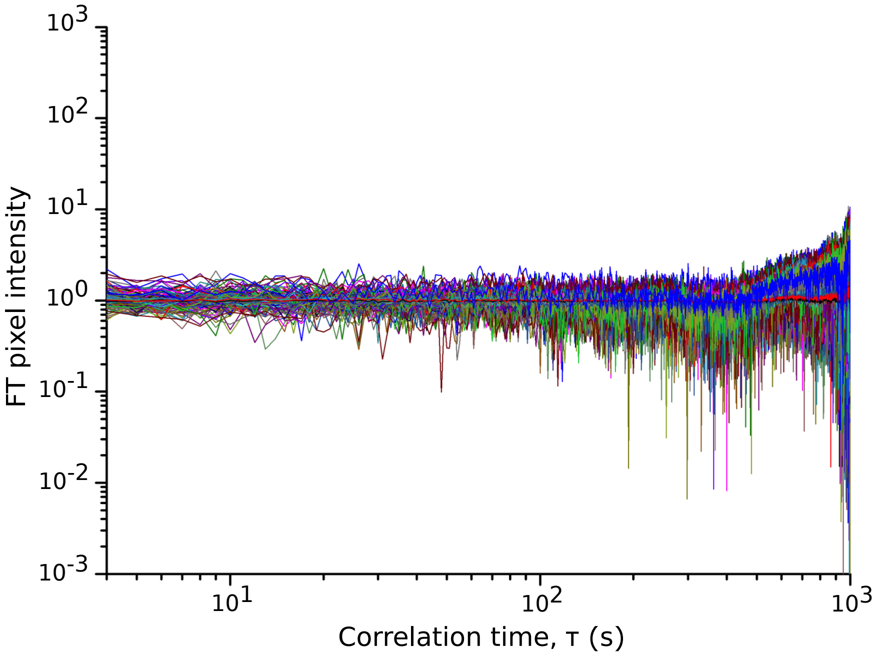


Figure 5.27: Fixed cells to do not exhibit a decay with correlation time: Plot to exemplify the fact that serum mESCs that are fixed before imaging do not decay with correlation time and as such cannot be fitted to obtain any measure of motion. Shown in different colours are the different *q*-values.)

Fixed cells do not exhibit a decay in correlation function

The first question I addressed was: could I measure a correlation function decay over increasing time in fixed cells or not? It was expected that owing to the MBD1-eGFP protein being immobile in fixed cells, it would not be possible to measure a decay above the threshold of noise. In keeping with this idea it was not possible to obtain a measurable decay coefficient in the fixed cells (Figure 5.27).

Blank images within the cell culture dish were also taken to assess background noise. These images showed little gain in pixel intensity with correlation time and so were not FFT processed (Figure 5.28). These results gave us confidence that what was being measured in the live-cell experiments were true correlation function decays resulting from the motion of the protein within the cells.

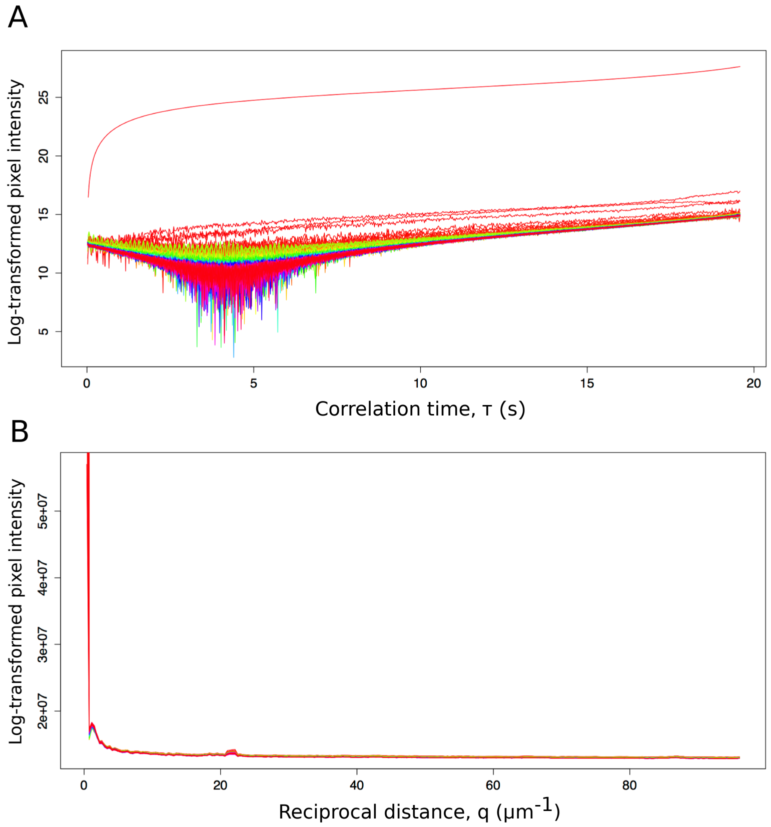


Figure 5.28: Pixel intensity of blank images does not increase with correlation time: (A) Pixel intensity plotted as a function of correlation time (τ). (B) Pixel intensity plotted as a function of reciprocal distance (q).

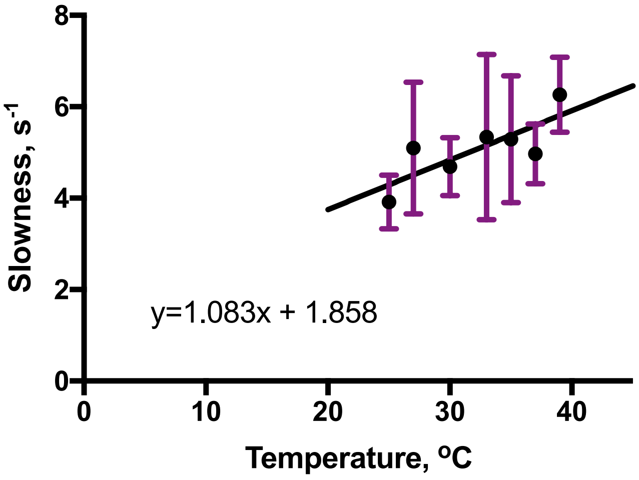


Figure 5.29: DDM measurements increase with temperature: Plot of slowness against temperature.

Temperature-relation of the speed parameter

As I am measuring the ensemble motion of protein molecules predominantly freely diffusing in the cell, I wanted to ensure that my measure of effective speed increased with an increase in thermal energy in the system, as would be expected for diffusive motion as described by the Stokes-Einstein equation (Dill and Bromberg, 2010; Equation 5.23).

$$D = \frac{K_{\mathrm{B}}T}{6\pi\eta r} \tag{5.23}$$

As such I measured the effective speed for MBD1-eGFP serum-grown mESCs at a range of temperatures (25-39°C). To ensure that the microscope and the stage had reached the desired temperature, the microscope was left for a number of hours to equilibrate and the cells within the live cell chamber on the microscope were left for 30 minutes to adjust from the incubator temperature of 37°C to the experimental temperature.

The results from this experiment showed that there was a correlation between temperature and effective speed as would be expected. However, the correlation itself is relatively weak. This is likely due to a number of factors, including the inherent difficulty in controlling the temperature within the cell. In addition, the motion of the MBD1-eGFP protein is known to be on average sub-diffusive, suggesting that there are multiple forms of motion being measured. Although one of these forms of motion would be free diffusion, the other aspects of the ensemble motion may not be so linearly related to temperature. Lastly, mESCs are inherently not behaving “naturally” at any temperatures that deviate from 37°C, as such I could partially be measuring the physiological effects of temperature on the cell, which could be confounding my measurement.

5.2.16 Measurement of DNA methylation in live cells using DDM

Having seen that I am able to measure differences between the MBD1-eGFP construct and the eGFP construct in live cells, and having validated that what I am measuring is consistent with the motion of a protein in a sub-diffusive manner, I wanted to assess whether the method could detect the differences between the presence and absence of DNA methylation.

To answer this question, I assessed methylation using one of the DNMT-TKO cell lines that I previously utilized for my FRAP analysis (see section 5.2.6).

These results highlighted that, similar to FRAP, using my semi-automated analysis pipeline, I was able to determine that there was a marked difference in effective speed of the MBD1-eGFP protein in serum-grown DNMT-TKO-C mESCs and DNMT-TKO mESCs. This difference is consistent with the idea that more methylation would result in a reduction in effective speed and it matches what was previously seen for the FRAP experiments.

5.2.17 Measurement of different levels of global DNA methylation in live cells using DDM

Having seen that I could measure a difference in slowness between cells with and without DNA methylation, I next wanted to assess whether I could detect differences between differing levels of DNA methylation. To assess this, I compared serum-grown

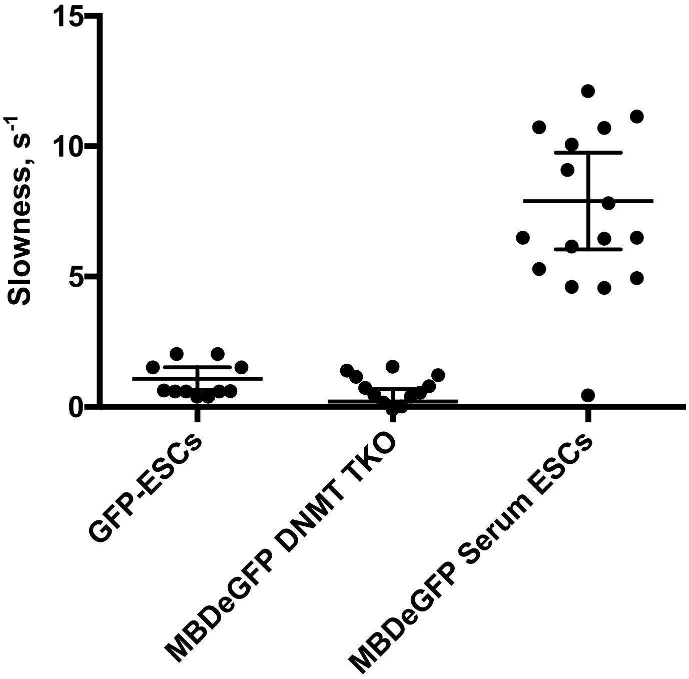


Figure 5.30: DDM can measure DNA methylation in serum mESCs: Plot depicting slowness of serum mESCs containing eGFP-only, control serum mESCs containing MBD1-eGFP and serum DNMT-TKO mESCs containing the MBD1-eGFP construct. The serum mESCs containing the MBD1-eGFP construct are significantly slower than the other conditions (p-value<0.05). There is no significant difference between the other conditions.

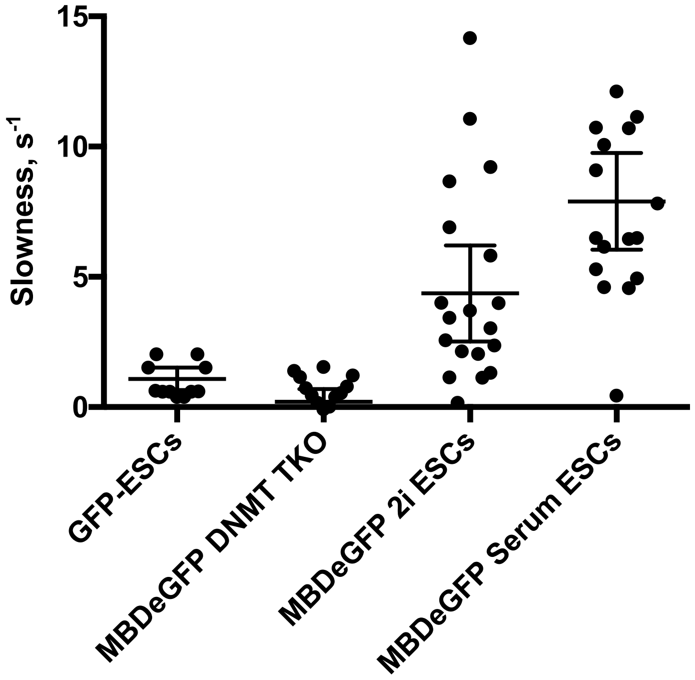


Figure 5.31: DDM measurements of serum and 2i mESCs: Plot depicting slowness of serum mESCs containing eGFP-only, serum DNMT-TKO mESCs containing the MBD1-eGFP construct and control mESCs containing MBD1-eGFP in both serum and 2i/LIF. The serum mESCs containing the MBD1-eGFP construct are significantly slower than the other conditions (p-value<0.05). There is no significant difference between the other conditions.

mESC lines (derived previously for the FRAP study) to the 2i/LIF-cultured versions of these mESCs. Serum mESCs contain ~70% CpG methylation globally and 2i/LIF mESCs contain less than 40% DNA methylation globally.

This experiment highlighted that using this method it was possible to define differences between serum and 2i/LIF mESCs and in the majority of cases it was possible to correctly predict the cell type based on this effective speed metric. However, it is clear from this experiment that in the present form of my analysis, the FRAP-based measurement of diffusion is better than the DDM measure.

5.2.18 Multiple measurements of global DNA methylation in live cells using DDM

Having shown that the method can in principle define differences between cells in terms of global levels of methylation, I next decided to assess whether this method could truly be used to define global DNA methylation levels at many snapshots. To assess this, I decided to track serum-grown mESCs across 8 hours and to image each individual cell every 30 minutes.

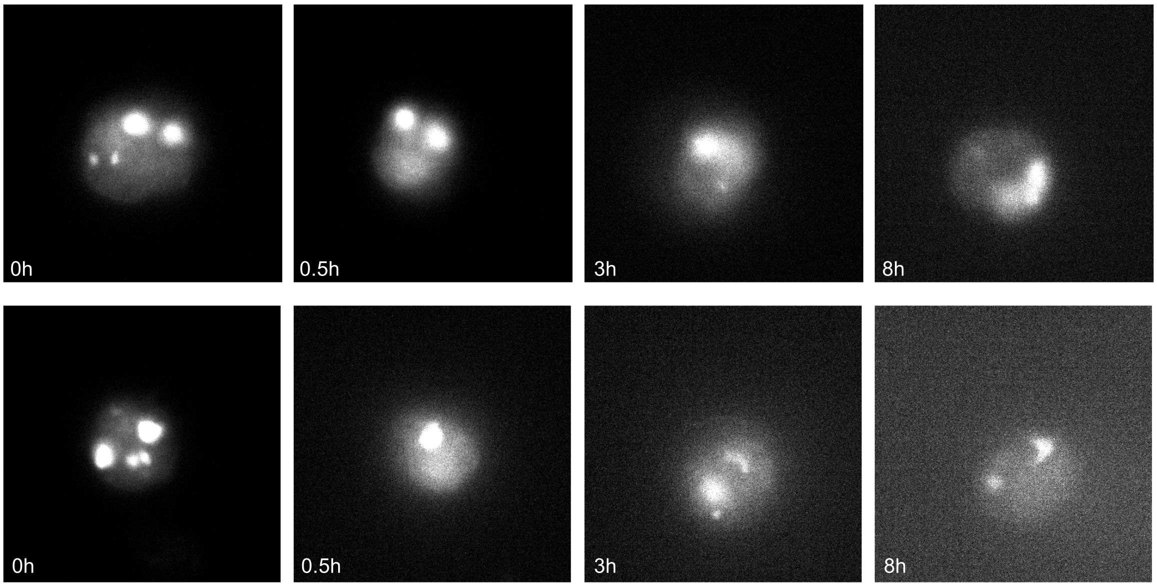


Figure 5.32: Multiple DDM measurements are possible across time: Two example cells are shown. These cells have had images taken of them every 20 minutes for >8hours. Four example images at different times are shown.

This experiment validated the utility of the DDM approach when multiple time points are required, since it was still possible to image cells following more than 10 snapshots. This is in contrast to FRAP where the cells are beginning to exhibit signs of cell death following two image acquisitions. In fact the limit to the number of snapshots that can be acquired in this instance, appear to be reliant on two main factors: the amount of incident light that is utilized for the imaging process and the recovery time allowed between snapshots. Although not shown here, I have further validated that it is possible to acquire images from the same cell using this DDM approach over time intervals from 20 seconds to more than 24 hours. Unfortunately, at present the error associated with the analysis is still far too large.

5.2.19 Conclusions from the DDM experiments

In the DDM section of this chapter, I have shown that it is possible to measure the motion of a protein within a single cell using a standard wide-field microscope. In addition, I have shown that it is possible to measure changes to the diffusion of said protein due to methylation. In addition, I have validated that I am not able to measure diffusion in fixed cells and that the measurements being made are proportionate to temperature. Most importantly, I have shown that it is possible to take a large number of images of the exact same live cell and that the time interval for such images can range from seconds to more than a day.

Unfortunately, it has not been possible as of yet to fully address the nuances of the secondary analysis. This has resulted in poor curve fitting and a reduction in both the precision and accuracy of my measurements. There are a number of practical reasons for this failing, in addition, to some more technical issues. I hope in future iterations of the scripts used to define this analysis that I will be able to effectively remove many of these issues. For instance, one aspect of the secondary analysis that needs to improve is the initialization of parameters during the first fit, such that I can incorporate additional terms to separate out the multiple forms of motion that are potentially occurring. Another issue that has not been mentioned, but that I hope to overcome is that presently, it is a requirement of the methodology that there is only one cell in the field of view. I hope in future iterations to be able to develop an analysis pipeline that will enable multiple cells to fall within the field of view. This would dramatically improve the ease of applicability of this approach to a large variety of systems under study, including but not limited to *in vivo* contexts, such as early development and ageing.

5.3 Discussion

I have developed two independent systems for the quantification of DNA methylation dynamics in live single cells. These approaches are non-overlapping and hence depending on the requirements of a given live cell experiment either could be more appropriate.

This FRAP-based system, is similar to previously described FRAP systems assessing different proteins and different motions (Conn et al., 2013 and Kang et al., 2015). It is able to discern very subtle differences in DNA methylation levels and appears to be independent of nuclear shape and size. In addition, the measure seems to be independent of the expression level of the construct. Unfortunately, this approach is not applicable to scenarios where multiple images required. To overcome this obstacle, I simultaneously adapted a novel imaging approach; DDM. At present, much work remains to be done to further improve the secondary analysis, and with it the accuracy of this approach. However, I think that this approach is exciting and that it will be widely applicable both within the field of DNA methylation, but also for assessing dynamics of other processes within cells.

Lastly, it will be of interest to see whether these approaches will be adaptable to more targeted live cell quantifications of DNA methylation, for example using FRET. To enable future development of such a system, I have generated a number of constructs detailed in the Materials and Methods section 2.5.4.

Chapter 6

Conclusions and Outlook

In this thesis I have developed and utilised novel experimental and computational methods for the study of DNA methylation dynamics. Taking DNA methylation as my model system, I have applied these methods to gain insights into the nature of ageing; both at the whole tissue level and at the single cell level.

In this thesis, I defined a multi-tissue epigenetic predictor in the mouse. This predictor is accurate with an error of 3.33 weeks and can determine deviations in biological age upon interventions including ovariectomy and high fat diet, both of which are known to reduce lifespan in mice. Next, I described the analysis of a homogeneous population of muscle satellite cells (MuSCs) that I interrogated at the single cell level, using single cell combined transcriptome and methylome sequencing (scM&T-Seq). I found that with age there was increased global transcriptional variability and a reduction in transcriptional network connectivity. I discovered that there was decreased feature-specific methylome homogeneity with age, and that this was in contrast to the increased predictability of methylation status from neighbouring positions within the same cell. These findings explain the loss of functionality of these cells with age and suggest a mechanism for how an epigenetic predictor may work at a single cell level. Lastly, I describe two imaging approaches to study DNA methylation dynamically in single cells. Using these methods, I demonstrate that it is possible to accurately determine methylation status across a wide spectrum of global methylation levels and that by using such approaches novel information about dynamic methylation processes can be obtained. These methods represent the first to study DNA methylation dynamically in a living system.

Since the publication of this multi-tissue epigenetic predictor of age in the mouse (Stubbs et al., 2017), I have been working to define an updated multi-tissue epigenetic predictor. There were a number of reasons for this as outlined in Chapter 3. At present, the dataset used for this new predictor contains >700 samples and the predictor itself has an accuracy of 5.33 weeks (Chapter 3). This accuracy represents a marked improvement over the current predictor when normalised by the maximal chronological age used in prediction (>8% vs <4% error). In the future, this predictor will utilise more than 1,000 samples that will span the whole lifespan of a mouse from pre-birth to 3 years. Included within the overall dataset will be samples derived from additional tissues not included in the published predictor including whole blood, kidney and skin. Alongside this, the predictor will be defined using data from three methods: WGBS, RRBS and scBS-Seq. Defining the predictor using these multiple sources of epigenetic information will not only make it more roust but also more amenable to different experimental systems.

In addition to developing an elastic-net regression model using this dataset, it would also be interesting in the future to utilise other modelling approaches to improve the robustness of this prediction of biological age. Machine Learning approaches such as Random Forrest are potentially interesting in this setting owing to the built-in redundancy inherent in these models potentially providing a solution to the problem of missingness in epigenetic data (particularly sequencing data). However, such models come with caveats, for instance, designing cheaper, higher throughput assays in the future may be difficult owing to the number of sites that information is interrogated over.

In the imminent future, it will be exciting to assess predictions of biological age in putative rejuvenation systems. In particular those that have so far precluded lifespan studies due to ethical or time considerations. Such systems include, the assessment of heterochronic parabionts and *in vivo* reprogrammed mice. Neither of which have been studied in a lifespan setting; due to ethical and time constraints respectively (McCay et al., 1957 and Mosteiro et al., 2016).

It will also be incumbent upon us in the future to develop a simple and affordable assay of the mouse epigenetic clock. This is integral to ensuring that the promise of such a predictor is realised. There are a number of ways that such a test could be developed. One such approach would be that of amplicon sequencing, this approach would enable precise targeting of the sites of interest, whilst maintaining the benefits of NGS. The disadvantage of such an approach would be that it is complex to set up in the first instance owing to the number of PCR primers that would be required for the current iterations of the multi-tissue predictors. An alternative approach would be to utilise a customised methylArray. This approach is more straight forward to develop in the first instance, owing to the simple requirement of hybridisation of the targets to the probes, without the necessity of targeted PCR. The disadvantage of such an approach relate to the reduced throughput and increased cost relative to the amplicon.

Utilising measurements derived from epigenetic age as a proxy for healthspan and/or lifespan combined with a cheap assay for epigenetic age in the mouse would greatly reduce the iteration time of mammalian ageing studies. Genetic, environmental and drug manipulations could be screened for reduced biological age, enabling screening for ageing in a mammalian context. This in turn would enable new and unexpected leads to be followed in a rational, open-minded fashion.

My work has shown that it is possible to define an epigenetic predictor in not just humans, the great apes and canines but also now the mouse. My work has now shifted the question of conservation further back in evolutionary time. Is the ability to define epigenetic predictors of age limited to placental mammals? Or is it conserved outside of mammalia? To address these questions future models would need to be derived from methylome datasets in other species. The implications of this conservation together with the finding that these predictors seem to be reflecting genome wide behaviour raise many interesting questions about the nature of lifespan and ageing. For instance, are the genomes of long lived species inherently more able to repair epimutations resulting from the ravages of time? Or are their genomes inherently less prone to epimutations owing to the nature of the DNA sequence or 3D organisation? Answers to these questions and others would greatly benefit from future work set upon defining a pan-species epigenetic predictor. Such a predictor, would also give insight into whether the epigenetic changes that are seen with age are causative in ageing or a passive bystander.

Deep insights into these questions will likely also come from future work defining single cell based epigenetic predictors, or experimentation-led single cell models for how such an ensemble measurement could arise from single cell information. Such information could test the hypothesis described by Horvath (Horvath, 2013) that these epigenetic predictors are the "cumulative effect of an epigenetic maintenance system". Consequently, one would expect the "burden of age" to be shared by all cells if this hypothesis were to be true. Should the "burden of age" be limited to a small number of cells, this would suggest a wholly different mechanism for the clock. Another important implication of defining a single cell predictor is that it would enable rare populations of cells to be assayed for biological age. This would be important for understanding the changes to the epigenetic predictor during early development that could inform when and how such a predictor is initialised. Furthermore, it would allow the assessment of rare stem cell populations with age, such as the TA-Hi MuSCs assessed in this thesis. This is of enormous interest because such rare cell populations are potentially the most amenable to regenerative interventions and rejuvenation.

The work described in Chapter 4 on the TA-Hi MuSC single cell combined epigenomic and transcriptomic dataset has resulted in the development of two novel computational approaches for the study of single cell DNA methylation data. In the future, it will be interesting to further develop and test these approaches in this and other systems. For instance, further analysis of the current dataset could interrogate the differences between young and old cells using experimentally determined feature annotations. Such annotations are available from chromatin immunoprecipitation (ChIP) and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-Seq) methods performed on these cells, and cell-culture models thereof (Asp et al., 2011, Blum et al., 2012 and Lilja et al., 2017) and more differentiated counterparts of them. More broadly, it will be of interest to apply these computational approaches to other single cell bisulfite datasets. For instance, in the context of the zygote and early embryo where there are large-scale dynamic changes in DNA methylation (Amouroux et al., 2016 and Santos et al., 2002), the mechanics of which are still not fully understood.

The analysis of this single cell dataset has also alluded to a potential mechanism for the epigenetic predictors: a mechanism of increasing intracellular homogeneity and resultant changes in cell-to-cell heterogeneity. In the future it will be exciting to understand whether this model that was developed from analysis of specific features in the genome, is true for the large majority of the genome. In addition, it will be important to develop a model that is able to predict the nature of these changes from initial methylation states in single cells from young individuals. Such a model would provide insights as to the nature of ageing and may help in deriving a pan-species epigenetic predictor of age.

Together with such a model, it would be of great interest to know whether this behaviour holds in more differentiated cell types in the muscle lineage. For instance, do more differentiated cells exhibit this tendency to increase intracellular epigenetic homogeneity whilst increasing cell-to-cell heterogeneity too? Will the heterogeneity be increased due to the asymmetric cell division known to occur in these TA-Hi MuSCs? To answer these questions and more it will be of interest in the future to assess activated MuSCs, myoblasts and nuclei from fused myofibers at the single cell combined level. In addition, such a study would be invaluable for asking more fundamental questions about the biology of these cells. For instance, does the correlation between transcription and DNA methylation increase through differentiation? At what stage do DNA methylation levels reach those expected for somatic tissues and known to exist in the bulk muscle? This last question could be addressed using the MethylRO mouse (contains an MBD1-fluorophore construct; Ueda et al., 2014) and the dynamic measures of DNA methylation that I developed in Chapter 5. Utilising the live cell approaches would be of particular advantage should the global gain in methylation only occur upon cell fusion with pre-existing muscle fibers. The live-cell method would allow these newly incorporated nuclei to be assessed directly even within the fibre, negating the need for complicated extraction protocols.

To enable wider applicability of the DDM live cell measure of DNA methylation dynamics, it will be important in the future to increase the robustness of the image analysis and to enable it to become more user friendly. It will also be exciting, when this analysis is made more robust to apply it to the study of other dynamic live cell processes as the approach itself is not limited to DNA methylation. Such systems could include histone marks or transcription factors. In addition, using either DDM or FRAP it is possible to make measurements of different fluorophores at the same time. One system that would be interesting to assess in the context of development and ageing would be the dynamics of PRC2 and H3K27me3. Such a study would provide insights into the nature of the relationship between the two with age. Do the global protein levels of PRC2 change? Or do the dynamics with which it interacts with the DNA change?

At a more fundamental level, it will be fascinating to study DNA methylation dynamics in different cell types in a relatively high throughput manner. This would enable screening of systems of interest that could then be followed up with more intensive and higher resolution approaches such as scM&T-Seq. As mentioned, the currently available MethylRO mouse (Ueda et al., 2014) already makes the application of these methods in an *in vivo* setting possible. Initial studies using this system have found that there are large differences in the patterning of methylation in different cell types. As such, I think that in the future it will be exciting to employ these quantitative dynamic measures to assess DNA methylation in an *in vivo* context. In particular, I think that it will be fascinating to assess these dynamics during early development and ageing.

In addition to utilising the quantitative and dynamic measures of DNA methylation described in this thesis, I think that in the future it will be important to adapt these methods to enable loci-specific measurements of DNA methylation dynamics to be made. One way this could be achieved is through a FRET-based system using the constructs derived in Materials and Methods section 2.5.4. Such an adapted system could potentially enable dynamic single cell measurements of the epigenetic predictor sites that could be followed across time. However, there will be a limit to the number of independent sites that can be uniquely identified from any one image, owing to the number of separable excitation/emission spectra. As such, it may perhaps require the adaptation of the predictor weights to a singular positive weight and singular negative weight that would require maximally 4 fluorophors.

In this thesis I have developed multiple tools to measure and analyse DNA methylation. I have shown that these tools can provide exciting insights into the nature of both DNA methylation dynamics and ageing. I hope that the additions I have made to the DNA methylation toolbox will now enable DNA methylation to be a model system for the study of ageing.

Bibliography

Acosta, J. C., A. Banito, T. Wuestefeld, A. Georgilis, P. Janich, J. P. Morton, D. Athineos, T.-W. Kang, F. Lasitschka, M. Andrulis, et al.

2013. A complex secretory program orchestrated by the inflammasome controls paracrine senescence. *Nature cell biology*, 15(8):978.

Acosta, J. C., A. O'Loghlen, A. Banito, M. V. Guijarro, A. Augert, S. Raguz, M. Fumagalli, M. Da Costa, C. Brown, N. Popov, et al.

2008. Chemokine signaling via the cxcr2 receptor reinforces senescence. *Cell*, 133(6):1006–1018.

Allsopp, R. C., H. Vaziri, C. Patterson, S. Goldstein, E. V. Younglai, A. B. Futcher, C. W. Greider, and C. B. Harley

1992. Telomere length predicts replicative capacity of human fibroblasts. *Proceedings of the National Academy of Sciences*, 89(21):10114–10118.

Amouroux, R., B. Nashun, K. Shirane, S. Nakagawa, P. W. Hill, Z. D'Souza, M. Nakayama, M. Matsuda, A. Turp, E. Ndjetehe, et al.

2016. De novo dna methylation drives 5hmc accumulation in mouse zygotes. *Nature cell biology*, 18(2):225.

Angermueller, C., S. J. Clark, H. J. Lee, I. C. Macaulay, M. J. Teng, T. X. Hu, F. Krueger, S. A. Smallwood, C. P. Ponting, T. Voet, et al.

2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature methods*, 13(3):229–232.

Angermueller, C., H. J. Lee, W. Reik, and O. Stegle

2017. Deepcpg: accurate prediction of single-cell dna methylation states using deep learning. *Genome biology*, 18(1):67.

Aoki, A., I. Suetake, J. Miyagawa, T. Fujio, T. Chijiwa, H. Sasaki, and S. Tajima
2001. Enzymatic properties of de novo-type mouse dna (cytosine-5) methyltransferases. *Nucleic acids research*, 29(17):3506–3512.

Asdell, S., H. Doornenbal, S. Joshi, and G. Sperling
1967. The effects of sex steroid hormones upon longevity in rats. *Journal of reproduction and fertility*, 14(1):113–120.

Asp, P., R. Blum, V. Vethantham, F. Parisi, M. Micsinai, J. Cheng, C. Bowman, Y. Kluger, and B. D. Dynlacht
2011. Genome-wide remodeling of the epigenetic landscape during myogenic differentiation. *Proceedings of the National Academy of Sciences*, 108(22):E149–E158.

Auclair, G., J. Borgel, L. A. Sanz, J. Vallet, S. Guibert, M. Dumas, P. Cavelier, M. Girardot, T. Forné, R. Feil, et al.
2016. Ehmt2 directs dna methylation for efficient gene silencing in mouse embryos. *Genome research*, 26(2):192–202.

Austad, S. N. and K. E. Fischer
2016. Sex differences in lifespan. *Cell metabolism*, 23(6):1022–1033.

Bahar, R., C. H. Hartmann, K. A. Rodriguez, A. D. Denny, R. A. Busuttil, M. E. Dollé, R. B. Calder, G. B. Chisholm, B. H. Pollock, C. A. Klein, et al.
2006. Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature*, 441(7096):1011.

Baker, D. J., B. G. Childs, M. Durik, M. E. Wijers, C. J. Sieben, J. Zhong, R. Saltness, K. B. Jeganathan, G. C. Versoza, A.-M. Pezeshki, et al.
2016. Naturally occurring p16ink4a-positive cells shorten healthy lifespan. *Nature*, 530(7589):184.

Baker, D. J., T. Wijshake, T. Tchkonia, N. K. LeBrasseur, B. G. Childs, B. Van De Sluis, J. L. Kirkland, and J. M. Van Deursen
2011. Clearance of p16ink4a-positive senescent cells delays ageing-associated disorders. *Nature*, 479(7372):232.

Barakat, T. S. and J. Gribnau
2010. X chromosome inactivation and embryonic stem cells. In *The Cell Biology of Stem Cells*, Pp. 132–154. Springer.

Barau, J., A. Teissandier, N. Zamudio, S. Roy, V. Nalesso, Y. Hérault, F. Guillou, and D. Bourc'his

2016. The dna methyltransferase dnmt3c protects male germ cells from transposon activity. *Science*, 354(6314):909–912.

Baris, O. R., S. Ederer, J. F. Neuhaus, J.-C. von Kleist-Retzow, C. M. Wunderlich, M. Pal, F. T. Wunderlich, V. Peeva, G. Zsurka, W. S. Kunz, et al.

2015. Mosaic deficiency in mitochondrial oxidative metabolism promotes cardiac arrhythmia during aging. *Cell metabolism*, 21(5):667–677.

Barski, A., S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao

2007. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837.

Baubec, T., R. Ivánek, F. Lienert, and D. Schübeler

2013. Methylation-dependent and-independent genomic targeting principles of the mbd protein family. *Cell*, 153(2):480–492.

Beauséjour, C. M., A. Krtolica, F. Galimi, M. Narita, S. W. Lowe, P. Yaswen, and J. Campisi

2003. Reversal of human cellular senescence: roles of the p53 and p16 pathways. *The EMBO journal*, 22(16):4212–4222.

Beerman, I., D. Bhattacharya, S. Zandi, M. Sigvardsson, I. L. Weissman, D. Bryder, and D. J. Rossi

2010. Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. *Proceedings of the National Academy of Sciences*, 107(12):5465–5470.

Bernet, J. D., J. D. Doles, J. K. Hall, K. K. Tanaka, T. A. Carter, and B. B. Olwin

2014. p38 mapk signaling underlies a cell-autonomous loss of stem cell self-renewal in skeletal muscle of aged mice. *Nature medicine*, 20(3):265–271.

Bernstein, B. E., T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, et al.

2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–326.

Bestor, T., A. Laudano, R. Mattaliano, and V. Ingram

1988. Cloning and sequencing of a cdna encoding dna methyltransferase of mouse cells: the carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *Journal of molecular biology*, 203(4):971–983.

Bestor, T. H.

2000. The dna methyltransferases of mammals. *Human molecular genetics*, 9(16):2395–2402.

Bigot, A., W. J. Duddy, Z. G. Ouandaogo, E. Negroni, V. Mariot, S. Ghimbovschi, B. Harmon, A. Wielgosik, C. Loiseau, J. Devaney, et al.

2015. Age-associated methylation suppresses spry1, leading to a failure of re-quiescence and loss of the reserve stem cell pool in elderly muscle. *Cell reports*, 13(6):1172–1182.

Bjornsson, H. T., M. I. Sigurdsson, M. D. Fallin, R. A. Irizarry, T. Aspelund, H. Cui, W. Yu, M. A. Rongione, T. J. Ekström, T. B. Harris, et al.

2008. Intra-individual change over time in dna methylation with familial clustering. *Jama*, 299(24):2877–2883.

Blackledge, N. P., A. M. Farcas, T. Kondo, H. W. King, J. F. McGouran, L. L. Hanssen, S. Ito, S. Cooper, K. Kondo, Y. Koseki, et al.

2014. Variant prc1 complex-dependent h2a ubiquitylation drives prc2 recruitment and polycomb domain formation. *Cell*, 157(6):1445–1459.

Blum, R., V. Vethantham, C. Bowman, M. Rudnicki, and B. D. Dynlacht

2012. Genome-wide identification of enhancers in skeletal muscle: the role of myod1. *Genes & development*, 26(24):2763–2779.

Bock, J.-O., H.-H. König, H. Brenner, W. E. Haefeli, R. Quinzler, H. Matschinger, K.-U. Saum, B. Schöttker, and D. Heider

2016. Associations of frailty with health care costs–results of the esther cohort study. *BMC health services research*, 16(1):128.

Bocklandt, S., W. Lin, M. E. Sehl, F. J. Sánchez, J. S. Sinsheimer, S. Horvath, and E. Vilain

2011. Epigenetic predictor of age. *PloS one*, 6(6):e14821.

Bostick, M., J. K. Kim, P.-O. Estève, A. Clark, S. Pradhan, and S. E. Jacobsen

2007. Uhrf1 plays a role in maintaining dna methylation in mammalian cells. *Science*, 317(5845):1760–1764.

Bowles, J.

1998. The evolution of aging: a new approach to an old problem of biology. *Medical hypotheses*, 51(3):179–221.

Bowles, J.

2000. Sex, kings and serial killers and other group-selected human traits. *Medical hypotheses*, 54(6):864–894.

Boyer, L. A., K. Plath, J. Zeitlinger, T. Brambrink, L. A. Medeiros, T. I. Lee, S. S. Levine, M. Wernig, A. Tajonar, M. K. Ray, et al.

2006. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *nature*, 441(7091):349.

Brack, A. S., H. Bildsoe, and S. M. Hughes

2005. Evidence that satellite cell decrement contributes to preferential decline in nuclear number from large fibres during murine age-related muscle atrophy. *Journal of cell science*, 118(20):4813–4821.

Brack, A. S., M. J. Conboy, S. Roy, M. Lee, C. J. Kuo, C. Keller, and T. A. Rando

2007. Increased wnt signaling during aging alters muscle stem cell fate and increases fibrosis. *Science*, 317(5839):807–810.

Brack, A. S. and P. Muñoz-Cánoves

2016. The ins and outs of muscle stem cell aging. *Skeletal muscle*, 6(1):1.

Brack, A. S. and T. A. Rando

2012. Tissue-specific stem cells: lessons from the skeletal muscle satellite cell. *Cell stem cell*, 10(5):504–514.

Bracken, A. P., N. Dietrich, D. Pasini, K. H. Hansen, and K. Helin

2006. Genome-wide mapping of polycomb target genes unravels their roles in cell fate transitions. *Genes & development*, 20(9):1123–1136.

Braig, M., S. Lee, C. Loddenkemper, C. Rudolph, et al.

2005. Oncogene-induced senescence as an initial barrier in lymphoma development. *Nature*, 436(7051):660.

Brien, G. L., G. Gambero, D. J. O'connell, E. Jerman, S. A. Turner, C. M. Egan, E. J.

Dunne, M. C. Jurgens, K. Wynne, L. Piao, et al.

2012. Polycomb phf19 binds h3k36me3 and recruits prc2 and demethylase no66 to embryonic stem cell genes during differentiation. *Nature structural & molecular biology*, 19(12):1273–1281.

Buenrostro, J. D., B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf

2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486.

Buscarlet, M., S. Provost, Y. F. Zada, A. Barhdadi, V. Bourgoin, G. Lépine, L. Mollica, N. Szuber, M.-P. Dubé, and L. Busque

2017. Dnmt3a and tet2 dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions. *Blood*, 130(6):753–762.

Busuttil, R., R. Bahar, and J. Vijg

2007. Genome dynamics and transcriptional deregulation in aging. *Neuroscience*, 145(4):1341–1347.

Cannon, M. V., D. A. Buchner, J. Hester, H. Miller, E. Sehayek, J. H. Nadeau, and D. Serre

2014. Maternal nutrition induces pervasive gene expression changes but no detectable dna methylation differences in the liver of adult offspring. *PloS one*, 9(3):e90335.

Carlson, M. E., M. Hsu, and I. M. Conboy

2008. Imbalance between psmad3 and notch induces cdk inhibitors in old muscle stem cells. *Nature*, 454(7203):528–532.

Carroll, J. E., M. R. Irwin, M. Levine, T. E. Seeman, D. Absher, T. Assimes, and S. Horvath

2017. Epigenetic aging and immune senescence in women with insomnia symptoms: findings from the women's health initiative study. *Biological psychiatry*, 81(2):136–144.

Cedar, H. and Y. Bergman

2012. Programming of dna methylation patterns. *Annual review of biochemistry*, 81:97–117.

Cerbino, R. and V. Trappe

2008. Differential dynamic microscopy: probing wave vector dependent dynamics with a microscope. *Physical review letters*, 100(18):188102.

Cerletti, M., Y. C. Jang, L. W. Finley, M. C. Haigis, and A. J. Wagers

2012. Short-term calorie restriction enhances skeletal muscle stem cell function. *Cell stem cell*, 10(5):515–519.

Chakkalakal, J. V., K. M. Jones, M. A. Basson, and A. S. Brack

2012. The aged niche disrupts muscle stem cell quiescence. *Nature*, 490(7420):355.

Challen, G. A., D. Sun, A. Mayle, M. Jeong, M. Luo, B. Rodriguez, C. Mallaney, H. Celik, L. Yang, Z. Xia, et al.

2014. Dnmt3a and dnmt3b have overlapping and distinct functions in hematopoietic stem cells. *Cell stem cell*, 15(3):350–364.

Chen, B., L. A. Gilbert, B. A. Cimini, J. Schnitzbauer, W. Zhang, G.-W. Li, J. Park, E. H. Blackburn, J. S. Weissman, L. S. Qi, et al.

2013a. Dynamic imaging of genomic loci in living human cells by an optimized crispr/cas system. *Cell*, 155(7):1479–1491.

Chen, B. H., R. E. Marioni, E. Colicino, M. J. Peters, C. K. Ward-Caviness, P.-C. Tsai, N. S. Roetker, A. C. Just, E. W. Demerath, W. Guan, et al.

2016. Dna methylation-based measures of biological age: meta-analysis predicting time to death. *Aging (Albany NY)*, 8(9):1844.

Chen, C.-C., K.-Y. Wang, and C.-K. J. Shen

2013b. Dna 5-methylcytosine demethylation activities of the mammalian dna methyltransferases. *Journal of Biological Chemistry*, 288(13):9084–9091.

Chiang, P. K., R. K. Gordon, J. Tal, G. Zeng, B. Doctor, K. Pardhasaradhi, and P. P. McCann

1996. S-adenosylmethionine and methylation. *The FASEB journal*, 10(4):471–480.

Christensen, B. C., E. A. Houseman, C. J. Marsit, S. Zheng, M. R. Wrensch, J. L. Wiemels, H. H. Nelson, M. R. Karagas, J. F. Padbury, R. Bueno, et al.

2009. Aging and environmental exposures alter tissue-specific dna methylation dependent upon cpg island context. *PLoS genetics*, 5(8):e1000602.

Chu, C., R. C. Spitale, and H. Y. Chang

2015. Technologies to probe functions and mechanisms of long noncoding rnas. *Na-*

ture structural & molecular biology, 22(1):29–35.

Cimmino, L., M. M. Dawlaty, D. Ndiaye-Lobry, Y. S. Yap, S. Bakogianni, Y. Yu, S. Bhattacharyya, R. Shaknovich, H. Geng, C. Lobry, et al.
2015. Tet1 is a tumor suppressor of hematopoietic malignancy. *Nature immunology*, 16(6):653–662.

Clark, S. J., S. A. Smallwood, H. J. Lee, F. Krueger, W. Reik, and G. Kelsey
2017. Genome-wide base-resolution mapping of dna methylation in single cells using single-cell bisulfite sequencing (scbs-seq). *Nature protocols*, 12(3):534–547.

Collins, C. A., P. S. Zammit, A. P. Ruiz, J. E. Morgan, and T. A. Partridge
2007. A population of myogenic stem cells that survives skeletal muscle aging. *Stem cells*, 25(4):885–894.

Colman, R. J., T. M. Beasley, J. W. Kemnitz, S. C. Johnson, R. Weindruch, and R. M. Anderson
2014. Caloric restriction reduces age-related and all-cause mortality in rhesus monkeys. *Nature communications*, 5:3557.

Conboy, I. M., M. J. Conboy, A. J. Wagers, E. R. Girma, et al.
2005. Rejuvenation of aged progenitor cells by exposure to a young systemic environment. *Nature*, 433(7027):760.

Conboy, I. M. and T. A. Rando
2002. The regulation of notch signaling controls satellite cell activation and cell fate determination in postnatal myogenesis. *Developmental cell*, 3(3):397–409.

Conboy, M. J., A. O. Karasov, and T. A. Rando
2007. High incidence of non-random template strand segregation and asymmetric fate determination in dividing stem cells and their progeny. *PLoS biology*, 5(5):e102.

Conn, K. L., M. J. Hendzel, and L. M. Schang
2013. The differential mobilization of histones h3. 1 and h3. 3 by herpes simplex virus 1 relates histone dynamics to the assembly of viral chromatin. *PLoS pathogens*, 9(10):e1003695.

Coppé, J.-P., C. K. Patil, F. Rodier, Y. Sun, D. P. Muñoz, J. Goldstein, P. S. Nelson, P.-Y. Desprez, and J. Campisi
2008. Senescence-associated secretory phenotypes reveal cell-nonautonomous func-

tions of oncogenic ras and the p53 tumor suppressor. *PLoS biology*, 6(12):e301.

Cosgrove, B. D., P. M. Gilbert, E. Porpiglia, F. Mourkioti, S. P. Lee, S. Y. Corbel, M. E. Llewellyn, S. L. Delp, and H. M. Blau
2014. Rejuvenation of the muscle stem cell population restores strength to injured aged muscles. *Nature medicine*, 20(3):255–264.

Crespi, B. J. and R. Teo
2002. Comparative phylogenetic analysis of the evolution of semelparity and life history in salmonid fishes. *Evolution*, 56(5):1008–1020.

Dabney, A., J. D. Storey, and G. Warnes
2010. qvalue: Q-value estimation for false discovery rate control. *R package version*, 1(0).

Das, C. and J. K. Tyler
2012. Histone exchange and histone modifications during transcription and aging. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1819(3):332–342.

Davis, F. M., B. Lloyd-Lewis, O. B. Harris, S. Kozar, D. J. Winton, L. Muresan, and C. J. Watson
2016. Single-cell lineage tracing in the mammary gland reveals stochastic clonal dispersion of stem/progenitor cell progeny. *Nature communications*, 7:13053.

Day, C. A., L. J. Kraft, M. Kang, and A. K. Kenworthy
2012. Analysis of protein and lipid dynamics using confocal fluorescence recovery after photobleaching (frap). *Current protocols in cytometry*, Pp. 2–19.

de Ronde, J. J., M. J. Bonder, E. H. Lips, S. Rodenhuis, and L. F. Wessels
2014. Breast cancer subtype specific classifiers of response to neoadjuvant chemotherapy do not outperform classifiers trained on all subtypes. *PloS one*, 9(2):e88551.

De Sandre-Giovannoli, A., R. Bernard, P. Cau, C. Navarro, J. Amiel, I. Boccaccio, S. Lyonnet, C. L. Stewart, A. Munnich, M. Le Merrer, et al.
2003. Lamin a truncation in hutchinson-gilford progeria. *Science*, 300(5628):2055–2055.

Decker, M. L., E. Chavez, I. Vulto, and P. M. Lansdorp
2009. Telomere length in hutchinson-gilford progeria syndrome. *Mechanisms of*

ageing and development, 130(6):377–383.

Demontis, F., V. K. Patel, W. R. Swindell, and N. Perrimon

2014. Intertissue control of the nucleolus via a myokine-dependent longevity pathway. *Cell reports*, 7(5):1481–1494.

Di Croce, L. and K. Helin

2013. Transcriptional regulation by polycomb group proteins. *Nature structural & molecular biology*, 20(10):1147–1155.

Di Leonardo, A., S. P. Linke, K. Clarkin, and G. M. Wahl

1994. Dna damage triggers a prolonged p53-dependent g1 arrest and long-term induction of cip1 in normal human fibroblasts. *Genes & development*, 8(21):2540–2551.

Dill, K. and S. Bromberg

2010. *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience*. Garland Science.

Dolle, M. E., W. K. Snyder, J. A. Gossen, P. H. Lohman, and J. Vijg

2000. Distinct spectra of somatic mutations accumulated with age in mouse heart and small intestine. *Proceedings of the National Academy of Sciences*, 97(15):8403–8408.

Dollé, M. E. and J. Vijg

2002. Genome dynamics in aging mice. *Genome research*, 12(11):1732–1738.

Domcke, S., A. F. Bardet, P. A. Ginno, D. Hartl, L. Burger, and D. Schübeler

2015. Competition between dna methylation and transcription factors determines binding of nrf1. *Nature*, 528(7583):575.

Dong, A., J. A. Yoder, X. Zhang, L. Zhou, T. H. Bestor, and X. Cheng

2001. Structure of human dnmt2, an enigmatic dna methyltransferase homolog that displays denaturant-resistant binding to dna. *Nucleic acids research*, 29(2):439–448.

Dorn, G. W. and S. J. Matkovich

2015. Epitranscriptional regulation of cardiovascular development and disease. *The Journal of physiology*, 593(8):1799–1808.

Du, Q., Z. Wang, and V. L. Schramm

2016. Human dnmt1 transition state structure. *Proceedings of the National Academy* *of Sciences*, 113(11):2916–2921.

Ebrahimi, B.

2015. Reprogramming barriers and enhancers: strategies to enhance the efficiency and kinetics of induced pluripotency. *Cell Regeneration*, 4(1):10.

Ehrlich, M., M. A. Gama-Sosa, L.-H. Huang, R. M. Midgett, K. C. Kuo, R. A. McCune, and C. Gehrke

1982. Amount and distribution of 5-methylcytosine in human dna from different types of tissues or cells. *Nucleic acids research*, 10(8):2709–2721.

Elabd, C., W. Cousin, P. Upadhyayula, R. Y. Chen, M. S. Chooljian, J. Li, S. Kung, K. P. Jiang, and I. M. Conboy

2014. Oxytocin is an age-specific circulating hormone that is necessary for muscle maintenance and regeneration. *Nature communications*, 5:4082.

Eskeland, R., M. Leeb, G. R. Grimes, C. Kress, S. Boyle, D. Sproul, N. Gilbert, Y. Fan, A. I. Skoultchi, A. Wutz, et al.

2010. Ring1b compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Molecular cell*, 38(3):452–464.

Evans, W. J. and W. W. Campbell

1993. Sarcopenia and age-related changes in body composition and functional capacity. *The Journal of nutrition*, 123(2 Suppl):465–468.

Ficz, G., T. A. Hore, F. Santos, H. J. Lee, W. Dean, J. Arand, F. Krueger, D. Oxley, Y.-L. Paul, J. Walter, et al.

2013. Fgf signaling inhibition in escs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell stem cell*, 13(3):351–359.

Fischer, J. M., P. P. Calabrese, A. J. Miller, N. M. Muñoz, W. M. Grady, D. Shibata, and R. M. Liskay

2016. Single cell lineage tracing reveals a role for tgfβr2 in intestinal stem cell dynamics and differentiation. *Proceedings of the National Academy of Sciences*, P. 201611980.

Florath, I., K. Butterbach, H. Müller, M. Bewerunge-Hudler, and H. Brenner

2013. Cross-sectional and longitudinal changes in dna methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated cpg sites. *Human* *molecular genetics*, 23(5):1186–1201.

Fong, A. P., Z. Yao, J. W. Zhong, Y. Cao, W. L. Ruzzo, R. C. Gentleman, and S. J. Tapscott
2012. Genetic and epigenetic determinants of neurogenesis and myogenesis. *Developmental cell*, 22(4):721–735.

Fontana, L., L. Partridge, and V. D. Longo
2010. Extending healthy life span—from yeast to humans. *science*, 328(5976):321–326.

Foote, A. D.
2008. Mortality rate acceleration and post-reproductive lifespan in matrilineal whale species. *Biology letters*, 4(2):189–191.

Francis, N. J., R. E. Kingston, and C. L. Woodcock
2004. Chromatin compaction by a polycomb group protein complex. *Science*, 306(5701):1574–1577.

Frenk, S., G. Pizza, R. V. Walker, and J. Houseley
2017. Aging yeast gain a competitive advantage on non-optimal carbon sources. *Aging cell*, 16(3):602–604.

Friedman, J., T. Hastie, and R. Tibshirani
2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Frobel, J., H. Hemeda, M. Lenz, G. Abagnale, S. Joussen, B. Denecke, T. Šarić, M. Zenke, and W. Wagner
2014. Epigenetic rejuvenation of mesenchymal stromal cells derived from induced pluripotent stem cells. *Stem Cell Reports*, 3(3):414–422.

Gal-Yam, E. N., G. Egger, L. Iniguez, H. Holster, S. Einarsson, X. Zhang, J. C. Lin, G. Liang, P. A. Jones, and A. Tanay
2008. Frequent switching of polycomb repressive marks and dna hypermethylation in the pc3 prostate cancer cell line. *Proceedings of the National Academy of Sciences*, 105(35):12979–12984.

Garagnani, P., M. G. Bacalini, C. Pirazzini, D. Gori, C. Giuliani, D. Mari, A. M.

Di Blasio, D. Gentilini, G. Vitale, S. Collino, et al.

2012. Methylation of elovl2 gene as a new epigenetic marker of age. *Aging cell*, 11(6):1132–1134.

Gendrel, A.-V. and E. Heard

2011. Fifty years of x-inactivation research. *Development*, 138(23):5049–5055.

Ghahramani, N. M., T. C. Ngun, P.-Y. Chen, Y. Tian, S. Krishnan, S. Muir, L. Rubbi, A. P. Arnold, G. J. de Vries, N. G. Forger, et al.

2014. The effects of perinatal testosterone exposure on the dna methylome of the mouse brain are late-emerging. *Biology of sex differences*, 5(1):8.

Girard, S. L., C. V. Bourassa, L.-P. L. Perreault, M.-A. Legault, A. Barhdadi, A. Ambalavanan, M. Brendgen, F. Vitaro, A. Noreau, G. Dionne, et al.

2016. Paternal age explains a major portion of de novo germline mutation rate variability in healthy individuals. *PloS one*, 11(10):e0164212.

Goldman, R. D., D. K. Shumaker, M. R. Erdos, M. Eriksson, A. E. Goldman, L. B. Gordon, Y. Gruenbaum, S. Khuon, M. Mendez, R. Varga, et al.

2004. Accumulation of mutant lamin a causes progressive changes in nuclear architecture in hutchinson–gilford progeria syndrome. *Proceedings of the National Academy of Sciences of the United States of America*, 101(24):8963–8968.

Goldsmith, T. C.

2004. Aging as an evolved characteristic–weismann's theory reconsidered. *Medical hypotheses*, 62(2):304–308.

Goldsmith, T. C.

2006. *The Evolution of Aging: How new theories will change the future of medicine*. Azinet.

Goldsmith, T. C.

2008. Aging, evolvability, and the individual benefit requirement; medical implications of aging theory controversies. *Journal of theoretical biology*, 252(4):764–768.

Gowher, H. and A. Jeltsch

2001. Enzymatic properties of recombinant dnmt3a dna methyltransferase from mouse: the enzyme modifies dna in a non-processive manner and also methylates non-cpa sites. *Journal of molecular biology*, 309(5):1201–1208.

Gowher, H. and A. Jeltsch

2002. Molecular enzymology of the catalytic domains of the dnmt3a and dnmt3b dna methyltransferases. *Journal of Biological Chemistry*, 277(23):20409–20414.

Gravina, S., X. Dong, B. Yu, and J. Vijg

2016. Single-cell genome-wide bisulfite sequencing uncovers extensive heterogeneity in the mouse liver methylome. *Genome biology*, 17(1):150.

Gray, M. D., J.-C. Shen, A. S. Kamath-Loeb, A. Blank, B. L. Sopher, G. M. Martin, J. Oshima, and L. A. Loeb

1997. The werner syndrome protein is a dna helicase. *Nature genetics*, 17(1):100–103.

Grossniklaus, U. and R. Paro

2014. Transcriptional silencing by polycomb-group proteins. *Cold Spring Harbor perspectives in biology*, 6(11):a019331.

Hackett, J. A. and M. A. Surani

2013. Dna methylation dynamics during the mammalian life cycle. *Phil. Trans. R. Soc. B*, 368(1609):20110328.

Hahn, M. A., T. Hahn, D.-H. Lee, R. S. Esworthy, B.-w. Kim, A. D. Riggs, F.-F. Chu, and G. P. Pfeifer

2008. Methylation of polycomb target genes in intestinal cancer is mediated by inflammation. *Cancer research*, 68(24):10280–10289.

Hahn, O., S. Grönke, T. M. Stubbs, G. Ficz, O. Hendrich, F. Krueger, S. Andrews, Q. Zhang, M. J. Wakelam, A. Beyer, et al.

2017. Dietary restriction protects from age-associated dna methylation and induces epigenetic reprogramming of lipid metabolism. *Genome biology*, 18(1):56.

Hamilton, W. D.

1964. The genetical evolution of social behaviour. ii. *Journal of theoretical biology*, 7(1):17–52.

Hamilton, W. D.

1966. The moulding of senescence by natural selection. *Journal of theoretical biology*, 12(1):12–45.

Handa, V. and A. Jeltsch

2005. Profound flanking sequence preference of dnmt3a and dnmt3b mammalian dna methyltransferases shape the human epigenome. *Journal of molecular biology*, 348(5):1103–1112.

Hannum, G., J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J.-B. Fan, Y. Gao, et al.
2013. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell*, 49(2):359–367.

Haugstetter, A., C. Loddenkemper, D. Lenze, J. Gröne, C. Standfuss, I. Petersen, B. Dörken, and C. Schmitt
2010. Cellular senescence predicts treatment outcome in metastasised colorectal cancer. *British journal of cancer*, 103(4):505.

Hayflick, L.
1965. The limited in vitro lifetime of human diploid cell strains. *Experimental cell research*, 37(3):614–636.

Hayflick, L. and P. S. Moorhead
1961. The serial cultivation of human diploid cell strains. *Experimental cell research*, 25(3):585–621.

Heintz, C., T. K. Doktor, A. Lanjuin, C. Escoubas, Y. Zhang, H. J. Weir, S. Dutta, C. G. Silva-García, G. H. Bruun, I. Morantte, et al.
2017. Splicing factor 1 modulates dietary restriction and torc1 pathway longevity in c. elegans. *Nature*, 541(7635):102.

Hernando-Herraez, I., J. Prado-Martinez, P. Garg, M. Fernandez-Callejo, H. Heyn, C. Hvilsom, A. Navarro, M. Esteller, A. J. Sharp, and T. Marques-Bonet
2013. Dynamics of dna methylation in recent human and great ape evolution. *PLoS genetics*, 9(9):e1003763.

Heyn, H., N. Li, H. J. Ferreira, S. Moran, D. G. Pisano, A. Gomez, J. Diez, J. V. Sanchez-Mut, F. Setien, F. J. Carmona, et al.
2012. Distinct dna methylomes of newborns and centenarians. *Proceedings of the National Academy of Sciences*, 109(26):10522–10527.

Hill, P. W., R. Amouroux, and P. Hajkova
2014. Dna demethylation, tet proteins and 5-hydroxymethylcytosine in epigenetic reprogramming: an emerging complex story. *Genomics*, 104(5):324–333.

Horrington, E. M., F. Pope, W. Lunsford, and C. M. McCay

1960. Age changes in the bones, blood pressure, and diseases of rats in parabiosis. *Gerontologia*, 4:21–31.

Horvath, S.

2013. Dna methylation age of human tissues and cell types. *Genome biology*, 14(10):3156.

Horvath, S., W. Erhart, M. Brosch, O. Ammerpohl, W. von Schönfels, M. Ahrens, N. Heits, J. T. Bell, P.-C. Tsai, T. D. Spector, et al.

2014. Obesity accelerates epigenetic aging of human liver. *Proceedings of the National Academy of Sciences*, 111(43):15538–15543.

Horvath, S., P. Garagnani, M. G. Bacalini, C. Pirazzini, S. Salvioli, D. Gentilini, A. M. Di Blasio, C. Giuliani, S. Tung, H. V. Vinters, et al.

2015a. Accelerated epigenetic aging in down syndrome. *Aging cell*, 14(3):491–495.

Horvath, S., M. Gurven, M. E. Levine, B. C. Trumble, H. Kaplan, H. Allayee, B. R. Ritz, B. Chen, A. T. Lu, T. M. Rickabaugh, et al.

2016a. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome biology*, 17(1):171.

Horvath, S., P. Langfelder, S. Kwak, J. Aaronson, J. Rosinski, T. F. Vogt, M. Eszes, R. L. Faull, M. A. Curtis, H. J. Waldvogel, et al.

2016b. Huntington's disease accelerates epigenetic aging of human brain and disrupts dna methylation levels. *Aging (Albany NY)*, 8(7):1485.

Horvath, S. and A. J. Levine

2015. Hiv-1 infection accelerates age according to the epigenetic clock. *The Journal of infectious diseases*, 212(10):1563–1573.

Horvath, S., C. Pirazzini, M. G. Bacalini, D. Gentilini, A. M. Di Blasio, M. Delledonne, D. Mari, B. Arosio, D. Monti, G. Passarino, et al.

2015b. Decreased epigenetic age of pbmcs from italian semi-supercentenarians and their offspring. *Aging (Albany NY)*, 7(12):1159.

Horvath, S. and B. R. Ritz

2015. Increased epigenetic age and granulocyte counts in the blood of parkinson's disease patients. *Aging (Albany NY)*, 7(12):1130.

Hu, L., J. Lu, J. Cheng, Q. Rao, Z. Li, H. Hou, Z. Lou, L. Zhang, W. Li, W. Gong, et al.
2015. Structural insight into substrate preference for tet-mediated oxidation. *Nature*, 527(7576):118.

Huang, J., Q. Gan, L. Han, J. Li, H. Zhang, Y. Sun, Z. Zhang, and T. Tong
2008. Sirt1 overexpression antagonizes cellular senescence with activated erk/s6k1 signaling in human diploid fibroblasts. *PloS one*, 3(3):e1710.

Hunkapiller, J., Y. Shen, A. Diaz, G. Cagney, D. McCleary, M. Ramalho-Santos, N. Krogan, B. Ren, J. S. Song, and J. F. Reiter
2012. Polycomb-like 3 promotes polycomb repressive complex 2 binding to cpg islands and embryonic stem cell self-renewal. *PLoS genetics*, 8(3):e1002576.

Ingouff, M., B. Selles, C. Michaud, T. M. Vu, F. Berger, A. J. Schorn, D. Autran, M. Van Durme, M. K. Nowack, R. A. Martienssen, et al.
2017. Live-cell analysis of dna methylation during sexual reproduction in arabidopsis reveals context and sex-specific dynamics controlled by noncanonical rddm. *Genes & development*, 31(1):72–83.

Islam, S., A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson
2014. Quantitative single-cell rna-seq with unique molecular identifiers. *Nature methods*, 11(2):163–166.

Iurlaro, M., G. Ficz, D. Oxley, E.-A. Raiber, M. Bachman, M. J. Booth, S. Andrews, S. Balasubramanian, and W. Reik
2013. A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome biology*, 14(10):R119.

Iyer, L. M., S. Abhiman, and L. Aravind
2011. Natural history of eukaryotic dna methylation systems. *Prog Mol Biol Transl Sci*, 101:25–104.

Iyer, L. M., M. Tahiliani, A. Rao, and L. Aravind
2009. Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell cycle*, 8(11):1698–1710.

James, S. J., S. Melnyk, M. Pogribna, I. P. Pogribny, and M. A. Caudill

2002. Elevation in s-adenosylhomocysteine and dna hypomethylation: potential epigenetic mechanism for homocysteine-related pathology. *The Journal of nutrition*, 132(8):2361S–2366S.

Jang, Y., M. Sinha, M. Cerletti, C. Dall'Osso, and A. J. Wagers

2011. Skeletal muscle stem cells: effects of aging and metabolism on muscle regenerative function. In *Cold Spring Harbor symposia on quantitative biology*, volume 76, Pp. 101–111. Cold Spring Harbor Laboratory Press.

Janssen, I., S. B. Heymsfield, and R. Ross

2002. Low relative skeletal muscle mass (sarcopenia) in older persons is associated with functional impairment and physical disability. *Journal of the American Geriatrics Society*, 50(5):889–896.

Jenuwein, T. and C. D. Allis

2001. Translating the histone code. *Science*, 293(5532):1074–1080.

Jia, D., R. Z. Jurkowska, X. Zhang, A. Jeltsch, and X. Cheng

2007. Structure of dnmt3a bound to dnmt3l suggests a model for de novo dna methylation. *Nature*, 449(7159):248.

Johnson, K. C., D. C. Koestler, C. Cheng, and B. C. Christensen

2014. Age-related dna methylation in normal breast tissue and its relationship with invasive breast tumor methylation. *Epigenetics*, 9(2):268–275.

Jung, M. and G. P. Pfeifer

2015. Aging and dna methylation. *BMC biology*, 13(1):7.

Kalari, S., M. Jung, K. H. Kernstine, T. Takahashi, and G. P. Pfeifer

2013. The dna methylation landscape of small cell lung cancer suggests a differentiation defect of neuroendocrine cells. *Oncogene*, 32(30):3559.

Kalhor, R., P. Mali, and G. M. Church

2016. Rapidly evolving homing crispr barcodes. *bioRxiv*, P. 055863.

Kang, M., M. Andreani, and A. K. Kenworthy

2015. Validation of normalizations, scaling, and photofading corrections for frap data analysis. *PloS one*, 10(5):e0127966.

Katajisto, P., J. Döhla, C. L. Chaffer, N. Pentinmikko, N. Marjanovic, S. Iqbal,

R. Zoncu, W. Chen, R. A. Weinberg, and D. M. Sabatini

2015. Asymmetric apportioning of aged mitochondria between daughter cells is required for stemness. *Science*, 348(6232):340–343.

Katsimpardi, L., N. K. Litterman, P. A. Schein, C. M. Miller, F. S. Loffredo, G. R. Wojtkiewicz, J. W. Chen, R. T. Lee, A. J. Wagers, and L. L. Rubin

2014. Vascular and neurogenic rejuvenation of the aging mouse brain by young systemic factors. *Science*, 344(6184):630–634.

Kemp, C. J., J. M. Moore, R. Moser, B. Bernard, M. Teater, L. E. Smith, N. A. Rabaia, K. E. Gurley, J. Guinney, S. E. Busch, et al.

2014. Ctcf haploinsufficiency destabilizes dna methylation and predisposes to cancer. *Cell reports*, 7(4):1020–1029.

Kenyon, C. J.

2010. The genetics of ageing. *Nature*, 464(7288):504–512.

Khaidakov, M., E. R. Siegel, and R. J. S. Reis

2006. Direct repeats in mitochondrial dna and mammalian lifespan. *Mechanisms of ageing and development*, 127(10):808–812.

Kimura, H., Y. Hayashi-Takanaka, and K. Yamagata

2010. Visualization of dna methylation and histone modifications in living cells. *Current opinion in cell biology*, 22(3):412–418.

Kirkwood, T. B.

1977. Evolution of ageing. *Nature*, 270(5635):301–304.

Kirkwood, T. B. and S. Melov

2011. On the programmed/non-programmed nature of ageing within the life history. *Current Biology*, 21(18):R701–R707.

Kirschner, K., T. Chandra, V. Kiselev, D. Flores-Santa Cruz, I. C. Macaulay, H. J. Park, J. Li, D. G. Kent, R. Kumar, D. C. Pask, et al.

2017. Proliferation drives aging-related functional decline in a subpopulation of the hematopoietic stem cell compartment. *Cell reports*, 19(8):1503–1511.

Klein, C. and F. Waharte

2010. Analysis of molecular mobility by fluorescence recovery after photobleaching in living cells. *Microscopy: Science, Technology, Applications and Education, Formatex*

Research Center, Pp. 772–783.

Klimasauskas, S., S. Kumar, R. J. Roberts, and X. Cheng
1994. Hhal methyltransferase flips its target base out of the dna helix. *Cell*, 76(2):357–369.

Koushik, S. V., H. Chen, C. Thaler, H. L. Puhl, and S. S. Vogel
2006. Cerulean, venus, and venus y67c fret reference standards. *Biophysical journal*, 91(12):L99–L101.

Kriete, A.
2013. Robustness and aging—a systems-level perspective. *Biosystems*, 112(1):37–48.

Krueger, F. and S. R. Andrews
2011. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *bioinformatics*, 27(11):1571–1572.

Kuang, S., K. Kuroda, F. Le Grand, and M. A. Rudnicki
2007. Asymmetric self-renewal and commitment of satellite stem cells in muscle. *Cell*, 129(5):999–1010.

Kuilman, T., C. Michaloglou, L. C. Vredeveld, S. Douma, R. van Doorn, C. J. Desmet, L. A. Aarden, W. J. Mooi, and D. S. Peeper
2008. Oncogene-induced senescence relayed by an interleukin-dependent inflammatory network. *Cell*, 133(6):1019–1031.

Kulis, M., S. Heath, M. Bibikova, A. C. Queirós, A. Navarro, G. Clot, A. Martínez-Trillos, G. Castellano, I. Brun-Heath, M. Pinyol, et al.
2012. Epigenomic analysis detects widespread gene-body dna hypomethylation in chronic lymphocytic leukemia. *Nature genetics*, 44(11):1236–1242.

Laflamme, M. A. and C. E. Murry
2011. Heart regeneration. *Nature*, 473(7347):326.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al.
2001. Initial sequencing and analysis of the human genome.

Lauberth, S. M., T. Nakayama, X. Wu, A. L. Ferris, Z. Tang, S. H. Hughes, and R. G. Roeder
2013. H3k4me3 interactions with taf3 regulate preinitiation complex assembly and

selective gene activation. *Cell*, 152(5):1021–1036.

Lee, T.-f., J. Zhai, and B. C. Meyers

2010. Conservation and divergence in eukaryotic dna methylation. *Proceedings of the National Academy of Sciences*, 107(20):9027–9028.

Lee, T. I., R. G. Jenner, L. A. Boyer, M. G. Guenther, S. S. Levine, R. M. Kumar, B. Chevalier, S. E. Johnstone, M. F. Cole, K.-i. Isono, et al.

2006. Control of developmental regulators by polycomb in human embryonic stem cells. *Cell*, 125(2):301–313.

Leonard, A. E., S. L. Pereira, H. Sprecher, and Y.-S. Huang

2004. Elongation of long-chain fatty acids. *Progress in lipid research*, 43(1):36–54.

Lepper, C., T. A. Partridge, and C.-M. Fan

2011. An absolute requirement for pax7-positive satellite cells in acute injury-induced skeletal muscle regeneration. *Development*, 138(17):3639–3646.

Leroi, A. M., M. R. Rose, and G. V. Lauder

1994. What does the comparative method reveal about adaptation? *The American Naturalist*, 143(3):381–402.

Levine, M. E., H. D. Hosgood, B. Chen, D. Absher, T. Assimes, and S. Horvath

2015a. Dna methylation age of blood predicts future onset of lung cancer in the women's health initiative. *Aging (Albany NY)*, 7(9):690.

Levine, M. E., A. T. Lu, D. A. Bennett, and S. Horvath

2015b. Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and alzheimer's disease related cognitive functioning. *Aging (Albany NY)*, 7(12):1198.

Levine, S. S., A. Weiss, H. Erdjument-Bromage, Z. Shao, P. Tempst, and R. E. Kingston

2002. The core of the polycomb repressive complex is compositionally and functionally conserved in flies and humans. *Molecular and cellular biology*, 22(17):6070–6078.

Lewis, E. B.

1978. A gene complex controlling segmentation in drosophila. *Nature*, 276(5688):565–570.

Li, J., L. Miao, D. Shieh, E. Spiotto, J. Li, B. Zhou, A. Paul, R. J. Schwartz, A. B.

Firulli, H. A. Singer, et al.

2016. Single-cell lineage tracing reveals that oriented cell division contributes to trabecular morphogenesis and regional specification. *Cell reports*, 15(1):158–170.

Li, Z., H. Dai, S. N. Martos, B. Xu, Y. Gao, T. Li, G. Zhu, D. E. Schones, and Z. Wang

2015. Distinct roles of dnmt1-dependent and dnmt1-independent methylation patterns in the genome of mouse embryonic stem cells. *Genome biology*, 16(1):115.

Liao, Y., G. K. Smyth, and W. Shi

2013. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

Lilja, K. C., N. Zhang, A. Magli, V. Gunduz, C. J. Bowman, R. W. Arpke, R. Darabi, M. Kyba, R. Perlingeiro, and B. D. Dynlacht

2017. Pax7 remodels the chromatin landscape in skeletal muscle stem cells. *PloS one*, 12(4):e0176190.

Lim, Y. C., S. Y. Chia, S. Jin, W. Han, C. Ding, and L. Sun

2016. Dynamic dna methylation landscape defines brown and white cell specificity during adipogenesis. *Molecular metabolism*, 5(10):1033–1041.

Lin, A. W., M. Barradas, J. C. Stone, L. van Aelst, M. Serrano, and S. W. Lowe

1998. Premature senescence involving p53 and p16 is activated in response to constitutive mek/mapk mitogenic signaling. *Genes & development*, 12(19):3008–3019.

Linnane, A., A. Baumer, R. Maxwell, H. Preston, C. Zhang, and S. Marzuki

1990. Mitochondrial gene mutation: the ageing process and degenerative diseases. *Biochemistry international*, 22(6):1067–1076.

Lister, R., E. A. Mukamel, J. R. Nery, M. Urich, C. A. Puddifoot, N. D. Johnson, J. Lucero, Y. Huang, A. J. Dwork, M. D. Schultz, et al.

2013. Global epigenomic reconfiguration during mammalian brain development. *Science*, 341(6146):1237905.

Liu, B., J. Wang, K. M. Chan, W. M. Tjia, W. Deng, X. Guan, H. Jian-dong, K. M. Li, P. Y. Chau, D. J. Chen, et al.

2005. Genomic instability in laminopathy-based premature aging. *Nature medicine*, 11(7):780.

Liu, G.-H., B. Z. Barkho, S. Ruiz, D. Diep, J. Qu, S.-L. Yang, A. D. Panopoulos,

K. Suzuki, L. Kurian, C. Walsh, et al.

2011. Recapitulation of premature aging with ipscs from hutchinson-gilford progeria syndrome. *Nature*, 472(7342):221.

Liu, L., T. H. Cheung, G. W. Charville, B. M. C. Hurgo, T. Leavitt, J. Shih, A. Brunet, and T. A. Rando

2013. Chromatin modifications as determinants of muscle stem cell quiescence and chronological aging. *Cell reports*, 4(1):189–204.

Liu, Y., A. Rusinol, M. Sinensky, Y. Wang, and Y. Zou

2006. Dna damage responses in progeroid syndromes arise from defective maturation of prelamin a. *J Cell Sci*, 119(22):4644–4649.

Loffredo, F. S., M. L. Steinhauser, S. M. Jay, J. Gannon, J. R. Pancoast, P. Yalamanchi, M. Sinha, C. Dall'Osso, D. Khong, J. L. Shadrach, et al.

2013. Growth differentiation factor 11 is a circulating factor that reverses age-related cardiac hypertrophy. *Cell*, 153(4):828–839.

López-Otín, C., M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer

2013. The hallmarks of aging. *Cell*, 153(6):1194–1217.

Lorenzini, A., T. Stamato, and C. Sell

2011. The disposable soma theory revisited: time as a resource in the theories of aging. *Cell Cycle*, 10(22):3853–3856.

Lu, A. T., E. Hannon, M. E. Levine, K. Hao, E. M. Crimmins, K. Lunnon, A. Kozlenkov, J. Mill, S. Dracheva, and S. Horvath

2016. Genetic variants near mlst8 and dhx57 affect the epigenetic age of the cerebellum. *Nature communications*, 7.

Lu, P. J., F. Giavazzi, T. E. Angelini, E. Zaccarelli, F. Jargstorff, A. B. Schofield, J. N. Wilking, M. B. Romanowsky, D. A. Weitz, R. Cerbino, et al.

2012. Characterizing concentrated, multiply scattering, and actively driven fluorescent systems with confocal differential dynamic microscopy. *Physical review letters*, 108(21):218103.

Luger, K., A. W. Mader, R. K. Richmond, D. F. Sargent, and T. J. Richmond

1997. Crystal structure of the nucleosome core particle at 2.8 angstrom resolution. *Nature*, 389(6648):251.

Luo, M., M. Jeong, D. Sun, H. J. Park, B. A. Rodriguez, Z. Xia, L. Yang, X. Zhang, K. Sheng, G. J. Darlington, et al.
2015. Long non-coding rnas control hematopoietic stem cell function. *Cell Stem Cell*, 16(4):426–438.

Lynch, M. D., A. J. Smith, M. De Gobbi, M. Flenley, J. R. Hughes, D. Vernimmen, H. Ayyub, J. A. Sharpe, J. A. Sloane-Stanley, L. Sutherland, et al.
2012. An interspecies analysis reveals a key role for unmethylated cpg dinucleotides in vertebrate polycomb complex recruitment. *The EMBO journal*, 31(2):317–329.

Macaulay, I. C., W. Haerty, P. Kumar, Y. I. Li, T. X. Hu, M. J. Teng, M. Goolam, N. Saurat, P. Coupland, L. M. Shirley, et al.
2015. G&t-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature methods*, 12(6):519–522.

Mack, S., H. Witt, R. Piro, L. Gu, S. Zuyderduyn, A. Stütz, X. Wang, M. Gallo, L. Garzia, K. Zayne, et al.
2014. Epigenomic alterations define lethal cimp-positive ependymomas of infancy. *Nature*, 506(7489):445.

Maegawa, S., G. Hinkal, H. S. Kim, L. Shen, L. Zhang, J. Zhang, N. Zhang, S. Liang, L. A. Donehower, and J.-P. J. Issa
2010. Widespread and tissue specific age-related dna methylation changes in mice. *Genome research*, 20(3):332–340.

Magenis, R. E., M. G. Brown, D. A. Lacy, S. Budden, S. LaFranchi, J. M. Opitz, J. F. Reynolds, and D. H. Ledbetter
1987. Is angelman syndrome an alternate result of del (15)(qllql3)? *American Journal of Medical Genetics Part A*, 28(4):829–838.

Marioni, R. E., S. Shah, A. F. McRae, B. H. Chen, E. Colicino, S. E. Harris, J. Gibson, A. K. Henders, P. Redmond, S. R. Cox, et al.
2015a. Dna methylation age of blood predicts all-cause mortality in later life. *Genome biology*, 16(1):25.

Marioni, R. E., S. Shah, A. F. McRae, S. J. Ritchie, G. Muniz-Terrera, S. E. Harris, J. Gibson, P. Redmond, S. R. Cox, A. Pattie, et al.
2015b. The epigenetic clock is correlated with physical and cognitive fitness in the

lothian birth cohort 1936. *International journal of epidemiology*, 44(4):1388–1396.

Martin-Herranz, D. E., A. J. M. Ribeiro, F. Krueger, J. M. Thornton, W. Reik, and T. M. Stubbs
2017. currbs: simple and robust evaluation of enzyme combinations for reduced representation approaches. *Nucleic Acids Research*.

Martins, A. C.
2011. Change and aging senescence as an adaptation. *PLoS One*, 6(9):e24328.

Mason, J. B.
2003. Biomarkers of nutrient exposure and status in one-carbon (methyl) metabolism. *The Journal of nutrition*, 133(3):941S–947S.

Maunakea, A. K., R. P. Nagarajan, M. Bilenky, T. J. Ballinger, C. D'Souza, S. D. Fouse, B. E. Johnson, C. Hong, C. Nielsen, Y. Zhao, et al.
2010. Conserved role of intragenic dna methylation in regulating alternative promoters. *Nature*, 466(7303):253.

McCay, C. M., F. Pope, W. Lunsford, G. Sperling, and P. Sambhavaphol
1957. Parabiosis between old and young rats. *Gerontologia*, 1:7–17.

McKenna, A., G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, and J. Shendure
2016. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298):aaf7907.

Medawar, P. B.
1952. *An unsolved problem of biology*. College.

Meissner, A., A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander, and R. Jaenisch
2005. Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. *Nucleic acids research*, 33(18):5868–5877.

Meng, H., G. Chen, H.-M. Gao, X. Song, Y. Shi, and L. Cao
2014. The emerging nexus of active dna demethylation and mitochondrial oxidative metabolism in post-mitotic neurons. *International journal of molecular sciences*, 15(12):22604–22625.

Michaloglou, C., L. C. Vredeveld, M. S. Soengas, C. Denoyelle, et al.
2005. Brafe600-associated senescence-like cell cycle arrest of human naevi. *Nature*,

436(7051):720.

Milo, R.

2013. What is the total number of protein molecules per cell volume? a call to rethink some published values. *Bioessays*, 35(12):1050–1055.

Mitteldorf, J.

2004. Ageing selected for its own sake. *Evolutionary Ecology Research*, 6(7):937–953.

Mitteldorf, J.

2006. Chaotic population dynamics and the evolution of ageing. *Evolutionary Ecology Research*, 8(3):561–574.

Mitteldorf, J.

2010. Evolutionary origins of aging. In *The Future of Aging*, Pp. 87–126. Springer.

Mitteldorf, J. and J. Pepper

2009. Senescence as an adaptation to limit the spread of disease. *Journal of Theoretical Biology*, 260(2):186–195.

Mitteldorf, J. and D. S. Wilson

2000. Population viscosity and the evolution of altruism. *Journal of Theoretical Biology*, 204(4):481–496.

Mohammed, H., I. Hernando-Herraez, A. Savino, A. Scialdone, I. Macaulay, C. Mulas, T. Chandra, T. Voet, W. Dean, J. Nichols, et al.

2017. Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell reports*, 20(5):1215–1228.

Mooijman, D., S. S. Dey, J.-C. Boisset, N. Crosetto, and A. Van Oudenaarden

2016. Single-cell 5hmc sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nature biotechnology*, 34(8):852–856.

Mosteiro, L., C. Pantoja, N. Alcazar, R. M. Marión, D. Chondronasiou, M. Rovira, P. J. Fernandez-Marcos, M. Muñoz-Martin, C. Blanco-Aparicio, J. Pastor, et al.

2016. Tissue damage and senescence provide critical signals for cellular reprogramming in vivo. *Science*, 354(6315):aaf4445.

Mulqueen, R. M., D. Pokholok, S. Norberg, A. J. Fields, D. Sun, K. A. Torkenczy, J. Shendure, C. Trapnell, B. J. O'Roak, Z. Xia, et al.

2017. Scalable and efficient single-cell dna methylation sequencing by combinatorial indexing. *bioRxiv*, P. 157230.

Nabel, C. S., H. Jia, Y. Ye, L. Shen, H. L. Goldschmidt, J. T. Stivers, Y. Zhang, and R. M. Kohli
2012. Aid/apobec deaminases disfavor modified cytosines implicated in dna demethylation. *Nature chemical biology*, 8(9):751–758.

Nagano, T., Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, and P. Fraser
2013. Single cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469).

Nishiyama, A., L. Yamaguchi, J. Sharif, Y. Johmura, T. Kawamura, K. Nakanishi, S. Shimamura, K. Arita, T. Kodama, F. Ishikawa, et al.
2013. Uhrf1-dependent h3k23 ubiquitylation couples maintenance dna methylation and replication. *Nature*, 502(7470):249.

Norvil, A. B., C. J. Petell, L. Alabdi, L. Wu, S. Rossie, and H. Gowher
2016. Dnmt3b methylates dna by a noncooperative mechanism, and its activity is unaffected by manipulations at the predicted dimer interface. *Biochemistry.*

Ocampo, A., P. Reddy, P. Martinez-Redondo, A. Platero-Luengo, F. Hatanaka, T. Hishida, M. Li, D. Lam, M. Kurita, E. Beyret, et al.
2016. In vivo amelioration of age-associated hallmarks by partial reprogramming. *Cell*, 167(7):1719–1733.

Oey, H., L. Isbel, P. Hickey, B. Ebaid, and E. Whitelaw
2015. Genetic and epigenetic variation among inbred mouse littermates: identification of inter-individual differentially methylated regions. *Epigenetics & chromatin*, 8(1):54.

Ohki, I., N. Shimotake, N. Fujita, J.-G. Jee, T. Ikegami, M. Nakao, and M. Shirakawa
2001. Solution structure of the methyl-cpg binding domain of human mbd1 in complex with methylated dna. *Cell*, 105(4):487–497.

Ohm, J. E., K. M. McGarvey, X. Yu, L. Cheng, K. E. Schuebel, L. Cope, H. P. Mohammad, W. Chen, V. C. Daniel, W. Yu, et al.
2007. A stem cell-like chromatin pattern may predispose tumor suppressor genes to dna hypermethylation and silencing in adult cancers. *Nature genetics*, 39(2):237.

Olguin, H. C. and B. B. Olwin

2004. Pax-7 up-regulation inhibits myogenesis and cell cycle progression in satellite cells: a potential mechanism for self-renewal. *Developmental biology*, 275(2):375–388.

O'Rourke, R. W.

2014. Metabolic thrift and the genetic basis of human obesity. *Annals of surgery*, 259(4):642.

Partridge, L., S. D. Pletcher, and W. Mair

2005. Dietary restriction, mortality trajectories, risk and damage. *Mechanisms of ageing and development*, 126(1):35–41.

Pasini, D., P. A. Cloos, J. Walfridsson, L. Olsson, J.-P. Bukowski, J. V. Johansen, M. Bak, N. Tommerup, J. Rappsilber, and K. Helin

2010. Jarid2 regulates binding of the polycomb repressive complex 2 to target genes in es cells. *Nature*, 464(7286):306.

Petkovich, D. A., D. I. Podolskiy, A. V. Lobanov, S.-G. Lee, R. A. Miller, and V. N. Gladyshev

2017. Using dna methylation profiling to evaluate biological age and longevity interventions. *Cell Metabolism*, 25(4):954–960.

Picelli, S., Å. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg

2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10(11):1096.

Pishel, I., D. Shytikov, T. Orlova, A. Peregudov, I. Artyuhov, and G. Butenko

2012. Accelerated aging versus rejuvenation of the immune system in heterochronic parabiosis. *Rejuvenation research*, 15(2):239–248.

Pott, S.

2017. Simultaneous measurement of chromatin accessibility, dna methylation, and nucleosome phasing in single cells. *eLife*, 6.

Price, F. D., J. Von Maltzahn, C. F. Bentzinger, N. A. Dumont, H. Yin, N. C. Chang, D. H. Wilson, J. Frenette, and M. A. Rudnicki

2014. Inhibition of jak-stat signaling stimulates adult satellite cell function. *Nature medicine*, 20(10):1174–1181.

Price, G. R. et al.

1970. Selection and covariance. *Nature*, 227:520–521.

R Core Team

2017. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rakyan, V. K., T. A. Down, S. Maslau, T. Andrew, T.-P. Yang, H. Beyan, P. Whittaker, O. T. McCann, S. Finer, A. M. Valdes, et al.

2010. Human aging-associated dna hypermethylation occurs preferentially at bivalent chromatin domains. *Genome research*, 20(4):434–439.

Rangaraju, S., G. M. Solis, R. C. Thompson, R. L. Gomez-Amaro, L. Kurian, S. E. Encalada, A. B. Niculescu III, D. R. Salomon, and M. Petrascheck

2015. Suppression of transcriptional drift extends c. elegans lifespan by postponing the onset of mortality. *Elife*, 4:e08833.

Rauch, T., H. Li, X. Wu, and G. P. Pfeifer

2006. Mira-assisted microarray analysis, a new technology for the determination of dna methylation patterns, identifies frequent methylation of homeodomain-containing genes in lung cancer cells. *Cancer research*, 66(16):7939–7947.

Reddington, J. P., S. M. Perricone, C. E. Nestor, J. Reichmann, N. A. Youngson, M. Suzuki, D. Reinhardt, D. S. Dunican, J. G. Prendergast, H. Mjoseng, et al.

2013. Redistribution of h3k27me3 upon dna hypomethylation results in de-repression of polycomb target genes. *Genome biology*, 14(3):R25.

Reimand, J., M. Kull, H. Peterson, J. Hansen, and J. Vilo

2007. g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids research*, 35(suppl_2):W193–W200.

Reither, S., F. Li, H. Gowher, and A. Jeltsch

2003. Catalytic mechanism of dna-(cytosine-c5)-methyltransferases revisited: covalent intermediate formation is not essential for methyl group transfer by the murine dnmt3a enzyme. *Journal of molecular biology*, 329(4):675–684.

Reizel, Y., A. Spiro, O. Sabag, Y. Skversky, M. Hecht, I. Keshet, B. P. Berman, and H. Cedar

2015. Gender-specific postnatal demethylation and establishment of epigenetic memory. *Genes & development*, 29(9):923–933.

Renault, V., L.-E. Thorne, P.-O. Eriksson, G. Butler-Browne, and V. Mouly
2002. Regenerative potential of human skeletal muscle during aging. *Aging cell*, 1(2):132–139.

Reznick, D. N., M. J. Bryant, D. Roff, C. K. Ghalambor, and D. E. Ghalambor
2004. Effect of extrinsic mortality on the evolution of senescence in guppies. *Nature*, 431(7012):1095.

Richmond, T. J. and C. A. Davey
2003. The structure of dna in the nucleosome core. *Nature*, 423(6936):145.

Rimmelé, P., C. L. Bigarella, R. Liang, B. Izac, R. Dieguez-Gonzalez, G. Barbet, M. Donovan, C. Brugnara, J. M. Blander, D. A. Sinclair, et al.
2014. Aging-like phenotype and defective lineage specification in sirt1-deleted hematopoietic stem and progenitor cells. *Stem cell reports*, 3(1):44–59.

Rocheteau, P., B. Gayraud-Morel, I. Siegl-Cachedenier, M. A. Blasco, and S. Tajbakhsh
2012. A subpopulation of adult skeletal muscle stem cells retains all template dna strands after cell division. *Cell*, 148(1):112–125.

Rossi, D. J., D. Bryder, J. M. Zahn, H. Ahlenius, R. Sonu, A. J. Wagers, and I. L. Weissman
2005. Cell intrinsic alterations underlie hematopoietic stem cell aging. *Proceedings of the National Academy of Sciences of the United States of America*, 102(26):9194–9199.

Ruckh, J. M., J.-W. Zhao, J. L. Shadrach, P. van Wijngaarden, T. N. Rao, A. J. Wagers, and R. J. Franklin
2012. Rejuvenation of regeneration in the aging central nervous system. *Cell stem cell*, 10(1):96–103.

Ryazanova, A. Y., L. Abrosimova, T. Oretskaya, and E. Kubareva
2012. Diverse domains of (cytosine-5)-dna methyltransferases: structural and functional characterization. In *Methylation-From DNA, RNA and Histones to Diseases and Treatment*. InTech.

Sambasivan, R., B. Gayraud-Morel, G. Dumas, C. Cimper, S. Paisant, R. G. Kelly, and S. Tajbakhsh
2009. Distinct regulatory cascades govern extraocular and pharyngeal arch muscle

progenitor cell fates. *Developmental cell*, 16(6):810–821.

Sambasivan, R., R. Yao, A. Kissenpfennig, L. Van Wittenberghe, A. Paldi, B. Gayraud-Morel, H. Guenou, B. Malissen, S. Tajbakhsh, and A. Galy
2011. Pax7-expressing satellite cells are indispensable for adult skeletal muscle regeneration. *Development*, 138(17):3647–3656.

Santos, F., B. Hendrich, W. Reik, and W. Dean
2002. Dynamic reprogramming of dna methylation in the early mouse embryo. *Developmental biology*, 241(1):172–182.

Santos, F., V. Zakhartchenko, M. Stojkovic, A. Peters, T. Jenuwein, E. Wolf, W. Reik, and W. Dean
2003. Epigenetic marking correlates with developmental potential in cloned bovine preimplantation embryos. *Current Biology*, 13(13):1116–1121.

Sardo, V. L., W. Ferguson, G. A. Erikson, E. J. Topol, K. K. Baldwin, and A. Torkamani
2017. Influence of donor age on induced pluripotent stem cells. *Nature biotechnology*, 35(1):69–74.

Schillebeeckx, M., A. Schrade, A.-K. Löbs, M. Pihlajoki, D. B. Wilson, and R. D. Mitra
2013. Laser capture microdissection–reduced representation bisulfite sequencing (lcm-rrbs) maps changes in dna methylation associated with gonadectomy-induced adrenocortical neoplasia in the mouse. *Nucleic acids research*, 41(11):e116–e116.

Schindelin, J., I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, et al.
2012. Fiji: an open-source platform for biological-image analysis. *Nature methods*, 9(7):676–682.

Schlesinger, Y., R. Straussman, I. Keshet, S. Farkash, M. Hecht, J. Zimmerman, E. Eden, Z. Yakhini, E. Ben-Shushan, B. E. Reubinoff, et al.
2007. Polycomb-mediated methylation on lys27 of histone h3 pre-marks genes for de novo methylation in cancer. *Nature genetics*, 39(2):232.

Schrack, J. A., N. D. Knuth, E. M. Simonsick, and L. Ferrucci
2014. “ideal” aging is associated with lower resting metabolic rate: the baltimore longitudinal study of aging. *Journal of the American Geriatrics Society*, 62(4):667–

672.

Schulman, E. et al.

1956. Dendroclimatic changes in semiarid america. *Dendroclimatic changes in semiarid America.*

Scialdone, A., K. N. Natarajan, L. R. Saraiva, V. Proserpio, S. A. Teichmann, O. Stegle, J. C. Marioni, and F. Buettner

2015. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61.

Seisenberger, S., J. R. Peat, T. A. Hore, F. Santos, W. Dean, and W. Reik

2013. Reprogramming dna methylation in the mammalian life cycle: building and breaking epigenetic barriers. *Phil. Trans. R. Soc. B*, 368(1609):20110330.

Serrano, M., A. W. Lin, M. E. McCurrach, D. Beach, and S. W. Lowe

1997. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16 ink4a. *Cell*, 88(5):593–602.

Shao, Z., F. Raible, R. Mollaaghababa, J. R. Guyon, C.-t. Wu, W. Bender, and R. E. Kingston

1999. Stabilization of chromatin structure by prc1, a polycomb complex. *Cell*, 98(1):37–46.

Sharif, J., T. A. Endo, M. Nakayama, M. M. Karimi, M. Shimada, K. Katsuyama, P. Goyal, J. Brind'Amour, M.-A. Sun, Z. Sun, et al.

2016. Activation of endogenous retroviruses in dnmt1-/- escs involves disruption of setdb1-mediated repression by np95 binding to hemimethylated dna. *Cell Stem Cell*, 19(1):81–94.

Sharif, J., M. Muto, S.-i. Takebayashi, I. Suetake, A. Iwamatsu, T. A. Endo, J. Shinga, Y. Mizutani-Koseki, T. Toyoda, K. Okamura, et al.

2007. The sra protein np95 mediates epigenetic inheritance by recruiting dnmt1 to methylated dna. *Nature*, 450(7171):908.

Shen, J.-C., J.-M. Zingg, A. S. Yang, C. Schmutte, and P. A. Jones

1995. A mutant hpall methyltransferase functions as a mutator enzyme. *Nucleic acids research*, 23(21):4275–4282.

Shumaker, D. K., T. Dechat, A. Kohlmaier, S. A. Adam, M. R. Bozovsky, M. R. Erdos,

M. Eriksson, A. E. Goldman, S. Khuon, F. S. Collins, et al.

2006. Mutant nuclear lamin a leads to progressive alterations of epigenetic control in premature aging. *Proceedings of the National Academy of Sciences*, 103(23):8703–8708.

Siddique, A. N., R. Z. Jurkowska, T. P. Jurkowski, and A. Jeltsch

2011. Auto-methylation of the mouse dna-(cytosine c5)-methyltransferase dnmt3a at its active site cysteine residue. *The FEBS journal*, 278(12):2055–2063.

Sierro, N., J. N. Battey, S. Ouadi, N. Bakaher, L. Bovet, A. Willig, S. Goepfert, M. C. Peitsch, and N. V. Ivanov

2014. The tobacco genome sequence and its comparison with those of tomato and potato. *Nature communications*, 5.

Sinha, M., Y. C. Jang, J. Oh, D. Khong, E. Y. Wu, R. Manohar, C. Miller, S. G. Regalado, F. S. Loffredo, J. R. Pancoast, et al.

2014. Restoring systemic gdf11 levels reverses age-related dysfunction in mouse skeletal muscle. *Science*, 344(6184):649–652.

Skulachev, V. P.

2001. The programmed death phenomena, aging, and the samurai law of biology. *Experimental gerontology*, 36(7):995–1024.

Slieker, R. C., M. van Iterson, R. Luijk, M. Beekman, D. V. Zhernakova, M. H. Moed, H. Mei, M. Van Galen, P. Deelen, M. J. Bonder, et al.

2016. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome biology*, 17(1):191.

Smallwood, S. A., H. J. Lee, C. Angermueller, F. Krueger, H. Saadeh, J. Peat, S. R. Andrews, O. Stegle, W. Reik, and G. Kelsey

2014. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, 11(8):817–820.

Smith, J. M. and G. A. Parker

1976. The logic of asymmetric contests. *Animal behaviour*, 24(1):159–175.

Sohal, R. S. and M. J. Forster

2014. Caloric restriction and the aging process: a critique. *Free Radical Biology and Medicine*, 73:366–382.

Song, J., O. Rechkoblit, T. H. Bestor, and D. J. Patel

2011. Structure of dnmt1-dna complex reveals a role for autoinhibition in maintenance dna methylation. *Science*, 331(6020):1036–1040.

Soojin, V. Y.

2012. Birds do it, bees do it, worms and ciliates do it too: Dna methylation from unexpected corners of the tree of life. *Genome biology*, 13(10):174.

Sousa-Victor, P., S. Gutarra, L. Garcia-Prat, J. Rodriguez-Ubreva, L. Ortet, V. Ruiz-Bonilla, M. Jardi, E. Ballestar, S. Gonzalez, A. L. Serrano, et al.

2014. Geriatric muscle stem cells switch reversible quiescence into senescence. *Nature*, 506(7488):316.

Spiers, H., E. Hannon, S. Wells, B. Williams, C. Fernandes, and J. Mill

2016. Age-associated changes in dna methylation across multiple tissues in an inbred mouse model. *Mechanisms of ageing and development*, 154:20–23.

Stölzel, F., M. Brosch, S. Horvath, M. Kramer, C. Thiede, M. von Bonin, O. Ammerpohl, M. Middeke, J. Schetelig, G. Ehninger, et al.

2017. Dynamics of epigenetic age following hematopoietic stem cell transplantation. *haematologica*, 102(8):e321–e323.

Stubbs, T. M., M. J. Bonder, A.-K. Stark, F. Krueger, F. von Meyenn, O. Stegle, and W. Reik

2017. Multi-tissue dna methylation age predictor in mouse. *Genome Biology*, 18(1):68.

Stuelsatz, P. and Z. Yablonka-Reuveni

2016. Isolation of mouse periocular tissue for histological and immunostaining analyses of the extraocular muscles and their satellite cells. *Skeletal Muscle Regeneration in the Mouse: Methods and Protocols*, Pp. 101–127.

Suetake, I., F. Shinozaki, J. Miyagawa, H. Takeshima, and S. Tajima

2004. Dnmt3l stimulates the dna methylation activity of dnmt3a and dnmt3b through a direct interaction. *Journal of Biological Chemistry*, 279(26):27816–27823.

Sun, D., M. Luo, M. Jeong, B. Rodriguez, Z. Xia, R. Hannah, H. Wang, T. Le, K. F. Faull, R. Chen, et al.

2014. Epigenomic profiling of young and aged hscs reveals concerted changes during aging that reinforce self-renewal. *Cell stem cell*, 14(5):673–688.

Swindell, W. R.

2012. Dietary restriction in rats and mice: a meta-analysis and review of the evidence for genotype-dependent effects on lifespan. *Ageing research reviews*, 11(2):254–270.

Tahiliani, M., K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, et al.

2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian dna by mll partner tet1. *Science*, 324(5929):930–935.

Takahashi, K. and S. Yamanaka

2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126(4):663–676.

Tanas, A. S., M. E. Borisova, E. B. Kuznetsova, V. V. Rudenko, K. O. Karandasheva, M. V. Nemtsova, V. L. Izhevskaya, O. A. Simonova, S. S. Larin, D. V. Zaletaev, et al.

2017. Rapid and affordable genome-wide bisulfite dna sequencing by xmai-reduced representation bisulfite sequencing. *Epigenomics*, (00).

Taylor, P. D.

1992. Altruism in viscous populations—an inclusive fitness model. *Evolutionary ecology*, 6(4):352–356.

Tchkonia, T., Y. Zhu, J. Van Deursen, J. Campisi, and J. L. Kirkland

2013. Cellular senescence and the senescent secretory phenotype: therapeutic opportunities. *The Journal of clinical investigation*, 123(3):966.

Teschendorff, A. E., U. Menon, A. Gentry-Maharaj, S. J. Ramus, D. J. Weisenberger, H. Shen, M. Campan, H. Noushmehr, C. G. Bell, A. P. Maxwell, et al.

2010. Age-dependent dna methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome research*, 20(4):440–446.

Thinggaard, M., M. McGue, B. Jeune, M. Osler, J. W. Vaupel, and K. Christensen

2016. Survival prognosis in very old adults. *Journal of the American Geriatrics Society*, 64(1):81–88.

Thompson, M. J. et al.

2017. An epigenetic aging clock for dogs and wolves. *Aging (Albany NY)*, 9(3):1055.

Thum, T., C. Gross, J. Fiedler, T. Fischer, S. Kissler, M. Bussen, P. Galuppo, S. Just, W. Rottbauer, S. Frantz, et al.
2008. Microrna-21 contributes to myocardial disease by stimulating map kinase signalling in fibroblasts. *Nature*, 456(7224):980.

Tierney, M. T., T. Aydogdu, D. Sala, B. Malecova, S. Gatto, P. L. Puri, L. Latella, and A. Sacco
2014. Stat3 signaling controls satellite cell expansion and skeletal muscle repair. *Nature medicine*, 20(10):1182–1186.

Travis, J. M.
2004. The evolution of programmed death in a spatially structured population. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 59(4):B301–B305.

Ueda, J., K. Maehara, D. Mashiko, T. Ichinose, T. Yao, M. Hori, Y. Sato, H. Kimura, Y. Ohkawa, and K. Yamagata
2014. Heterochromatin dynamics during the differentiation process revealed by the dna methylation reporter mouse, methylro. *Stem cell reports*, 2(6):910–924.

Ulrey, C. L., L. Liu, L. G. Andrews, and T. O. Tollefsbol
2005. The impact of metabolism on dna methylation. *Human molecular genetics*, 14(suppl_1):R139–R147.

van der Maaten, L. and G. Hinton
2008. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Van Rossum, G. and F. L. Drake
2011. *The python language reference manual.* Network Theory Ltd.

Varley, K. E., J. Gertz, K. M. Bowling, S. L. Parker, T. E. Reddy, F. Pauli-Behn, M. K. Cross, B. A. Williams, J. A. Stamatoyannopoulos, G. E. Crawford, et al.
2013. Dynamic dna methylation across diverse human cell lines and tissues. *Genome research*, 23(3):555–567.

Versteeg, R.
2014. Tumours outside the mutation box. *Nature*, 506(7489):438–440.

Vidal-Bralo, L., Y. Lopez-Golan, A. Mera-Varela, I. Rego-Perez, S. Horvath, Y. Zhang, Á. del Real, G. Zhai, F. J. Blanco, J. A. Riancho, et al.

2016. Specific premature epigenetic aging of cartilage in osteoarthritis. *Aging (Albany NY)*, 8(9):2222.

Vilkaitis, G., I. Suetake, S. Klimašauskas, and S. Tajima

2005. Processive methylation of hemimethylated cpg sites by mouse dnmt1 dna methyltransferase. *Journal of Biological Chemistry*, 280(1):64–72.

Villeda, S. A., J. Luo, K. I. Mosher, B. Zou, M. Britschgi, G. Bieri, T. M. Stan, N. Fainberg, Z. Ding, A. Eggel, et al.

2011. The ageing systemic milieu negatively regulates neurogenesis and cognitive function. *Nature*, 477(7362):90–94.

Villeda, S. A., K. E. Plambeck, J. Middeldorp, J. M. Castellano, K. I. Mosher, J. Luo, L. K. Smith, G. Bieri, K. Lin, D. Berdnik, et al.

2014. Young blood reverses age-related impairments in cognitive function and synaptic plasticity in mice. *Nature medicine*, 20(6):659–663.

von Maltzahn, J., A. E. Jones, R. J. Parks, and M. A. Rudnicki

2013. Pax7 is critical for the normal function of satellite cells in adult skeletal muscle. *Proceedings of the National Academy of Sciences*, 110(41):16474–16479.

von Meyenn, F., M. Iurlaro, E. Habibi, N. Q. Liu, A. Salehzadeh-Yazdi, F. Santos, E. Petrini, I. Milagre, M. Yu, Z. Xie, et al.

2016. Impairment of dna methylation maintenance is the main cause of global demethylation in naive embryonic stem cells. *Molecular cell*, 62(6):848–861.

Waddington, C. H.

1953. Genetic assimilation of an acquired character. *Evolution*, 7(2):118–126.

Wagers, A. J., R. I. Sherwood, J. L. Christensen, and I. L. Weissman

2002. Little evidence for developmental plasticity of adult hematopoietic stem cells. *Science*, 297(5590):2256–2259.

Wang, T., B. Tsui, J. F. Kreisberg, N. A. Robertson, A. M. Gross, M. K. Yu, H. Carter, H. M. Brown-Borg, P. D. Adams, and T. Ideker

2017. Epigenetic aging signatures in mice livers are slowed by dwarfism, calorie restriction and rapamycin treatment. *Genome biology*, 18(1):57.

Waterston, R. H. and L. Pachter

2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562.

Weidner, C., P. Ziegler, M. Hahn, T. Brümmendorf, A. Ho, P. Dreger, and W. Wagner
2015. Epigenetic aging upon allogeneic transplantation: the hematopoietic niche does not affect age-associated dna methylation. *Leukemia*, 29(4):985.

Weidner, C. I., Q. Lin, C. M. Koch, L. Eisele, F. Beier, P. Ziegler, D. O. Bauerschlag, K.-H. Jöckel, R. Erbel, T. W. Mühleisen, et al.
2014. Aging of blood can be tracked by dna methylation changes at just three cpg sites. *Genome biology*, 15(2):R24.

Weismann, A.
1889. Essays on heredity and kindred biological subjects.

Widschwendter, M., H. Fiegl, D. Egle, E. Mueller-Holzner, G. Spizzo, C. Marth, D. J. Weisenberger, M. Campan, J. Young, I. Jacobs, et al.
2007. Epigenetic stem cell signature in cancer. *Nature genetics*, 39(2):157.

Willbanks, A., M. Leary, M. Greenshields, C. Tyminski, S. Heerboth, K. Lapinska, K. Haskins, and S. Sarkar
2016. The evolution of epigenetics: from prokaryotes to humans and its biological consequences. *Genetics & epigenetics*, 8:25.

Williams, G. C.
2001. Pleiotropy, natural selection, and the evolution of senescence. *Science's SAGE KE*, 2001(1):13.

Wilson, D. S., G. Pollock, and L. A. Dugatkin
1992. Can altruism evolve in purely viscous populations? *Evolutionary ecology*, 6(4):331–341.

Wilson, L. G., V. A. Martinez, J. Schwarz-Linek, J. Tailleur, G. Bryant, P. Pusey, and W. C. Poon
2011. Differential dynamic microscopy of bacterial motility. *Physical review letters*, 106(1):018101.

Woodberry, O. G., K. B. Korb, and A. E. Nicholson
2007. A simulation study of the evolution of ageing. *Evolutionary Ecology Research*, 9(7):1077–1096.

Woodworth, M. B., K. M. Girskis, and C. A. Walsh
2017. Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nature Reviews Genetics*, 18(4):230–244.

Wright, D. E., A. J. Wagers, A. P. Gulati, F. L. Johnson, and I. L. Weissman
2001. Physiological migration of hematopoietic stem and progenitor cells. *Science*, 294(5548):1933–1936.

Wu, X. and Y. Zhang
2017. Tet-mediated active dna demethylation: mechanism, function and beyond. *Nature Reviews Genetics*, 18(9):517–534.

Xia, L., S. Ma, Y. Zhang, T. Wang, M. Zhou, Z. Wang, and J. Zhang
2015. Daily variation in global and local dna methylation in mouse livers. *PloS one*, 10(2):e0118101.

Xiao, Y., B. Word, A. Starlard-Davenport, A. Haefele, B. D. Lyn-Cook, and G. Hammons
2008. Age and gender affect dnmt3a and dnmt3b expression in human liver. *Cell biology and toxicology*, 24(3):265–272.

Xu, X., Y. Hou, X. Yin, L. Bao, A. Tang, L. Song, F. Li, S. Tsang, K. Wu, H. Wu, et al.
2012. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, 148(5):886–895.

Yamagata, K.
2010. Dna methylation profiling using live-cell imaging. *Methods*, 52(3):259–266.

Yamazaki, T., K. Yamagata, and T. Baba
2007. Time-lapse and retrospective analysis of dna methylation in mouse preimplantation embryos by live cell imaging. *Developmental biology*, 304(1):409–419.

Yang, H., H. Lin, H. Xu, L. Zhang, L. Cheng, B. Wen, J. Shou, K. Guan, Y. Xiong, and D. Ye
2014. Tet-catalyzed 5-methylcytosine hydroxylation is dynamically regulated by metabolites. *Cell research*, 24(8):1017–1020.

Yang, J., L. Lior-Hoffmann, S. Wang, Y. Zhang, and S. Broyde
2013. Dna cytosine methylation: structural and thermodynamic characterization of

the epigenetic marking mechanism. *Biochemistry*, 52(16):2828–2838.

Yang, J.-N.

2013. Viscous populations evolve altruistic programmed ageing in ability conflict in a changing environment. *Evolutionary Ecology Research*, 15(5):527–543.

Zammit, P. S., J. P. Golding, Y. Nagata, V. Hudon, T. A. Partridge, and J. R. Beauchamp

2004. Muscle satellite cells adopt divergent fates. *J Cell Biol*, 166(3):347–357.

Zhang, C., Y. Hoshida, and K. C. Sadler

2016a. Comparative epigenomic profiling of the dna methylome in mouse and zebrafish uncovers high interspecies divergence. *Frontiers in genetics*, 7.

Zhang, H. and J.-K. Zhu

2011. Rna-directed dna methylation. *Current opinion in plant biology*, 14(2):142–147.

Zhang, R., L. Liu, Y. Yao, F. Fei, F. Wang, Q. Yang, Y. Gui, and X. Wang

2017. High resolution imaging of dna methylation dynamics using a zebrafish reporter. *Scientific Reports*, 7.

Zhang, W., T. D. Spector, P. Deloukas, J. T. Bell, and B. E. Engelhardt

2015. Predicting genome-wide dna methylation using methylation marks, genomic position, and dna regulatory elements. *Genome biology*, 16(1):14.

Zhang, X., J. Su, M. Jeong, M. Ko, Y. Huang, H. J. Park, A. Guzman, Y. Lei, Y.-H. Huang, A. Rao, et al.

2016b. Dnmt3a and tet2 compete and cooperate to repress lineage-specific transcription factors in hematopoietic stem cells. *Nature genetics*, 48(9):1014–1023.

Zhu, C., Y. Gao, H. Guo, B. Xia, J. Song, X. Wu, H. Zeng, K. Kee, F. Tang, and C. Yi

2017. Single-cell 5-formylcytosine landscapes of mammalian early embryos and escs at single-base resolution. *Cell Stem Cell*, 20(5):720–731.

Zhu, J., M. Adli, J. Y. Zou, G. Verstappen, M. Coyne, X. Zhang, T. Durham, M. Miri, V. Deshpande, P. L. De Jager, et al.

2013. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell*, 152(3):642–654.

Zhu, J.-K.

2009. Active dna demethylation mediated by dna glycosylases. *Annual review of* *genetics*, 43:143–166.

Zou, X., W. Ma, I. A. Solov'yov, C. Chipot, and K. Schulten
2011. Recognition of methylated dna through methyl-cpg binding domain proteins. *Nucleic acids research*, 40(6):2747–2758.

Nomenclature

2HG	2-hydroxyglutarate
2OG	2-oxoglutarate
5caC	DNA carboxycytosin
5fC	DNA formylcytosin
5hmC	DNA hydroxymethylation
5mC	DNA methylation
AID	Activation-Induced cytidine Deaminase
AMAR	Apparent Methylomic Aging Rate
APOBEC3G	Apolipoprotein B mRNA Editing enzyme, Catalytic polypeptide-like 3G
BABR	Babraham
BAH	Bromo-Adjacent Homology
BER	Base Excision Repair
BMI	Body Mass Index
bp	base pairs
BS	Blocking Solution
BSA	Bovine Serum Albumin
CGI	CpG Island
ChIP	Chromatin Immuno-Precipitation
CLEVER-Seq	Chemical-Labeling-Enabled C-to-T conversion Sequencing
CMOS	Complementary Metal-Oxide Semiconductor
CpG	cytosine-phosphate-guanine
CPM	Counts Per Million
DDM	Differential Dynamic Microscopy
DICF	Differential Intensity Correlation Function
DMEM	Dulbecco's Modified Eagle's Medium
DMR	Differentially Methylated Region
DMSO	Dimethyl sulfoxide
DNMT	DNA Methyltransferase

DNMT−TKO	DNA Methyltransferase Triple Knock-Out
DNMT−TKO−C	DNA Methyltransferase Triple Knock-Out Control
DNMT1−iKO	DNA Methyltransferase 1 inducible Knock-Out
dNTP	deoxynucleotide triphosphate
dsOligo	double-stranded oligonucleotide
EED	Embryonic Ectoderm Development
eGFP	enhanced Green Fluorescent Protein
ELOVL2	Elongation of very long chain fatty acids protein 2
ENCODE	Encyclopedia of DNA Elements
ESC	Embryonic Stem Cell
EZH	Enhancer of Zeste
FACS	Fluorescence-Activated Cell Sorting
FBS	Fetal Bovine Serum
FCS	Fetal Calf Serum
FDR	False Discovery Rate
FFT	Fast Fourier Transform
FIJI	FIJI Is Just ImageJ
FITC	Fluorescein Isothiocyanate
FRAP	Fluorescence Recovery After Photobleaching
FRET	Fluorescence Resonance Energy Transfer
GDF-11	Growth Differentiation Factor 11
GEO	Gene Expression Omnibus
GFP	Green Fluorescent Protein
GHR-KO	Growth Hormone Receptor Knock-Out
GLMNET	Lasso and Elastic−Net Regularized Generalized Linear Models
GMM	Gaussian Mixture Model
GO	Gene Ontology
H3K27me3	Histone H3 Lysine K27 trimethylation
H3K4me	Histone H3 Lysine K4 methylation
HD	Hamming Distance
HGPS	Hutchinson-Gilford progeria syndrome
HIV-1	Human Immunodeficiency Virus-1
HSC	Haematopoietic Stem Cell
IAP	Intra-cisternal A-type Particle
IF	Immunofluorescence
iPSC	induced Pluripotent Stem Cell
KDM2B	Lysine Demethylase 2B

LB	Lysogeny broth
LIF	leukaemia inhibitory factor
LINE	Long Interspersed Nuclear Elements
MACS2	Model−based Analysis of ChIP-Seq 2
MAD	Median Absolute Deviation
MAE	Median Absolute Error
MBD	Methyl Binding Domain
MEF	Mouse Embryonic Fibroblast
meQTL	methylation quantitative trait loci
mESC	mouse Embryonic Stem Cell
mRNA	messenger Ribonucleic Acid
MuSC	Muscle Satellite Cell
NER	Nucleotide Excision Repair
NGS	Next Generation Sequencing
NLS	Nuclear Localisation Signal
p16$^{\mathrm{INK4A}}$	cyclin-dependent kinase inhibitor 2A
PBAT	Post-Bisulfite Adapter Tagging
PBS	Phosphate-Buffered Saline
PC	Principle Component
PCA	Principle Component Analysis
PcG	Polycomb Group
PCNA	Proliferating Cell Nuclear Antigen
PCR	Polymerase Chain Reaction
PE	Phycoerythrin
PFA	Paraformaldehyde
PRC	Polycomb Repressive Complex
PRD	PCNA Recognition Domain
PWM	Position Weight Matrix
QC	Quality Control
QSC	Quiescent Satellite Cell
RFP	Red Fluorescent Protein
RING	Really Interesting New Gene
ROI	Region Of Interest
RRBS	Reduced Representation Bisulfite Sequencing
SA-β-Gal	Senescence-Associated beta-Galactosidase
SAH	S-Adenosyl Homocysteine
SAM	S-Adenosyl Methionine

SASP	Senescence Associated Secretory Phenotype
scAba-Seq	single-cell AbaSI coupled with sequencing
scATAC-Seq	single-cell Assay for Transposase-Accessible Chromatin using sequencing
scBS-Seq	Single-cell bisulfite sequencing
scG&T-Seq	single-cell Genome and Transcriptome Sequencing
scM&T-Seq	single-cell Methylome and Transcriptome Sequencing
scNMT-Seq	single-cell Nucleosome, Methylation and Transcription Sequencing
scNOMe-Seq	single-cell Nucleosome Occupancy and Methylome sequencing
SMUG	Single-strand selective Monofunctional Uracil-DNA Glycosylase
SNP	Single Nucleotide Polymorphism
SOC	Super Optimal broth with Catabolite repression
SPRY1	Sprouty1
SRA	SET and Ring finger Associated
SRA	Sequence Read Archive
SUZ12	Suppressor of Zeste 12 homolog
TA	Tibialis anterior
TA-Hi MuSC	*Tibialis anterior* Pax7-highly-expressing muscle satellite cells
TCGA	The Cancer Genome Atlas
TDG	Thymine-DNA Glycosylase
TET	methylcytosine dioxygenase
TRD	Target Recognition Domain
tSNE	t-Distributed Stochastic Neighbor Embedding
TSS	Transcription Start Site
UHRF1	Ubiquitin-like, containing PHD and RING finger domains, 1
UMI	Unique Molecular Identifier
UNG	Uracil-DNA Glycosylase
WGBS	Whole Genome Bisulfite Sequencing