Title	Characterisation of core histone sequences and nuclear mobility using a reproducible research approach
Author(s)	Susano Pinto, David Miguel
Publication Date	2017-09-19
Item record	http://hdl.handle.net/10379/6841

Characterisation of core histone sequences and nuclear mobility using a reproducible research approach

A thesis presented for the degree of Doctor of Philosophy by

David Miguel Susano Pinto

Supervisor:
Dr. Andrew Flaus

Co-supervisor:
Prof. Kevin Sullivan

Discipline of Biochemistry
School of Natural Sciences
National University of Ireland, Galway

August 2017

Contents

Contents iii

List of Figures ix

List of Tables xi

Acknowledgements xiii

Summary xv

1 Introduction 1

1.1 Chromatin 1

Histones 3

Chromatin Dynamics 8

1.2 Fluorescence microscopy 12

Fluorescent proteins 12

FRAP 14

1.3 Bioinformatics 17

Free Software and Scientific Advances 18

Free Software Projects in Academia 18

Meta-research 19

1.4 Aims and Objectives 20

2 Materials and Methods 21

2.1 Source of chemicals, enzymes, and solutions 21

2.2 DNA methods 21

Bacterial cultures 21

Preparation of competent bacteria 23

Transformation of competent cells 23
Plasmid DNA preparation 23
Phenol:chloroform extraction 24
Ethanol precipitation 24
Agarose gel electrophoresis 24
DNA sequencing and oligonucleotide preparation 24
Polymerase Chain Reaction 25
Plasmid construction 27
2.3 Protein methods 31
Western blotting 31
2.4 Cell methods 32
Cell culture 32
Cell stock storage 32
Whole cell extract 33
Genomic DNA extraction 33
Viable cells count 33
Kill curve 34
Transfection by lipofection 34
Transfection by electroporation 34
Generation of stable cell lines 35
Fixation and staining 35
Flow cytometry 35
2.5 Software used 36
2.6 Source Code and Data 37
2.7 Developed software 37

3 Human Canonical Core Histone Catalogue 39
3.1 Introduction 40
Biochemical perspective 40
Functional perspective 41
Genomic perspective 44
Cataloguing of canonical core histone diversity 44
3.2 Histone genes 45
Canonical core histone gene nomenclature 45
Histone gene clustering 47
Histone gene sequences 48
Histone variant genes 49

3.3 Histone transcripts 49
3.4 Histone proteins 51
Canonical H2A isoforms 52
Canonical H2B isoforms 54
Canonical H3 isoforms 54
Canonical H4 isoforms 56
3.5 Reproducible research 56
3.6 Conclusion 59
3.7 Materials and methods 60

4 Application of FRAP to Histones in Human Cell Nuclei 61
4.1 Introduction 62
Nucleosome dynamics and histone SIN mutants 62
Fluorescence Recovery After Photobleaching 63
FRAP measurements of histones 64
Aims and objectives 65
4.2 Materials and Methods 65
4.3 Results 66
Inhibition of cell motility 68
Image-based tracking of cell movement 70
Chromatin movement within nuclei 70
Selection of G_1 cells 71
Chromatin movement observed by photoactivation 72
4.4 Discussion 74
4.5 Conclusion 78

5 Software Tools for Image and Sequence Analysis 81
5.1 Introduction 82
5.2 GNU Octave 83
Reading and Writing of Image Files 84
Bio-Formats and Octave Java interface 86
5.3 Octave Forge Image package 86
Image Thresholding Algorithms 87
Mathematical Morphology 89
Image Regions of Interest 89
5.4 Octave FRAP package 90
5.5 BioPerl 92

Dist::Zilla and Pod::Weaver 92

Bio-EUtilities 94

Debian packaging 95

5.6 Build systems for reproducible research 97

5.7 Discussion 98

5.8 Conclusion 100

6 Discussion and Future Perspectives 101

6.1 Data in Flux 102

6.2 Data not in Flux 104

6.3 Runtime Environment 105

6.4 Continuous Delivery 106

6.5 Version Control 107

6.6 Software Engineering and Biology 108

6.7 Conclusion 109

A Human Canonical Core Histone Catalogue 111

B Mouse Canonical Core Histone Catalogue 127

C List of primers 149

D bp_genbank_ref_extractor manual 155

D.1 NAME 155

D.2 VERSION 155

D.3 SYNOPSIS 155

D.4 DESCRIPTION 155

D.5 OPTIONS 156

D.6 EXAMPLES 158

D.7 NON-BUGS 160

D.8 NOTES ON USAGE 160

E Structure and function of histone H2AX 163

E.1 Introduction 164

Chromatin Structure and Genome Stability 164

H2AX and DNA repair 165

E.2 Structural Properties of H2AX 166

Definition of H2AX 166

H2AX Gene 168

H2AX Transcripts 169
H2AX Protein 170
H2AX Post-Translational Modifications 173
H2AX Distribution in Chromatin 176
E.3 Functional Roles of H2AX 181
Initiation of H2AX Phosphorylation as a Reporter of DSB Events 181
Spreading of H2AX Phosphorylation as a Damage Signal Amplifier 182
γH2AX and Chromatin Structural Remodelling 183
γH2AX and Localisation of DSB Repair Proteins 184
γH2AX and Maintenance of Proximity of Break Ends . . . 186
γH2AX and Complementary Damage Signalling Via Ubiquitylation 186
E.4 Conclusion 186
Bibliography **189**

List of Figures

1.1 Overview of the chromatin hierarchical structure 2

1.2 Classic "beads on a string" view of chromatin 3

1.3 Structure of the nucleosome core particle 5

1.4 Histone fold domain organisation 6

1.5 Concept of a FRAP experiment 16

2.1 Fluorescent intensity of HeLa cell populations by flow citometry 36

3.1 Histone gene nomenclature 46

3.2 Stem-loop and HDE alignment 51

4.1 Circle FRAP of H2B–EGFP in HeLa 67

4.2 Movement of confluent HeLa cells during a FRAP experiment 68

4.3 Movement of confluent primary horse cells during a circle FRAP experiment 69

4.4 Automatic tracking and alignment of moving cells 70

4.5 Movement of nuclear features and bleach spot distortion . . . 71

4.6 Photoactivation experiment demonstrating complex chromatin movement 73

4.7 HeLa cell cycle phases and timing 76

5.1 Automatic selection of regions for FRAP analysis 91

5.2 Frapinator visual log files for batch processing 93

A.1 Human H2A CDS slignment 115

A.2 Human H2B CDS slignment 116

A.3 Human H3 CDS slignment 117

A.4 Human H4 CDS slignment 118

B.1 Mouse H2A CDS alignment 140
B.2 Mouse H2B CDS alignment 141
B.3 Mouse H3 CDS alignment 142
B.4 Mouse H4 CDS alignment 143
B.5 Stem-loop and HDE alignment 147

E.1 Alignment of *Caenorhabditis* CENP-A homologues 168
E.2 H2AFX gene and transcripts 169
E.3 Sequence logo of human canonical H2A isoforms and H2AX . 172
E.4 Nucleosome structure highlighting differences between H2A and H2AX 173
E.5 Structural environment of H2A residues that differ from H2AX 174
E.6 H2AX distribution in the chromatin 177
E.7 Simulations of H2AX spacing distributions 179

List of Tables

1.1 Historical perspective of nucleosome structure 9

2.1 Commonly used buffers and media 22
2.2 PCR mixtures and conditions used 26

3.1 Properties distinguishing canonical and variant core histone proteins 41
3.2 Count of human canonical core histone coding genes and pseudogenes by histone cluster and type 41
3.3 Terminology describing histone variation 42
3.4 Canonical H2A encoded protein isoforms 53
3.5 Canonical H2B encoded protein isoforms 55
3.6 Canonical H3 encoded protein isoforms 56
3.7 Canonical H4 encoded protein isoforms 57

5.1 New or improved functions in GNU Octave 85
5.2 New or improved functions in the Octave Forge Image package 88
5.3 Perl module distributions packaged for Debian 96

A.1 Changes in human histones since *Marzluff et al.* (2002) 111
A.2 Human histone catalogue 112
A.3 Annotated histone variants 119
A.4 Histone codon substitution 120
A.5 Histone codon usage 121
A.6 Human histone variants catalogue 123
A.7 Anomalies in the annotation of human histones 125

B.1 Mouse catalogue build information 127

B.2 Count of histone mouse genes 128
B.3 Mouse histone catalogue 129
B.4 Canonical mouse H2A encoded protein isoforms 132
B.5 Canonical mouse H2B encoded protein isoforms 133
B.6 Canonical mouse H3 encoded protein isoforms 134
B.7 Canonical mouse H4 encoded protein isoforms 135
B.8 Catalogue of variant core histone genes and products 136
B.9 Anomalies in the annotation of mouse histones 137
B.10 Histone codon substitution 144
B.11 Histone codon usage 145
B.12 Changes in mouse histones since *Marzluff et al.* (2002) 148

E.1 Localisation of all human canonical and variant H2A genes and proteins 168
E.2 Reported Post-Translational Modifications (PTMs) for H2AX . 176

Acknowledgements

A recurrent theme on this thesis is how science builds on top of previous scientific discoveries. But science is built by people, and people build themselves on top of each other. Thank you to everyone who gave me the support and enabled me to build this high.

Thank you to my supervisor Andrew Flaus for all the guidance, concern, and support during this whole venture. How you care for your students and the science is an inspiration. I have found myself repeating your words to others that I'm now in charge of.

Thank you to everyone in the Flaus lab, especially to Martin, Jonathan, Indu, and Holger who have also been good friends and made it all light work.

I wish also to thank my co-supervisor Kevin Sullivan for all the discussions, ideas, and sharing of knowledge.

The Centre for Chromosome Biology was a very pleasant place to work, a truly open space for sharing of knowledge over a cup of coffee or tea. Thanks to everyone that made it so. Special thanks to Chelly, Emma, Lisa, Louise, Ronan, and Tiago.

I'm also grateful to all the random people on the internet that made for an odd extended collaboration. From IRC channels, internet forums, and mailing list, it's amazing the number and the size of communities involved in Free Software, not only working in the open between themselves, but also open to newcomers. Special thanks to Jordi and Juan from the Octave community.

Thank you for the encouragement to all my friends in Galway, Porto, and now in Oxford, that waited and prodded me to go on. I'm sorry for what was lost as well.

Thank you to Rei who carried me the most in the last few years and probably lost more sleep worrying about this than anybody else.

Finally, thank you to my family. My mother, father, and brothers for their constant support and friendship all this years.

Summary

Chromatin is a dynamic complex that controls access to genetic information by undergoing structural reconfigurations. Understanding this dynamic can provide insights into the biological implications of chromatin organisation.

We have undertaken a detailed catalogue of the human core histone genes and contributed to their annotations. Based on the reproducible research concept, we produced this catalogue with a system that is capable to generate new up to date manuscripts as a model for similar projects which can be continually improved along with genome annotations. As proof of concept, we used the same project to produce a catalogue of the current mouse histone genes.

Quantitative fluorescence microscopy has been used extensively to obtain insights into the dynamics of multiple proteins in live cells. Despite many advances in model design, fluorophores, and imaging capabilities, limitations are still encountered that can lead to misinterpretation of data. By using histone proteins with extremely slow exchange rates we have tested the limitations of Fluorescence Recovery After Photobleaching (FRAP) and developed approaches to overcome some of them. Importantly, we show that movement of chromatin precludes measurements of histone dynamics on the FRAP time scale.

To achieve these results we made contributions to multiple free software projects including Octave, BioPerl, and Debian. This included implementing new algorithms, refactoring code for efficiency and consistency, creating maintenance support tools, and packaging software for ease of installation by users. A core theme of this work was to create build systems capable of processing primary data from public databases or microscopy in a completely transparent way to generate complete manuscripts as implementations of the reproducible research concept. This thesis itself is an example of the approach.

In these studies we have tested the limits and developed new approaches to existing methods of chromatin analysis by designing novel reagents and software for the field of chromatin dynamics.

CHAPTER 1

Introduction

1.1 Chromatin

The human genome has a length of some 3.2×10^9 base pairs (*Lander et al.*, 2001) meaning that a diploid human cell will typically have 6.4 Gbp of genomic information inside its nucleus. All of this is organised in a large DNA–protein complex known as chromatin.

Access to the genomic information is not uniform in chromatin. It is organised in a hierarchical structure with multiple levels of compaction (Figure 1.1) where the level of compaction increases as access to the underlying genomic information decreases. This is often simplistically distinguished as euchromatin and heterochromatin for low and high compaction respectively, although more detailed categories of chromatin spatial organisation have been proposed (*Brehm et al.*, 2004; *Dixon et al.*, 2016). The binary distinction is in turn associated with active and inactive genes since a high level of compaction prevents DNA transcription by blocking the access of the required machinery (*Ball*, 2003).

Chromatin function is not limited by the compaction of DNA to fit into the cell nucleus. Instead, chromatin controls access to DNA by mediating transcription, replication, and recombination, as well as protecting DNA from damage. Hence, controlled and efficient access to genomic information is not simply a barrier problem for the sake of compaction but involves chromatin structure as an active participant in genomic mechanisms (*Felsenfeld and Groudine*, 2003).

The nucleosome is the basic repeating unit in the hierarchical structure of chromatin. An array of nucleosomes is then folded into a chromatin fibre by short-range interactions between neighbouring nucleo-

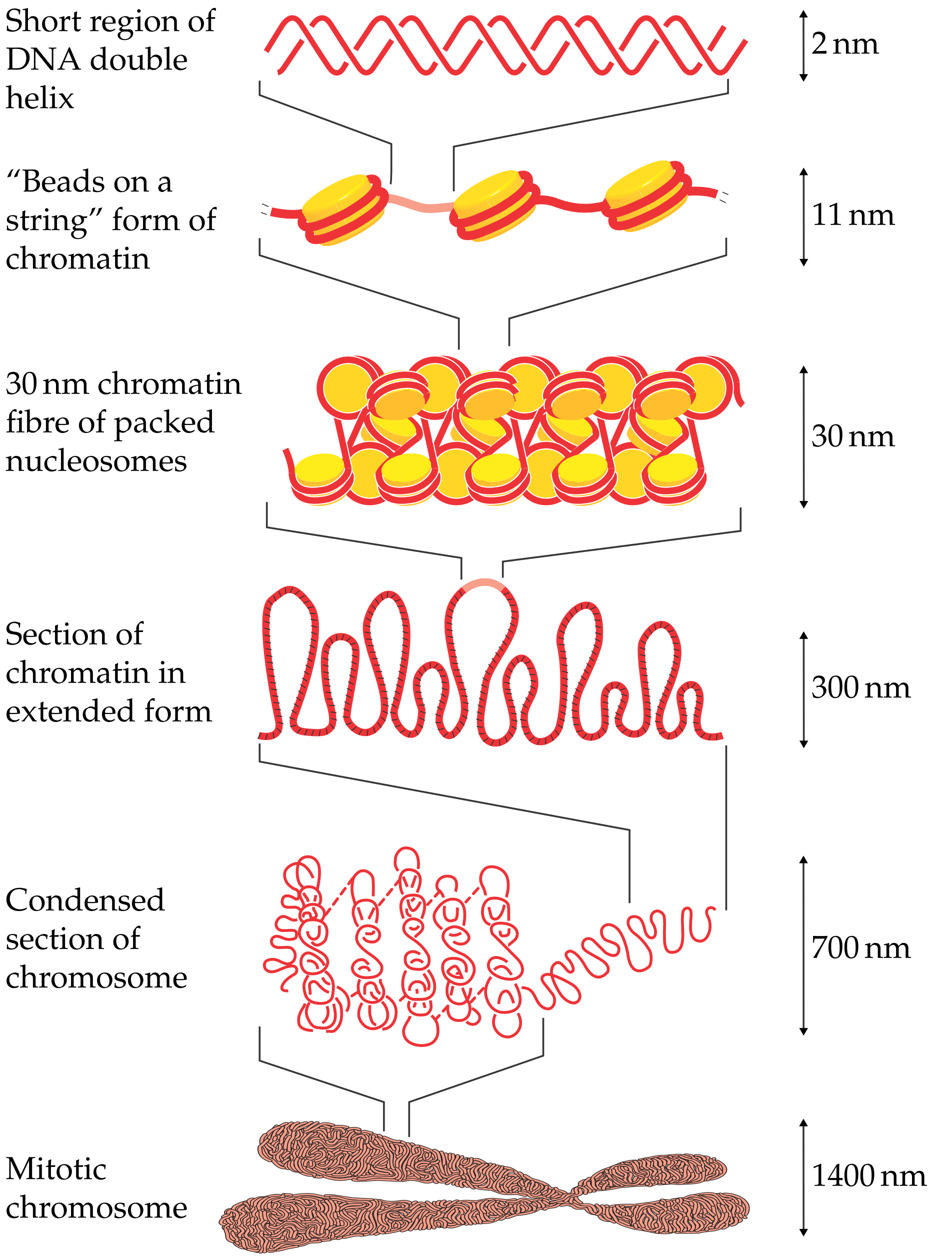

Short region of DNA double helix
2 nm
"Beads on a string" form of chromatin
11 nm
30 nm chromatin fibre of packed nucleosomes
30 nm
Section of chromatin in extended form
300 nm
Condensed section of chromosome
700 nm
Mitotic chromosome
1400 nm

Figure 1.1: **Overview of the chromatin hierarchical structure.** At the lowest level are the nucleosomes around which DNA is wrapped. This basic unit is repeated multiple times to form the classical view of "beads on a string" (Figure 1.2), which is folded on itself to form the 30 nm fibre, which is also folded on itself to form higher compaction fibres until it finally forms the mitotic chromosome. Figure adapted from *Alberts et al.* (2010) and *Lodish et al.* (2003).

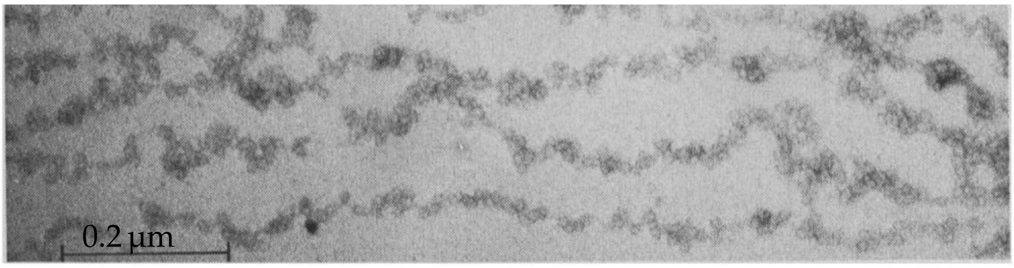


Figure 1.2: **Classic "beads on a string" view of chromatin.** Chicken erythrocyte chromatin negatively stained. Chromatin fibers are seen spilling out of ruptured nuclei depicting the classic "beads on a string". First published report on visualisation of nucleosomes, originally named ν bodies (*Olins and Olins*, 1974).

somes and the subsequent interactions between chromatin fibres shape the condensed chromosome structure (Figure 1.1). While the nucleosome structure is well established (*Luger et al.*, 1997), details of the intermediate assemblies are still controversial and some recent models eschew them altogether (*Fussner et al.*, 2011; *Luger et al.*, 2012).

The nucleosome itself is composed of a core particle consisting of 147 bp of DNA and a histone protein octamer, with a variable length linker DNA that connects flanking core particles (Figure 1.3(a)). Altogether the repeating nucleosome unit includes 165 bp to 245 bp of DNA depending on species and cell type (*Widom*, 1992).

Nucleosomes act as the substrate for almost all nuclear processes that require access to genomic DNA (*Felsenfeld and Groudine*, 2003). Reconfigurations of local chromatin necessary for these processes can be accomplished by changing nucleosome structure, or altering nucleosome composition by incorporation of variant histones or post-translational modifications. Understanding chromatin structure and the nature of changes which can occur during genomic activities requires understanding the biochemical properties of nucleosomes.

Histones

Histones are a family of proteins that are the principal protein components of chromatin. There are five histone types: H1, H2A, H2B, H3, and H4. Histone H1 binds to the linker DNA and is not part of the nucleosome core particle. The other four histones form the core nucleosome and are therefore referred to as core histones. The DNA polymer is negatively charged and the highly basic histone proteins bind to it, neutralising the negative charge and facilitating the required compaction of DNA. The histones interact with each other in a defined way to form a compact histone octamer structure that directs the DNA compaction.

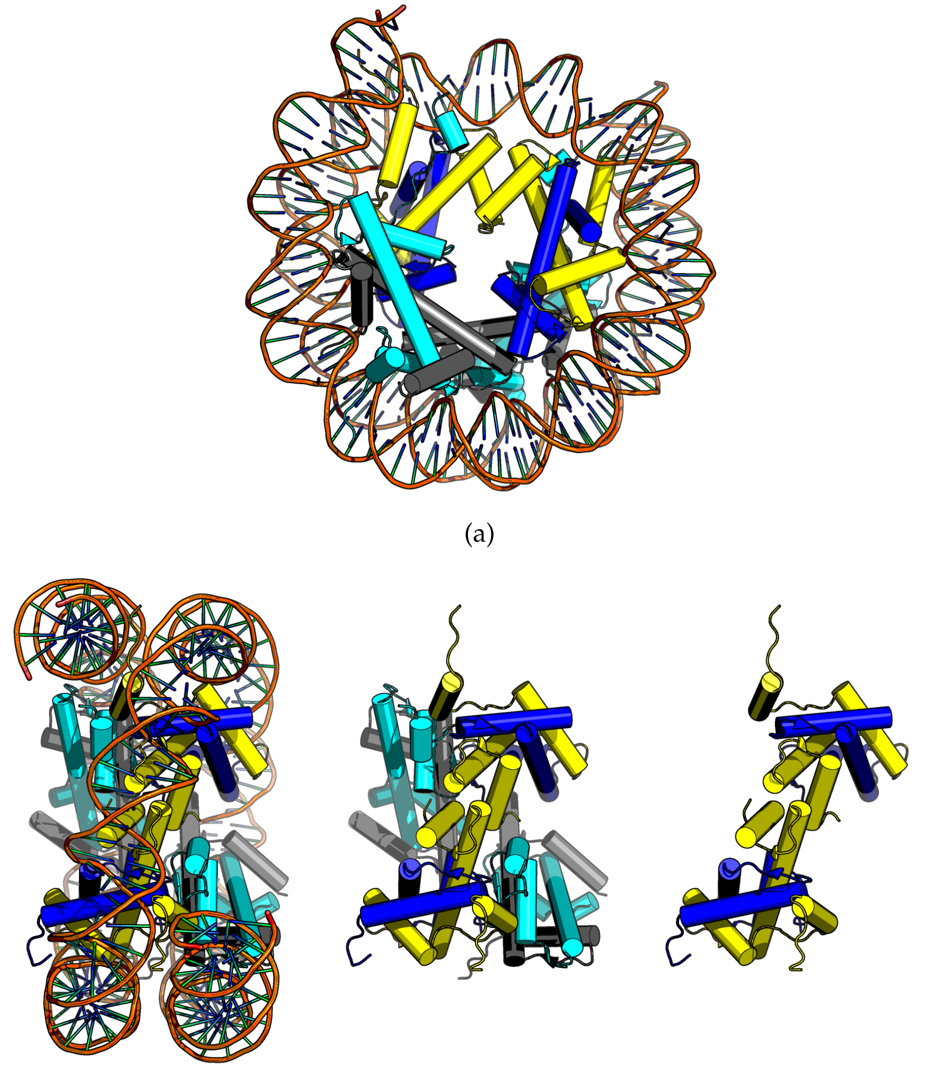
The histone octamer is composed of two of each of the core histones arranged in four dimers — two H2A–H2B and two H3–H4. The two H3–H4 dimers form a disk-shaped (H3–H4)$_2$ tetramer at the centre (Figure 1.3(d)) with one H2A–H2B dimer on each face of the disk (Figure 1.3(c)) and the 147 bp of DNA wrapped around the octamer in 1.65 turns to form the core particle (Figure 1.3(a) and 1.3(b)).

Histone and nucleosome structure

The linear sequence of each core histone protein can be divided into three parts: The histone fold domain is the central part and is common to all four core histones. It is involved in the formation of histone dimers and DNA binding. This is supplemented by histone fold extensions such as H3 αN helix and H2B αC helix, and the histone tails which are regions of polypeptide that extend beyond the nucleosome to interact with DNA and other nucleosomes amongst other functions.

Core histone proteins all share the same common histone fold domain, which is comprised of a long central α-helix flanked on both sides by loops and shorter α-helices (*Arents et al.*, 1991; *Arents and Moudrianakis*, 1995). This domain provides the characteristic "handshake" motif of histone fold dimer assembly whereby the central α-helix of each histone crosses its dimeric partner and fits between the shorter α-helices of the other (Figure 1.4(d) and 1.4(g)). This gives histones an extensive molecular contact interface within the dimer.

Each of the four core histones has unstructured N and/or C terminal tails that extend beyond the DNA in the nucleosome core, making inter-nucleosome interactions and therefore playing an important role in chromatin higher order structures. In addition, the tails are subject to a wide range of post-translational modifications including acetylation, methylation, ubiquitination, phosphorylation, and ADP–ribosylation (*Bannister and Kouzarides*, 2011). The vast number of these post-translational modifications has led to the histone code hypothesis that these modifications



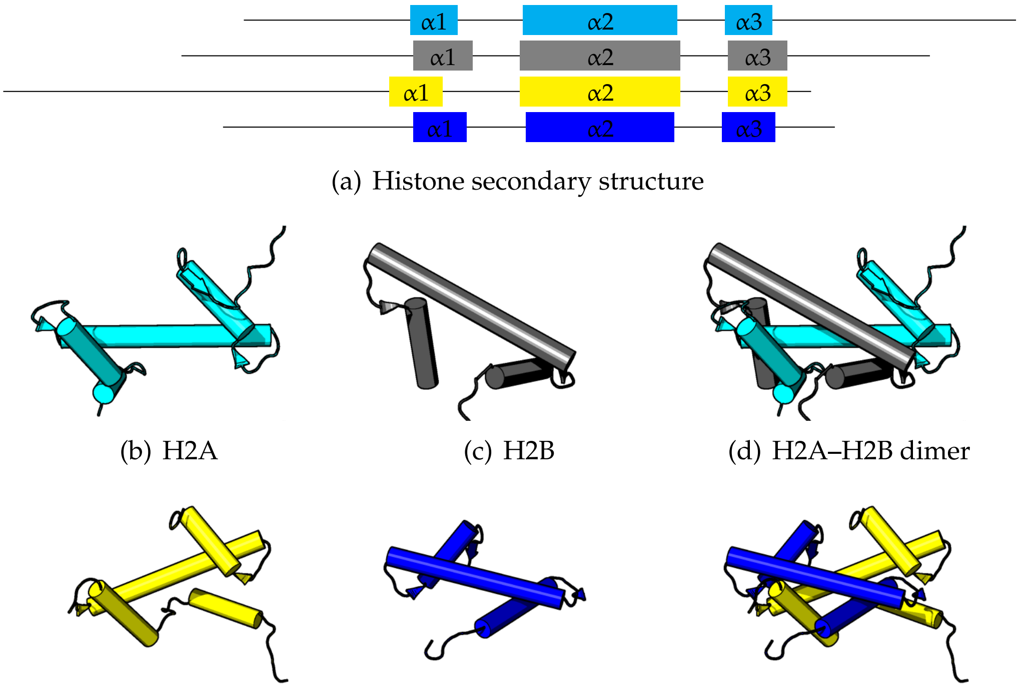
(a)

(b) Nucleosome core particle

(c) Histone octamer

(d) $(H3–H4)_2$ tetramer

Figure 1.3: **Structure of the nucleosome core particle.** Axial view of the nucleosome core particle facing towards the DNA entry/exit point (b) showing 1.65 turns of DNA wrapped around the histone octamer; (c) DNA hidden showing the histone octamer; (d) the two H2A–H2B dimers hidden to show the disk shaped $(H3–H4)_2$ tetramer centre. Figure generated from PDB structure 2CV5 (*Tsunaka et al.*, 2005). Histones H2A, H2B, H3, and H4 are coloured cyan, grey, yellow, and blue respectively.



(a) Histone secondary structure

(b) H2A

(c) H2B

(d) H2A–H2B dimer

(e) H3

(f) H4

(g) H3-H4 dimer

Figure 1.4: **Histone fold domain organisation.** (a) Secondary structure of the core histones with the α-helices that form the histone fold domain highlighted. (b)–(g) Tertiary structure of the individual core histones showing their histone fold domain, and quaternary structure of the H2A-H2B and H3-H4 dimers showing the histone fold domain in its characteristic "handshake" motif. Figures generated from PDB structure 2CV5 (*Tsunaka et al.*, 2005). Histones H2A, H2B, H3, and H4 are coloured cyan, grey, yellow, and blue respectively.

are the basis for regulation of chromatin-templated processes (*Jenuwein and Allis*, 2001).

Histone post-translational modifications have indeed been shown to regulate chromatin structure by recruiting different protein complexes. This influences DNA processes such as DNA replication, repair, recombination, and transcription. The most studied modifications have been related with transcriptional gene activation and gene silencing as outlined in the original proposal for the histone code (*Jenuwein and Allis*, 2001) that simplified chromatin structure effectively to either "on" or "off" states. For example, H3 K4 di-methylation and H3 K27 tri-methylation are well known markers for gene activation and silencing respectively. Similarly, phosphorylation of serine 139 in histone variant H2AX is involved in the recruitment of DNA repair-related proteins (*Pinto and Flaus*, 2010).

Histone variants

Each of the histone protein types is encoded by a family of genes in most genomes, resulting in several slightly different proteins. These are grouped into two classes, canonical isoforms and variant histones, which are based on their gene location, expression characteristics, and functional roles. Canonical isoforms contribute to the majority of histones in chromatin, have high sequence identify, and display largely equivalent functions as discussed in more detail on Chapter 3. Variant histones are present in smaller number than canonical histones, have more distinct sequences with unique structural features, and perform a variety of specialised functions. For example, the H2A variant H2AX has a longer C-terminal tail with unique locations for post-translational modifications involved in the DNA damage response (*Pinto and Flaus*, 2010). CENP-A is a H3 variant that forms a nucleosome with altered biophysical properties and has a unique N-terminal tail that is involved in centromere identity (*Black and Cleveland*, 2011).

Histone H1

"Histon" was the name given to the protein fraction of the original crude chromatin extracts (*Kossel*, 1884) and the name remains, including histone H1, even though histone H1 was the first species to be separated from the others (*Stedman and Stedman*, 1951).

Histone H1 is not part of the nucleosome core particle and is not required for the beads on a string structure. Instead, it binds to the linker DNA to stabilise the higher order chromatin structures and modulate the accessibility of regulatory proteins, chromatin remodelling factors, and histone modification enzymes to their target sites. Details of the contribution of H1 to chromatin structure and function remains unclear but the typically accepted view is that it binds at the nucleosome DNA entry/exit points and that it shortens the length of the linker DNA bringing nucleosomes close together to form the 30 nm fibres (*Harshman et al.*, 2013).

Similar to the canonical core histones, expression of histone H1 is replication dependent with variant H1 isoforms having specialised roles and tissue-specific expression. All canonical H1 genes are located in the main histone gene cluster on human chromosome 6, and the H1 protein is also the target of multiple post-translational modifications. Structurally, histone H1 isoforms consist of a short unstructured N terminal, a globular winged helix domain, and a variable length unstructured C terminal with a high lysine content, which is why these histones are also known as lysine-rich histones.

There are fundamental differences in the functional roles of the core histones and histone H1. There is no sequence similarity between H1 and the core histones, histone H1 does not have a histone fold, histone H1 is not evolutionary conserved like core histones, and histone H1 is not required for nucleosome formation.

Historical perspective on chromatin structure

The first chromatin preparation (*Kossel*, 1884) was named Nucleïn, and was separated in two components: one basic and protein-like which became known as histones, and one acid and unlike any cellular substance yet observed which became known as nucleic acid. This discovery was the precursor to the discovery of DNA as the hereditary molecule and genetic material, the discovery of DNA itself, the discovery of the nucleus as the recipient for genetic information, and even the discovery of proteins as polypeptides. The series of advances that led to the nucleosome structure form a timeline extending for over a century (Table 1.1, *van Holde* (1988)).

Chromatin Dynamics

Chromatin within the nucleus appears highly organised with each individual chromosome occupying a distinct territory (*Cremer et al.*, 2006). Early studies of bulk chromatin dynamics, including experiments using Fluorescence Recovery After Photobleaching (FRAP), indicated that bulk chromatin in interphase was essentially immobile (*Abney et al.*, 1997). Observation of labelled chromosome territories in live HeLa cells revealed no major mobility on the scale of individual chromosomes, with only small Brownian diffusion being recognised (*Edelmann et al.*, 2001).

Table 1.1: **Historical perspective of nucleosome structure.**

***Miescher* (1871)** First chromatin preparation, named Nucleïn.

***Miescher* (1874)** First DNA purification, also named Nucleïn, and identification of protamines from a sperm chromatin preparation. Demonstration that DNA exhibits acidic properties.

***Flemming* (1880)** Chromatin named for the first time as the readily stained material in cell nuclei. Only later would it be identified as the same as the Nucleïn in Miescher preparations.

***Kossel* (1884)** Histones named for the first time. They were the peptone-like substances with basic properties obtained by acid extraction from a chromatin preparation.

***Huiskamp* (1903)** After a series of studies on chromatin chemical properties, including the demonstration that Nucleïn is negatively charged and histones positively charged, it was proposed that histone and DNA binding is by electrostatic interactions.

***Stedman and Stedman* (1951)** Demonstration that histones are not homogeneous and can be separated into main and subsidiary histones, which are now known as core histones and histone H1. Further demonstration that these two fractions are not homogeneous themselves and that their distribution is tissue specific. Proposal that histones play a role in gene regulation based on the tissue specificity of certain histones.

***Allfrey et al.* (1964)** Proposal of a new model of transcription regulation by histone acetylation and methylation that modifies the interactions of histones with DNA in a reversible and dynamic manner.

***Phillips and Johns* (1965)** Culmination of a long series of papers about fractionation and identification of histones leading to the identification of the five histones types known today. This was made possible by the development of chromatographic fractionation methods and gel electrophoresis techniques.

***Hewish and Burgoyne* (1973)** Digestion of chromatin by nucleases yields 200 bp fragments. This led to the proposal that chromatin has a repeating substructure with a repetitive spacing.

***Olins and Olins* (1974)** First view of the "beads on a string" structure by electron microscopy (Figure 1.2). These results were initially regarded as artefacts (*Pardon and Wilkins*, 1972).

***D'Anna Jr and Isenberg* (1974)** Demonstration of strong association between H2A–H2B.

***Kornberg and Thomas* (1974)** Demonstration that histones form a (H3–H4)$_2$ tetramer and a H2A–H2B dimer.

***Kornberg* (1974)** Regular structure of the nucleosome is proposed as a repeating unit of chromatin with eight histones, two of each, and about 200 bp of DNA.

***Bradbury* (1975)** The current histone nomenclature is presented and submitted to IUPAC for official standardisation.

***Richmond et al.* (1984)** Crystal structure of the nucleosome solved at intermediate resolution of 7 Å. This reveals a disk shaped (H3–H4)$_2$ symmetric tetramer at the centre of the nucleosome, H2A–H2B dimers on each face of the disk, and the DNA sequence around it.

***Luger et al.* (1997)** Crystal structure of the nucleosome solved at atomic resolution of 2.8 Å showing detail of interactions between the histones and their internal structure.

***Davey et al.* (2002)** Crystal structure of the nucleosome solved at 1.9 Å resolution, the highest to date, showing details on the interactions between the DNA double helix sequence and histone octamer.

Some authors described curvilinear movements of chromatin but they were attributed to nuclear rotation. This apparent immobility was explained as chromatin attachment to subnuclear structures such as nucleolus and nuclear envelope, structural RNAs, and the nuclear matrix (*De Boni and Mintz*, 1986; *Parvinen and Söderström*, 1976).

However, increasing evidence indicates that chromatin motility is very complex and encompasses several levels of dynamic processes operating on different spacial scales and time frames (*Hübner and Spector*, 2010).

The yeast *Saccharomyces cerevisiae* centromeric chromatin movement is mainly caused by Brownian motion and is restricted to a small nuclear volume, apparently restrained by crowding, entanglement, and the presence of immobile structures such as microtubules (*Marshall et al.*, 1997). In contrast to centromeric and telomeric chromatin, internal chromosome mobility is dependent on the cell cycle. Noncentromeric chromatin is highly mobile including not only small-scale displacements (<0.2 µm) but also occasional large (>0.5 µm) displacements occurring in a time frame of seconds. The large movements are restricted in G_1 phase possibly due to the crowding effects of a small nucleus. The mobility is also reduced in S-phase and is speculated to be caused by chromatin attachment to nuclear structures such as the nuclear envelope or internal nuclear structures (*Heun et al.*, 2001).

Furthermore, transcriptional regulation has been shown to affect chromatin mobility. Inactive loci generally found in the nuclear periphery are relocated to the interior of the nucleus upon activation. For example, a locus that was transcriptionally active was shown to move towards the interior of the nucleus at 0.1 $\mu m\,min^{-1}$ to 0.9 $\mu m\,min^{-1}$ over 1 µm to 5 µm (*Chuang et al.*, 2006). Furthermore, inter and intra-chromosomal interactions in transcription complexes that co-regulate distal genes were shown to require locus movement towards transcription factories (*Osborne et al.*, 2004). Another example of chromatin motion is the contraction and looping of the 655 kbp T cell receptor (TCR) beta loci to form close interactions with the TCR alpha–delta loci during TCR recombination (*Skok et al.*, 2007).

A more recent study developed the Displacement Correlation Spectroscopy (DCS) technique for bulk chromatin dynamics and showed that chromatin movement is organised in large regions in the scale of 4 µm to 5 µm but that these are not chromosomal territories (*Zidovska et al.*, 2013).

All these recent studies provide a more dynamic picture of the interphase chromatin than was appreciated even a decade ago.

1.2 Fluorescence microscopy

Fluorophores are molecules that absorb light at one range of wavelengths, the excitation wavelengths, and emit light at a range of longer wavelengths, the emission wavelengths. Illumination of a sample with the excitation wavelength, and filtering out all but the emission wavelength, allows the localisation of fluorophores in a dark background. If specific molecules and cells can be associated with a fluorophore, their localisation can be inferred, and by choosing fluorophores with different emission wavelengths multiple molecules can be simultaneously identified.

In a typical widefield microscope configured for such experiments, light in the excitation wavelengths is selected by an optical filter, the excitation filter, from a wide-spectrum and high-intensity light source. The excitation light is then reflected by a dichroic mirror, and directed through an objective onto the sample, some of which is absorbed by the fluorophores causing some of them to be excited and emit light at a higher wavelength. Some of this emitted light is directed back through the same objective and passes through the dichroic mirror to produce the visualised image. The dichroic mirror is the central element in the configuration since it reflects the excitation light but allows passage of the emitted light. This is known as epifluorescence microscopy.

Fluorescent proteins

Countless fluorescent labels have been developed with different properties suited to specific microscopy techniques (*Rizzo et al.*, 2009; *Olenych et al.*, 2007). These are often split into two major classes: Synthetic fluorescent molecules such as DAPI or fluorescein that, either bind to the molecule of interest or can be chemically linked to it. In contrast, Fluorescent Proteins (FPs) are genetically encodable and can be expressed by the cells themselves. FPs can be fused to a protein of interest inside live cells, enabling visualisation of dynamics by real-time live-cell imaging (*Rizzo et al.*, 2009).

The *Aequorea victoria* jellyfish Green Fluorescent Protein (GFP) was the first FP to be expressed recombinantly. Although originally discovered in 1962 (*Shimomura et al.*, 1962), it was only in 1994 that it was used as a gene expression marker under the control of *T7* and *mec-7* promoters, in *Escherichia coli* and *Caenorhabditis elegans* respectively (*Chalfie et al.*, 1994). Since then many GFP variants have been engineered by mutating the original nucleotide sequence. These not only increase fluorescence quantum yield, photostability, and improved folding, but also change the excitation and emission spectrum, providing a wide range of GFP derivatives with different colours. FPs in other organisms were also discovered and further engineered to provide improved variants and wider range of fluorescent wavelengths (*Olenych et al.*, 2007).

A GFP fused H2B was the first histone to be fused with a fluorescent protein. Despite the large size of GFP when compared to H2B, a 27 kDa tag in a 14 kDa protein, it was shown that H2B–GFP is efficiently incorporated into nucleosomes and its distribution is comparable to the endogenous H2B without affecting cell cycle (*Kanda et al.*, 1998). Fusion of GFP with an histone H2A and its variants followed (*Perche et al.*, 2000), as well as histones H3 and H4 (*Kimura and Cook*, 2001). However, while the original tagging of H2B made use of a linker sequence of six amino acids, tagging of H3 and H4 required a longer linker of 23 amino acids, possibly due to their location at the centre of the nucleosome core particle. In addition, while the new biochemical assays further validated that GFP tagging did not affect the stability of nucleosomes and its incorporation in chromatin, microscopy experiments showed that the distribution of tagged H3 and H4 did not mimic the endogenous protein. The different distribution was explained as an effect of the unregulated expression of the tagged histones, which coupled with the slow exchange of H3 and H4 in chromatin, favoured their incorporation in the early phases of S phase (*Kimura and Cook*, 2001).

An important class of FPs that have been developed are the photocontrollable FPs (*Shcherbakova et al.*, 2014). These have fluorescent properties that can be modulated by excitation with specific wavelengths allowing for individual cells and proteins to be optically labelled. This strategy is especially useful for studies of cell lineage and protein movement. The photocontrollable FPs can be divided in three categories: Photo-activatable, photo-switchable, and reversible photo-switchable.

Photo-activatable FPs undergo irreversible dark to bright state conversion. PA–GFP (Photo Activatable GFP), the first PA–FP to be reported, was developed by mutating the original GFP (*Patterson and*

Lippincott-Schwartz, 2002) and is still the only green PA–FP. It allows simple marking and tracking of subsets of molecules within cells.

Photo-switchable FPs undergo irreversible conversion from one bright state to another bright state with a different emission wavelength. Like PA–FPs, they allow the tracking of molecules with the added advantages that the whole set is visible before the switch, and that the non-switched molecules continue to be visible. The development of proteins of this category started with Kaede FP which is an obligate homotetrameric complex (*Ando et al.*, 2002), thus making it unsuitable for use as a genetically encoded fusion tag. Monomeric Kaede-like FPs have since been developed including mEos2 (*McKinney et al.*, 2009).

Reversible photo-switchable FPs undergo interchangeable conversion between dark and bright states. These FPs, such as Dronpa, have mostly been used in single molecule localization microscopy, a type of super-resolution microscopy, due to the low energy required for the transition between states.

FRAP

Fluorescence Recovery After Photobleaching (FRAP) is an optical microscopy technique to measure the dynamics of fluorescently tagged molecules. Originally developed in the 1970s for quantitative dynamics of lipids in cell membrane under the name Fluorescence Photobleaching Recovery (FPR) (*Axelrod et al.*, 1976), it has been extensively used to obtain qualitative and quantitative insight into the kinetic properties of proteins since the development of GFP tagging.

FRAP is a widespread technique in the field of cell biology, enabling a wide range of studies into dynamics including cell membrane protein diffusion, dynamics of protein complex assembly, and protein aggregate degradation, to flow of drugs in extracellular matrices for drug delivery, fat migration in chocolate manufacturing, and plasticizers into food from packaging (*Sprague and McNally*, 2005; *Mueller et al.*, 2010; *Lorén et al.*, 2015).

In the FRAP technique, fluorescently tagged molecules in a small region are photobleached and the fluorescent intensity in the bleached region is measured over time to obtain a recovery curve which provides an insight into the dynamics of the tagged molecule. The recovery is a function of many parameters, specific to each FRAP experiment. For example, freely diffusing fluorescent molecules in a cellular compartment can be bleached and recovery happens though transport of unbleached molecules from the cytoplasm providing an estimate to transport rates. Alternatively, fluorescent molecules are bound to an immobile complex and the recovery is due to unbleached molecules from outside the bleached region moving into the photobleached region and exchanging into the complexes there. This provides estimates for association and dissociation rates.

Two factors contribute to the widespread use of FRAP. Firstly, it does not require specialised instruments beyond a standard confocal or widefield microscope with a laser module. Secondly, recovery curves are intuitive to understand. This is not the case with alternative techniques such as Fluorescence Correlation Microscopy (FCS) or single molecule tracking.

In the simplest case, a FRAP recovery curve can be analysed qualitatively. Using the plotted data of the qualitative recovery rate, the presence of an immobile fraction and binding interactions, and how this changes between different cases, can be interpreted (Figure 1.5).

In quantitative FRAP, data analysis involves fitting the recovery data to an idealized mathematical model. Many different models exist with different assumptions and dependencies on parameters (*Mueller et al.*, 2010). Kinetic parameters are typically k_{on} and k_{off}, for binding and unbinding rate constants, and D_f for the diffusion constant although a variety of additional parameters can be included in the design of FRAP models (*Mueller et al.*, 2010).

The compartment geometry must also be considered. This includes deciding whether there are boundaries to the compartment or whether it is an infinite space. A simplification of the compartment from a three dimensional volume to a two dimensional area can be applied.

The bleaching event is a source of considerable variability. The time interval between the bleaching and the acquisition of the first image is a window where recovery of a mobile fraction might happen. Likewise, the bleaching event is not instantaneous and will occur while molecules are exchanging. The shape of the bleached area is a parameter and the profile of the laser beam will have an effect that can be modelled as either a gaussian or a flat-top distribution.

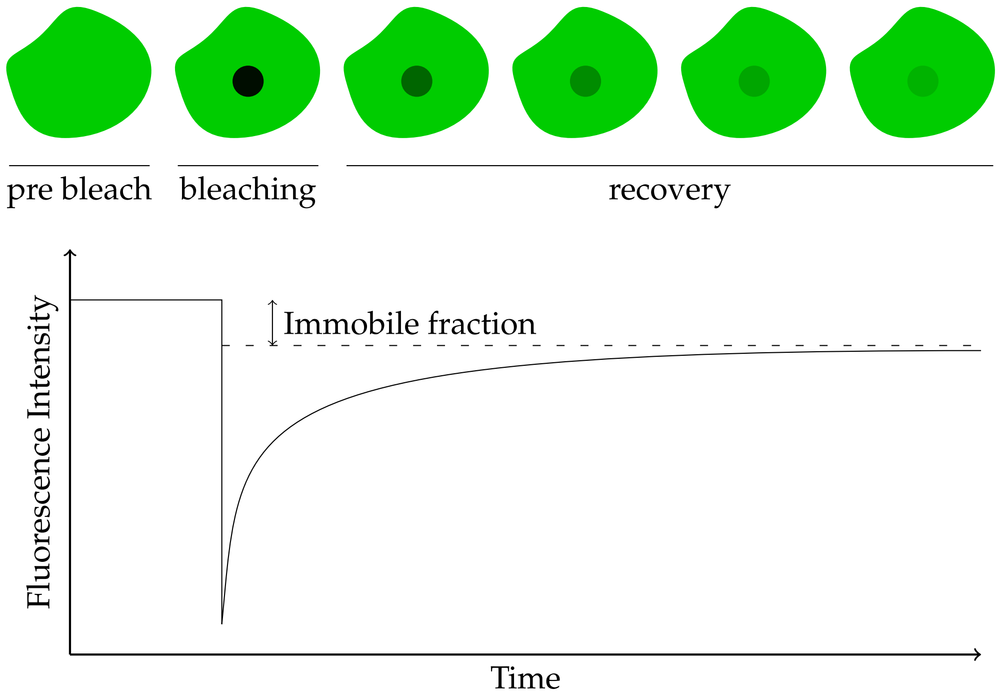


Figure 1.5: **Concept of a FRAP experiment.** The fluorescent proteins of a small region are permanently bleached by a laser event. The intensity in the region is measured so that the plot shows the recovery of fluorescence in the bleached area. The recovery curve reflects the diffusion and binding dynamics of the proteins tagged with the fluorescent proteins. The fraction of fluorescence that does not recover is the fraction of immobile proteins.

Diffusion rates may be fixed, fitted, or assumed to be instantaneous. A single diffusion rate may be assumed for the whole compartment, or non-homogeneous regions in the compartment can have different diffusion rates.

Finally, if binding to a complex is considered, a plethora of additional potential parameters are introduced. Multiple binding states may be present when binding reactions occur in a series of steps involving multiple proteins, and the distribution of the binding sites may be not equal throughout the compartment.

While all these many features may appear as parameters in the mathematical model for fitting to a FRAP curve, they also affect how the experiment is performed.

FRAP has a number of implicit assumptions which are applicable for most biological situations. Firstly, the biological system is assumed to have reached equilibrium and the equilibrium is maintained throughout the experiment. This means the kinetic parameters must remain constant throughout the experiment. Secondly, the distribution of the tagged protein mimics the endogenous protein. And finally, the binding sites should be part of a large, relatively immobile complex on the time and length scale of the recovery. These assumptions become difficult to maintain as the experiment times increases for slow moving molecules.

1.3 Bioinformatics

With the advent of genomic and transcriptomic sequencing, and with the development of super resolution microscopy, a biology laboratory can easily generate several gigabytes of data in a single set of experiments. Even laboratories that lack access to the required equipment can easily access these amounts of data in public repositories. The analysis of this data is becoming a bottleneck for research advances (*Marx*, 2013) so the development of tools to handle large datasets is a race against the technology that generates it (*DeLisi*, 1988). This has led to the rise of a new field of science, bioinformatics, for the research of software tools for the acquisition, storage, and analysis of biological data.

In this context, computers have become an essential tool in all areas of biological research (*Wren*, 2016). A researcher will use a computer on a daily basis as an enabling device for their research. DNA and protein sequences are computer files to be analysed by specialised software. Cloning strategies are planned with the help of software. Sequences are searched for in large databases and compared against other sequences for similarities. Likewise, the structure of biomolecules is predicted by software and visualised in 3D in a computer. Microscopes are controlled by computers, such that the images acquired by electronic cameras mean eyepieces are now optional. Even literature review is performed in a computer with new publications revealed by searching in online databases rather than by browsing issues in libraries. Software is thus essential for science and is a daily tool of the scientist.

Free Software and Scientific Advances

Scientific discoveries are almost always incremental on previous work. This is true not only for new theories but also for methods, since new techniques are usually generated by a constant iteration of improvements. Software used by scientists is also a tool and should allow for such a series of iterative improvements.

Outside of academia, lack of access to the source code of the software is also an issue. The problem has led to multiple organisations being formed to promote the ability of users to modify the software they use so it works for their own purposes.

Two of the main proponents of this philosophy are the Free Software Foundation and the Open Source Initiative. The Free Software Foundation defines free software as providing the four basic freedoms: Freedom to run a program for any purpose, freedom to study the source code of the program, freedom to distribute the program, and freedom to distribute modified versions of the program (*Free Software Foundation*, 2015). In fact, the free software movement began in academia itself, by Richard Stallman at the MIT Artifical Intelligence lab. This was triggered by a series of events caused by non-free software, when an AI lab spin-off made its software non-free, and the lab mainframe was replaced with another that ran on non-free software (*Stallman*, 2015).

Free Software Projects in Academia

This growing openness and sharing of code to enable incremental development of scientific software is becoming recognised as an important driver of progress in highly data driven research fields.

Despite the lack of computational details in the data analysis, free software is widely used in biological research, and multiple projects exist for enabling data analysis in molecular cell biology. This demonstrates that free software does enable research. For example, BioPerl is a large project for the handling of biological data in the Perl programming language and was heavily involved in facilitating the sequencing of the human genome (*Stajich et al.*, 2002). Following its success, similar projects have been created for the Python and Java languages, Biopython (*Cock et al.*, 2009) and BioJava (*Prlić et al.*, 2012) respectively. Many programming languages used by bioscience researchers are themselves free software, including R (*R Core Team*, 2014), GNU Octave (*Eaton et al.*, 2016), and Julia (*Bezanson et al.*, 2014). Similarly, interactive applications have also been created for a wide range of purposes such as ImageJ for image analysis (*Schneider et al.*, 2012), GBrowse for annotation and browsing genomic data (*Stein*, 2013), and OMERO for data storage (*Allan et al.*, 2012).

Meta-research

The lack of access to sufficient code in publications not only delays research by obliging investigators to redo the work of others to build on it, but also increases the difficulty of validating research by making published experiments unable to be reproduced. For example, for a survey of 18 published microarray gene-expression studies, it took a team of data analysts an average of half a week to replicate a single figure or table, with some figures taking more than one week. Only two of the 18 figures were "reproducible in principle", that is, with differences that were considered to be minor (*Ioannidis et al.*, 2009). Difficulty with reproducibility does not imply that results have been fabricated. The main problem identified was the lack of availability of data, and the second was the lack of detail on how to perform the analysis.

This reproducibility gap has become a growing issue in recent years and led to the creation of the new field of meta-research, for the investigation of the research practice (*Ioannidis et al.*, 2015; *Collins*, 2014).

With computers being central to analysis, one would expect that analysis could be readily shared in the form of source code. However, that is not the case. Publications favour a natural language description of the analysis which lacks the level of detail that would allow others to repeat or improve on work (*Ince et al.*, 2012). An indirect cause of this anomaly is that scientists share their work through print-centred articles and books that are far from ideal media for complicated source code or procedural instructions. There is a growing interest within the scientific community to improve the situation and journals have been changing their policies to promote sharing of source code (Nature Editorial, 2014; *GMD Executive Editors*, 2013; *Stodden et al.*, 2013).

1.4 Aims and Objectives

Histones organise the fundamental nucleosomal level of chromatin and although the static structure of nucleosomes is known, their dynamic properties are poorly understood. Current models of nucleosome dynamics are largely based on *in vitro* experiments and observations in unicellular yeast which are largely composed of euchromatin. We wished to apply FRAP to test models of nucleosome dynamics in human cells and their dependence on amino acid variations in core histones.

While canonical core histones are usually referred to as single proteins by their type H2A, H2B, H3, or H4 only, they actually comprise a family of isoforms whose sequence differences are rarely recognised. We required a catalogue of the human histones to understand this sequence variation as a basis for the FRAP investigation but the last version of such an inventory had been performed in 2002, before the finalisation of the human genome (*Marzluff et al.*, 2002). Therefore, we assembled an up to date catalogue of all human core histones presented in Chapter 3. Recognising the inevitable changes to such catalogues over time, we automated its creation so that a constantly up to date version could be generated with little effort.

As a first attempt to apply FRAP as a tool for measuring the effect of histone sequences on mobility in cell nuclei we chose to investigate the stability of core histone SIN (SWI/SNF INdependence) mutations, a set of mutations which has been show *in vitro* to have a large effect on nucleosome mobility. We faced complications with long time frame FRAP measurements. In Chapter 4 we detail our step-wise approach to solving these issues.

Through these studies, I developed many software components and strove to improve existing tools that would be available to all researchers. In Chapter 5 we describe the multiple tools created and improved, all of which have been released under a free software licence.

Inspired by the principles of reproducible research, all the source for this thesis is also available with a build system to automate its construction from the raw data. This thesis is itself part of our broader effort to investigate the application of software development tools for the maintenance of reproducible research.

CHAPTER **2**

Materials and Methods

2.1 Source of chemicals, enzymes, and solutions

All chemicals used were purchased from Sigma unless otherwise stated. All solutions were prepared according to Table 2.1 with Milli-Q purified water, and autoclaved prior to use when appropriate.

Restriction enzymes, T4 DNA ligase, T4 Polynucleotide Kinase (PNK), and DNA ladders were obtained from New England BioLabs. Protein ladders were obtained from Invitrogen.

For use in tissue culture, Fetal Calf Serum (FCS), and Dulbecco's Modified Eagle's Medium (DMEM) without phenol red and L-glutamine, were obtained from Lonza. Non-Essential Amino Acid (NEAA) solution, Dulbecco's Phosphate-Buffered Saline (DPBS) without Ca^{2+} and Mg^{2+}, and DMEM supplemented with glucose, sodium pyruvate, L-glutamine, and phenol red, were obtained from Sigma. Trypsin–EDTA and Penicillin–Streptomycin solution were obtained from Gibco.

2.2 DNA methods

Bacterial cultures

E. coli cultures were prepared with either LB broth or agar at 37 °C. For antibiotic selection, ampicillin, kanamycin, and chloramphenicol, were used at concentrations of 100, 30, and 34 $\text{mg}\,\text{l}^{-1}$ respectively.

Table 2.1: **Commonly used buffers and media.**

DMEM
4.5 g l^{-1} glucose; 110 mg l^{-1} L-glutamine; 584 µg l^{-1} sodium pyruvate; 15.9 mg l^{-1} phenol red.

DNA loading buffer (10 ×**)**
25 % Ficoll (w/v); 100 mM Tris–HCl pH=7.4; 100 mM EDTA.

FACS buffer
96 % DPBS (v/v); 2 mM EDTA; 25 mM HEPES buffer pH=7.0; 1 % Fetal Calf Serum (FCS) (v/v).

Freezing media
90 % FCS (v/v); 10 % DMSO (v/v).

Growth medium (HeLa, HEp2)
89 % DMEM (v/v); 9 % FCS (v/v); 1 × NEAA solution; 50 units/ml penicillin; 50 µg ml^{-1} streptomycin.

Growth medium (horse)
81 % DMEM (v/v); 16 % FCS (v/v); 2 × NEAA solution; 50 units/ml penicillin; 50 µg ml^{-1} streptomycin.

LB agar
20 g l^{-1} LB broth powder; 7.5 g l^{-1} agar.

LB broth
20 g l^{-1} LB broth powder.

Running buffer
1 × TG; 0.1 % SDS (w/v).

TAE (Tris Acetate EDTA)
40 mM Tris; 20 mM acetic acid; 1 mM EDTA.

TBS (Tris Buffered Saline)
50 mM Tris–HCl pH=7.5; 100 mM NaCl.

TBS-T (TBS-Tween)
99.95 % TBS (v/v); 0.05 % Tween 20 (v/v).

TG (Tris Glycine)
25 mM Tris; 192 mM glycine.

Transfer buffer
1 × TG; 15 % methanol (v/v).

Preparation of competent bacteria

Competent *E. coli* cells were prepared from a culture of Invitrogen's One Shot TOP10 Chemically Competent *E. coli*. LB cultures of 1 l were grown at 37 °C until an $\mathrm{OD}_{600\,\mathrm{nm}}$ of 0.4 to 0.5. Subsequent steps were carried at 4 °C with previously chilled equipment and solutions.

Cultures were centrifuged at 6000 g_n for 10 minutes, the pellet was resuspended in 500 ml of 0.1 mM $CaCl_2$, and incubated on ice for 30 minutes. The suspensions were centrifuged again at 6000 g_n for 10 minutes, and the resulting pellet resuspended in 100 ml of $CaCl_2$ with 15 % glycerol. Aliquots of competent cells were prepared and stored at −80 °C.

Transformation efficiencies were measured after preparation of each batch and discarded if less than 1×10^6 cfu mg^{-1} of plasmid was obtained. Absence of antibiotic-resistant contaminations was assessed by streaking the cells on selective plates.

Transformation of competent cells

Competent cells were thawed on ice and split into aliquots of 50 µl to prechilled 2 ml tubes where 1 µl of ≈100 ng $\mu\mathrm{l}^{-1}$ DNA was added. Cells were incubated on ice for 30 minutes, followed by a 60 seconds heat-shock at 42 °C, and 5 more minutes on ice. 300 µl of non-selective LB was added to each tube and the cultures incubated at 37 °C with vigorous shaking for 45 minutes. Samples from the cultures were then plated onto the appropriate antibiotic containing agar plates, and incubated overnight at 37 °C.

For concentrations of plasmid DNA higher than 500 ng $\mu\mathrm{l}^{-1}$, only 0.3 µl of DNA was used, and both the initial and final incubation steps were shortened to 10 minutes.

Plasmid DNA preparation

Plasmid DNA was prepared with QIAprep Spin Miniprep, QIAGEN Plasmid *Plus* Midi, QIAquick Gel Extraction, and QIAquick PCR purification kits from QIAGEN according the manufacturer's instructions. Once prepared, DNA was stored at −20 °C. DNA concentrations were measured with a NanoDrop 2000c spectrophotometer from Thermo Scientific.

Phenol:chloroform extraction

To purify DNA from whole cell extract, an equal volume of phenol:chloroform was added and the mixture centrifuged at 6000 g_n for 15 minutes. The top aqueous phase was pipetted to a new tube and the process repeated a total of 3 times.

Ethanol precipitation

The DNA solution was mixed with 2.5 volumes of 100 % ethanol and 1/10 volumes of Sodium Acetate (3 M, pH=5.2), and incubated at 4 °C for 15 minutes. When DNA concentrations were below 50 ng μl^{-1}, incubation was performed overnight. The DNA solution was then centrifuged at 18 000 g_n for 30 minutes at 4 °C and the supernatant discarded. The pellet was left to dry until all traces of solvent evaporated. The DNA pellet was resuspended in the desired solvent: H_2O when being used for transfection, EB buffer from QIAGEN in all other cases.

Agarose gel electrophoresis

Agarose gels with concentrations ranging from 0.6 % to 2.0 % were prepared with TAE buffer, and supplemented with ethidium bromide to a concentration of 200 $\mu g\, l^{-1}$. DNA samples were loaded into the gel with DNA loading buffer and a choice of loading dyes between bromophenol blue, cresol red, orange G, or xylene cyanol. In each case, the appropriate dye was chosen to avoid shadowing of the DNA bands. Electrophoresis was performed in Owl EasyCast electrophoresis chambers with 1 × TAE buffer at 80 V to 120 V until the required separation was achieved. Gels were visualised using a Alpha Innotech ChemiImager 5500 UV transilluminator.

DNA sequencing and oligonucleotide preparation

DNA sequencing was performed by LGC Genomics after cloning for sequence confirmation and avoiding unexpected mutations.

Oligonucleotides were ordered from Eurofins MWG operon in lyophilised format, dissolved in H_2O to a 100 µM concentration, and stored at −20 °C. A list of all designed oligonucleotides is shown in Appendix C.

Polymerase Chain Reaction

Different types of PCR experiments were performed for different purposes using a selection of DNA polymerases (Table 2.2). Taq polymerase with ThermoPol buffer was obtained from New England Biolabs. KOD Hot Start DNA polymerase with Mg^{2+} free buffer was obtained from Novagen. PCRs were performed on a Mastercycler epgradient thermocycler from Eppendorf.

Use of different DNA polymerases was based on their cost-benefit for each application. For example, screening clones requires a large number of reactions in tandem and the introduction of small mutations is of no consequence. For these two reasons, the much cheaper Taq DNA polymerase was used for screening despite its relatively low fidelity and amplification rates. However, in PCR mutagenesis the whole plasmid is synthesised anew and we have no reasonable method to verify the entire plasmid sequence. As a result, KOD DNA polymerase was used for PCR mutagenesis.

Colony PCR

When screening multiple clones after transformation and a set of appropriate primers was available, PCRs were carried out directly on the bacteria colonies by adding them directly to the reaction mixture (Table 2.2). Bacteria from individual colonies were used to simultaneously perform a PCR and start a small culture. Plasmid purification was performed on cultures whose sample was confirmed to be the desired product by a positive PCR result.

Gene cloning

PCRs were used to clone and subclone genes from genomic DNA and plasmids, into different vectors. Primers were usually extended to introduce restriction sites and create DNA linkers. Additional extensions of cytosine and guanine were added to the 5′ end of primers to account for the minimum required bp around the restriction sites as recommended by the restriction enzyme manufacturer, New England BioLabs.

Table 2.2: **PCR mixtures and conditions used.** Since each reaction was unique, with different pair of primers and template DNA, the optimal salt concentrations, temperature of the annealing step, and time of extension step actually used were sometimes different. Listed values correspond to the most common usage.

	cloning		colony	mutagenesis	screening
	genomic	plasmid			
Template (ng)	1000	50	n/a	750	50
DMSO (µl)	—	—	1	—	1
Buffer (×)	1	1	1	1	1
$MgSO_4$ (nmol)	1.5	1.5	—	1.5	—
dNTPs (mM each)	0.2	0.2	0.2	0.2	0.2
Primer (forward) (µM)	2	2	2	0.4	2
Primer (reverse) (µM)	2	2	2	0.4	2
DNA polymerase (U)	0.5 (KOD)	0.5 (KOD)	2.5 (Taq)	0.5 (KOD)	2.5 (Taq)
Total volume (µl)	25	25	25	25	25
Initialization	120s at 94°C	120s at 94°C	180s at 94°C	120s at 94°C	120s at 94°C
Denaturation	30s at 94°C	30s at 94°C	15s at 94°C	30s at 94°C	15s at 94°C
Annealing	20s at 58°C	20s at 58°C	15s at 58°C	20s at 58°C	15s at 62°C
Extension (per kbp)	30s at 68°C	20s at 72°C	60s at 72°C	30s at 68°C	60s at 72°C
Final extension	300s at 62°C	300s at 72°C	300s at 72°C	300s at 68°C	300s at 72°C
Final hold	4 °C	4 °C	4 °C	4 °C	4 °C
Number of cycles	30 ×	30 ×	30 ×	15 ×	20 ×

Site-directed mutagenesis

Site-directed mutagenesis by PCR was used to insert or correct mutations in plasmids. Oligonucleotides were designed by flanking the desired mutation with regions at least 16 bp long and equal predicted melting temperature (T_m) of at least 55 °C. When possible, the selected codon for mutation was the one with highest frequency in the organism used for expression in accordance with the codon usage database (*Nakamura et al.*, 2000). To remove the template DNA post amplification, 1.5 µl of the restriction enzyme DpnI was added directly to the PCR mixture and incubated overnight at 37 °C. Because DpnI requires the site to be methylated, it cleaves the template DNA which was synthesised in bacteria and leaves the newly *in vitro* synthesised DNA. 1 µl of the digestion product was used for transformation directly without any clean up step. Individual clones were screened by sequencing. In PCR site-directed mutagenesis, the newly synthesised strands will be linear and so cannot be used as template for the following PCR cycle. This means that the amplification is linear rather than exponential and so the PCR is not actually a chain reaction.

Screening plasmid

PCRs were frequently used to screen plasmids for DNA sequences in the absence of opportune restriction sites or even plasmid maps.

Plasmid construction

pBOS–H2B–EGFP D25G V118I Plasmid was provided by Prof. Kevin Sullivan (NUIG) from previous work (*Kanda et al.*, 1998). DNA sequencing identified gene *HIST1H2BJ* as the closest human H2B in RefSeq but with missense mutations D25G and V118I.

pBOS–H2B–EGFP The D25G and V118I mutations in pBOS–H2B–EGFP were corrected by PCR mutagenesis using primers AFG114 and AFG115, and AFG112 and AFG113 respectively. The resulting product encodes the *HIST1H2BJ* RefSeq gene.

pBOS–EGFP The pBOS–H2B–EGFP was digested with KpnI and BamHI and the band corresponding to the linearised vector, without the H2B sequence, was purified by gel extraction. This linearised vector was used as backbone vector for the other pBOS constructed plasmids.

pBOS–H2A–EGFP The *HIST1H2AB* gene sequence was amplified from HeLa genomic DNA with primers AFG116 and AFG118, the PCR product digested with KpnI and BamHI, and then ligated into the pBOS–EGFP vector. The *HIST1H2AB* was chosen because it encodes the same protein as the H2A used in previous work (*Flaus et al.*, 2004). The only other available alternative that encoded the same protein was *HIST1H2AE*. However, the *HIST1H2AE* sequence has a lower codon adaptation index and more stable predicted 5′ mRNA secondary structure. An accidental frameshift mutation near the stop codon was posteriorly fixed by PCR mutagenesis using primers AFG396 and AFG397.

pBOS–H2AX–EGFP The *H2AFX* gene sequence was amplified from HeLa genomic DNA with primers AFG130 and AFG131. The same strategy used in the cloning of pBOS–H2A–EGFP was used. An accidental frameshift mutation near the stop codon was posteriorly fixed by PCR mutagenesis using primers AFG400 and AFG401.

pBOS–H2AX–EGFP S139 mutants For H2AX S139 mutations, the *H2AFX* gene sequence was amplified from HeLa genomic DNA in the same reaction that the mutations were introduced since their location is close to the sequence 3′ end. Primers AFG132, AFG133, and AFG134, were used with AFG130 to introduce H2AX mutations S139A, S139D, and S139E respectively. The mutation to alanine blocks phosphorylation of S139, while mutation to aspartic and glutamic acid mimic phosphorylation of S139. The same strategy used in the cloning of pBOS–H2AX–EGFP was used, including the correction of a frameshift mutation with AFG400 and AFG401.

pBOS–H2A.Z–EGFP Due to the presence of introns in the *H2AFZ* gene sequence, the transcript sequence was amplified from HeLa cDNA provided by Dr. Nadine Quinn (*Quinn*, 2011). The amplification was performed with primers AFG121 and AFG122, and the product cloned into the pBOS–EGFP with the same strategy used for pBOS–H2A–EGFP. An accidental frameshift mutation near the stop codon was posteriorly fixed by PCR mutagenesis using primers AFG398 and AFG399.

pBOS–H3–EYFP.MC–N1 Plasmid was provided by Prof. Kevin Sullivan (NUIG). DNA sequencing identified the H3 sequence as the human RefSeq *HIST1H3B* gene, encoding H3.1 from histone cluster 1.

pBOS–H3–EYFP T45A and T45E Mutations to H3 T45 were inserted into pBOS–H3–EYFP.MC–N1 by PCR mutagenesis. The primers AFG151 and AFG152 were used to generate the T45E mutation, and AFG153 and AFG154 for T45A.

pBOS–H4–ECFP.M–N1 Plasmid was provided by Prof. Kevin Sullivan (NUIG). DNA sequencing identified the H4 sequence as the human RefSeq *HIST1H4J* or *HIST1H4K*, both of which have the same genomic sequence.

pBOS–H4–ECFP R45H The H4 R45H mutation was inserted into pBOS–H4–ECFP.M–N1 by PCR mutagenesis using the primers AFG124 and AFG125. The codon `CAC` was selected for the histidine amino acid due to its higher codon usage in the human genome (*Nakamura et al.*, 2000).

pBOS–H4–EYFP The plasmid pBOS–H3–EYFP.MC–N1 was digested with the restriction enzymes BamHI and NotI to extract the EYFP sequence which was purified by agarose gel extraction. The pBOS–H4 sequence was prepared in the same manner from pBOS–H4–ECFP.M–N1. The two fragments were ligated to construct pBOS-H4-EYFP.

pBOS–H4–EYFP R45H The same strategy used for the cloning of wild type pBOS–H4–EYFP was used for the R45H mutant using pBOS–H4–ECFP R45H.

PA-GFP The PA-GFP sequence used as an insert for other plamids was amplified from pPA-GFP–N1 with primers AFG478 and AFG479. The amplicon was purified by agarose gel extraction, digested with NotI and

BamHI, and then cleaned by PCR purification. The pPA-GFP–N1 plasmid was a kind gift from Chelly van Vuuren (NUIG).

pBOS–H2B–PA-GFP The plasmid pBOS–H2B–EGFP was digested with NotI and BamHI which removed the EGFP sequence. The vector was purified by agarose gel extraction and ligated with the PA-GFP insert. This strategy introduced a proline to arginine mutation in the linker sequence between H2B and PAGFP when compared to the linker sequence between H2B and EGFP.

pBOS–H3–PA-GFP The plasmid pBOS–H3–EYFP.MC–N1 was digested with NotI and BamHI which removed the EYFP sequence. The vector was purified by agarose gel extraction and ligated with the PA-GFP insert.

mCherry–α–tubulin Plasmid was a kind gift from Chelly van Vuuren (NUIG).

pMH3.2–614 The plasmid includes a mouse replication dependent histone H3.2 gene with upstream and downstream regulatory elements (*Taylor et al.*, 1986). It was provided by Prof. Kevin Sullivan. Due to the absence of convenient restriction sites in the plasmid, insertion of alternative histone gene sequence was performed by blunt-end ligation of PCR products. pMH3.2–614 was amplified with primers AFG417 and AFG418 which amplified the vector backbone, including the upstream and downstream regulatory elements but ignored the H3.2 coding sequence. The amplified sequence was purified by agarose gel extraction to be used in the cloning of pM-H2B–EGFP and pM-H3–EYFP.

pM-H2B–EGFP and pMH3–EYFP The H2B–EGFP insert sequence was generated by PCR amplification with primers AFG419 and AFG420, purified by agarose gel extraction, and ligated into the pMH3.2–614 vector. The primers for the insert, AFG419 and AFG420, were phosphorylated by T4 PNK prior to PCR since T4 PNK is more efficient on single stranded DNA. The same strategy was used for the cloning of pMH3–EYFP using primers AFG424 and AFG420.

2.3 Protein methods

Western blotting

Protein concentration determination

Concentration of protein was measured with Bradford reagent. 2 µl of the sample after sonication (§2.4) was mixed with 48 µl of H_2O and 50 µl of NaOH and incubated at 65 °C for 8 minutes before adding 900 µl of Bradford reagent from Pierce. The mixture was transferred to plastic cuvettes and the absorbance at 595 nm measured in a Shimadzu spectrophotometer. The values obtained were interpolated from a standard curve prepared using known concentrations of BSA.

SDS–PAGE

Resolving and stacking SDS–PAGE gels of 15 % to 5 % respectively, both with a cross-linking ratio of 37.5:1 as described in *Harlow and Lane* (1988) were used. The resolving gel was poured directly after addition of TEMED and it was covered with a layer of isopropanol during polymerisation to ensure a sharp interface between the resolving and stacking layers. Protein samples and markers were boiled at 99 °C for 3 minutes and each was loaded twice, with volumes for 3.3 µg and 16.5 µg of protein. Gels ran at 180 V for 1 hour in 1 × TG buffer.

Protein transfer

Protein transfer occurred through the wet transfer system. The SDS-PAGE gel was placed onto pre-cut nitrocellulose transfer membrane previously soaked in transfer buffer. It was then set between a pair of extra thick blotting paper and cushions before being placed inside a transfer apparatus. The transfer ran at 4 °C for 60 minutes at 100 V.

Probing of blot with antibody

Blocking of the membrane was performed with 10 % non-fat dry milk (Marvel) in 1 × TBS-T buffer at room temperature for 30 minutes. Blocking was followed by primary antibody incubation which occurred in 5 % non-fat dry milk in 1 × TBS-T overnight at 4 °C. Concentrations of antibody used were 1:500 and 1:20000 for anti-GFP (catalogue number 11 814 460 001 from Roche) and anti-H3 (code ab1791 from abcam). The membrane was then washed with 1 × TBS-T for 15 minutes 3 times before secondary antibody incubation in 5 % non-fat dry milk with 1 × TBS-T for 1 hour. The membrane was washed once more in the same conditions as before for the detection. All blocking, antibody incubation and washing steps occurred on a rocker.

Detection was performed using the SuperSignal West Pico Chemiluminescent Substrate from Pierce, adding 1:1 of the solutions and allowing it to incubate with the membrane for 5 minutes. The membrane was exposed to x-ray films for 10, 60, 5, 180 and 1800 seconds which were then developed.

2.4 Cell methods

Cell culture

HeLa (ATCC CCL-2) and HEp-2 cells were supplied by Dr. Agnieszka Kaczmarczyk (*Kaczmarczyk*, 2012) and Dr. Volker Döring (*Döring*, 2012) respectively. Primary horse fibrolasts were a gift from Prof. Elena Giulotto from the University of Pavia, Department of Genetics and Microbiology.

All cell lines were maintained at 37 °C and 5 % CO_2 in 10 cm diameter plates with 10 ml of their respective growth medium (Table 2.1). Once cells reached a confluence of 80 % to 90 %, they were trypsinised and split. First they were washed with DPBS and then incubated at 37 °C for 5 minutes in 2 ml of trypsin–EDTA. Finally, they were diluted 1:10 in fresh media and transferred to a new plate.

Cell stock storage

For long-term storage of HeLa cell lines, they were grown until they reached a confluence of 80 % to 90 % and then trypsinised. A volume of Freezing Media was added, equal to the volume of trypsin–EDTA, and 2 ml aliquots of cells transferred to cryotubes. Tubes were immediately wrapped in cotton and placed at –80 °C.

Whole cell extract

To obtain whole cell extracts, HeLa cells were trypsinised as usual. Growth medium added and the suspension was centrifuged at 900 g_n for 10 minutes. Subsequent steps were carried out at 4 °C and with previously chilled reagents. The supernatant was discarded and the pellet resuspended in 500 µl of chilled PBS before being sonicated 3 times at 40 % amplitude for 10 seconds. Avoiding the formation of foam at the top, 300 µl of suspension was transferred from the bottom of the tube to a new 1.5 ml tube and mixed with an equal volume of Laemmli buffer before being stored at −80 °C. 2 µl from the suspension was also transferred to a new tube for determination of protein concentration by Bradforf protein assay.

Genomic DNA extraction

To extract genomic DNA from HeLa cells, they were trypsinised and growth medium was added before counting with an hemocytometer. Cells were centrifuged at 1500 g_n for 10 minutes at 4 °C, the supernatant was discarded and the pellet resuspended in TE buffer to achieve a desired concentration of 4×10^7 cells/ml. 9 volumes of genomic lysis buffer was added and the mixture was incubated at 37 °C for 90 minutes. Proteinase K was then added to a final concentration of 100 µg ml^{-1} and the mixture was incubated at 50 °C for 3 hours and swirled every 20 minutes. DNA was then purified by phenol:chloroform extraction (§2.2) and ethanol precipitation (§2.2).

Viable cells count

Trypan blue was used to assess the number of viable cells. After trypsinisation, cells were diluted in growth media, to a final concentration of 2×10^5 cells/ml. To 0.5 ml of the cell suspension, 0.1 ml of 0.4 % Trypan Blue Stain was added and the mixture left for 5 minutes at room temperature before counting the cells in an hemocytometer and making a distinction between stained (non-viable) and non-stained (viable) cells.

Kill curve

Cells were trypsinised and plated at 25 % confluence on a 24-well plate. After 24 hours, medium was replaced using different concentrations of the antibiotic per well, 3 replicas for each. After 3 days, medium was replenished with the same antibiotic concentrations. After 4 days, a total of one week after addition of antibiotics, cells were trypsinised and a count of viable cells was performed with Trypan Blue.

The highest concentration tested with no viable cells in all replicas after 7 days was used for selection of transfected cells when establishing stable cell lines.

Transfection by lipofection

Cells were transfected using Lipofectamine 2000, a cationic lipid reagent, from Invitrogen. Cells were trypsinised as usual on the day before transfection and replated on 6 well plates (surface area of 9.5 cm^2/well) with 2.5 ml of growth medium so they would be 90 % confluent on the following day. For each well, two tubes with 250 µl of transfection medium were prepared, one with 7.5 µl of Lipofectamine 2000 and another with 3750 ng of DNA from a stock with concentration of 500 ng μl^{-1} and prepared by ethanol precipitation (§2.2). Both tubes were incubated at room temperature for 5 minutes, mixed together, and incubated again at room temperature for 20 minutes. Cells were washed with DPBS during this time and growth medium switched to 2 ml of transfection medium. The mixture was then added to the cells medium who were incubated at 37 °C for 6 hours after which time it was switched back to 0.5 ml of growth medium.

Transfection by electroporation

Cells were transfected by electroporation using an Amaxa nucleofector device with Ingenio Electroporation solution and cuvettes from Mirus Bio. Cells were grown to confluence and trypsinised as usual. Growth medium was added to a total volume of 8 ml, and 1 ml aliquots (for approximately 10×10^6 cells) were centrifuged at 160 g_n for 5 minutes. The supernatant was discarded and a volume of 2 µl of plasmid at a concentration of 500 ng μl^{-1} was added to the top of the pellet. The cell pellet was resuspended in 100 µl of Mirus Bio Ingenio Electroporation solution and transferred to 0.2 cm cuvettes for electroporation in an Amaxa nucleofector. According to the manufacturers instructions, the preset programs I-013 and O-17 were used to transfect HeLa and HEp-2 cells respectively.

Generation of stable cell lines

Cells were trypsinised and split to a low confluence (1:20) on 10 cm dishes 24 hours after transfection. After another 24 hours, the appropriate antibiotic was added to the medium with a final concentration as determined by performing an antibiotic kill-curve. Cell growth was observed daily followed and medium replaced every 3 days. As cell colonies started to be visible by the naked eye, approximately 3 weeks after plating, these were screened by fluorescence microscopy. Positive colonies were aspirated and moved into 24-well plates with 1 ml disposable pipette tips, and the thinnest extremity removed. Mixed populations were observed, so the cells with highest expression levels were FACS sorted to obtain homogeneous populations (Figure 2.1).

Fixation and staining

For microscopy visualisation, HeLa cells were grown directly on HCl washed coverslips as they have difficulty attaching to glass. At least 24 hours post plating and fixation, growth medium was removed and the cells washed with PBS once before incubation with 4 % formaldehyde in PBS for 4 minutes. This solution was then removed and the cells washed with H_2O two more times, after which coverslips were removed from the wells and left to air dry. For each coverslip, 2 µl of SlowFade Light Antifade kit from Molecular Probes was used for mounting the coverslip on a microscope slide. DAPI was added to the mounting media when needed. Coverslips were then sealed with a 1:1 mixture of clear nail polish and acetone and stored on a dark box at 4 °C.

Flow cytometry

Cells were trypsinised and after addition of medium, centrifuged at 160 g_n for 5 minutes. The supernatant was discarded and the pellet

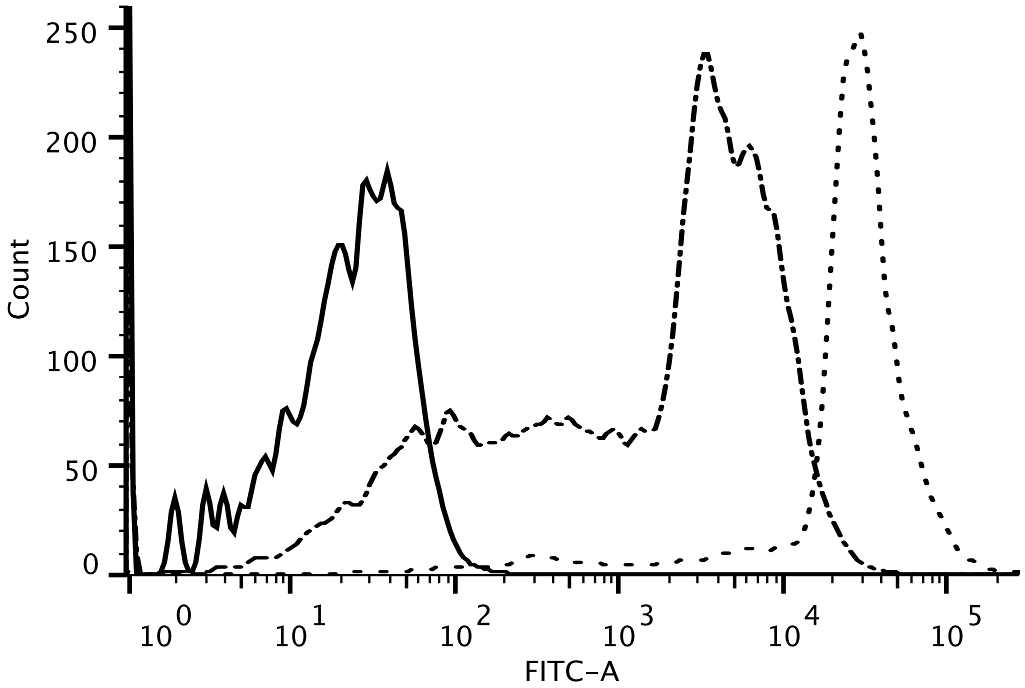


Figure 2.1: **Fluorescent intensity of HeLa cell populations by flow citometry.** The multiple populations with differing intensity values were sorted by FACS. The full line represents the intensity profile of HeLa wild type cells, the dash-dot line the mixed population of HeLa cells expressing H2B–EGFP, and the dotted line the homogeneous cell population after sorting.

washed with PBS. The sample was centrifuged one more time at $160\,g_n$ for 5 minutes. The supernatant was discarded again and the new pellet resuspended in Fluorescence-Activated Cell Sorting (FACS) buffer.

Populations with mixed levels of fluorescent intensity were frequently obtained while preparing stable cell lines. In such cases, cells with similar intensity of their corresponding fluorophore were sorted by FACS (Figure 2.1).

Both sample analysis and cell sorting were performed with live cells, the first with a BD FACSCanto II, and the later with a BD FACSAria II.

2.5 Software used

Image analysis was performed using GNU Octave version 4.2.0, and Octave Forge Image, Optim, Statistics, Bioformats, and Signal packages, versions 2.6.1, 1.5.2, 1.3.0, 5.3.3, and 1.3.2 respectively.

ImageJ, as distributed by the Fiji project, was routinely used for microscope image visualisation. PyMOL, by Schrödinger, LLC, was used for visualisation of protein structures. The European Molecular Biology Open Software Suite (EMBOSS) was used for analysis of codon usage, RNA folding, and reading of chromatogram files.

2.6 Source Code and Data

Source code to reproduce this thesis is available online, including its entire development history as a git repository, at `https://github.com/carandraug/phd-thesis.git`. The SCons build system is used to automate the build. Multiple dependencies will be required, namely GNU Octave, the Octave Forge Image package, BioPerl, and the Perl Bio-EUtilities module distribution, all of which will be confirmed by SCons. All the data required for the build is available online at Zenodo with record ID 377035 and DOI `10.5281/zenodo.377035`. The script `bootstrap.sh` is included in the thesis repository and can be used to download all data into the correct locations.

2.7 Developed software

Multiple software components were developed in the course of this thesis. When appropriate, these were contributed and merged into the software used (§2.5). Such work is described in Chapter 5.

Chapter **3**

Human Canonical Core Histone Catalogue

Abstract. Core histone proteins H2A, H2B, H3, and H4 are encoded by a large family of genes distributed across the human genome. Canonical core histones contribute the majority of proteins to bulk chromatin packaging, and are encoded in 4 clusters by 64 coding genes comprising 17 for H2A, 18 for H2B, 14 for H3, and 15 for H4, along with at least 18 total pseudogenes. The canonical core histone genes display coding variation that gives rise to 11 H2A, 15 H2B, 3 H3, and 2 H4 unique protein isoforms. Although histone proteins are highly conserved overall, these isoforms represent a surprising and seldom recognised variation with amino acid identity as low as 77 % between canonical histone proteins of the same type. The gene sequence and protein isoform diversity also exceeds commonly used subtype designations such as H2A.1 and H3.1, and exists in parallel with the well-known specialisation of variant histone proteins. RNA sequencing of histone transcripts shows evidence for differential expression of histone genes but the functional significance of this variation has not yet been investigated. To assist understanding of the implications of histone gene and protein diversity we have catalogued the entire human canonical core histone gene and protein complement. In order to organise this information in a robust, accessible, and accurate form, we applied software build automation tools to dynamically generate the canonical core histone repertoire based on current genome annotations and then to organise the information into a manuscript format. Automatically generated values are shown with a light grey background. Alongside recognition of the encoded protein diversity, this has led to multiple corrections to human histone annotations, reflecting the flux of the human genome as it is updated and enriched in reference databases. This dynamic manuscript approach is inspired by the aims of reproducible research and can be readily adapted to other gene families.

3.1 Introduction

Histones are among the most abundant proteins in eukaryotic cells and contribute up to half the mass of chromatin (*Alberts et al.*, 2014). The core histone types H2A, H2B, H3, and H4 define the structure and accessibility of the nucleosome as the fundamental repeating unit of genome organisation around which the DNA is wrapped (*Luger et al.*, 1997). In addition, the many chemically reactive sidechains of histones are post-translationally modified as a nexus for signalling and heritable epigenetics (*Kouzarides*, 2007).

Core histones are delineated as either canonical or variant based on their gene location, expression characteristics, and functional roles (Table 3.1). Canonical core histones contribute the majority of proteins to the bulk structure and generic function of chromatin, and are encoded by 82 genes in 4 clusters named HIST1-HIST4 in the human genome, of which 64 are coding genes and 18 are pseudogenes (Table 3.2).

Relationships within the histone family have been described using a variety of terminologies reflecting biochemical, functional, and genomic perspectives that are briefly described below and summarised in Table 3.3.

Biochemical perspective

Abundant histone proteins are readily isolated using their highly basic chemical character. Successive improvements in fractionation ultimately revealed 5 main histone types with nomenclature H1, H2A, H2B, H3, and H4 (*Bradbury*, 1975). An additional H1-related histone H5 is recognised in avian erythrocytes (*Kowalski and Pałyga*, 2011).

The demonstration of the nucleosome as the fundamental repeating unit of chromatin (*Kornberg*, 1974) showed that H2A, H2B, H3, and H4 associate as an octamer of two copies each within the nucleosome core particle. These four histones are referred to as core histones. In contrast, H1 associates with the linker DNA between nucleosome core particles and is referred to as a linker histone. The somatic H1 isoforms and tissue-specific variants are described elsewhere (*Harshman et al.*, 2013).

Arginine and lysine content was used as an early distinction between the histones (*Elgin and Weintraub*, 1975). The H1 linker histone has a low

Table 3.1: **Properties distinguishing canonical and variant core histone proteins.**

	Canonical	Variants
Expression timing	Replication dependent	Replication independent
Sequence identity	High	Low
Functional relationships	Isoforms	Specialised functions
Transcript stabilisation	Stem-loop	poly(A) tail
Gene distribution	Clusters	Scattered

Table 3.2: **Count of human canonical core histone coding genes and pseudogenes by histone cluster and type.** ψ indicates pseudogenes.

	H2A	H2B	H3	H4	Total
HIST1	12 + 6ψ	15 + 3ψ	10 + 1ψ	12 + 1ψ	49 + 11ψ
HIST2	4 + 0ψ	2 + 4ψ	3 + 2ψ	2 + 0ψ	11 + 6ψ
HIST3	1 + 0ψ	1 + 1ψ	1 + 0ψ	0 + 0ψ	3 + 1ψ
HIST4	0 + 0ψ	0 + 0ψ	0 + 0ψ	1 + 0ψ	1 + 0ψ
Total	17 + 6ψ	18 + 8ψ	14 + 3ψ	15 + 1ψ	64 + 18ψ

arginine/lysine ratio of 0.09 and became known as lysine-rich whereas the 4 core histones are arginine-rich with high arginine/lysine ratios of 0.91 in H2A, 0.40 in H2B, 1.38 in H3, and 1.25 in H4 type isoforms. Nevertheless, the core histones contain many lysines particularly in their N-terminal tails.

Separating histones by polyacrylamide gel electrophoresis (PAGE) using the strongly anionic detergent sodium dodecyl sulphate and neutral buffers (SDS PAGE) gives single bands for each histone type (*Shechter et al.*, 2007). However, PAGE with non-ionic detergent Triton X–100 and urea as denaturants in acid buffers (TAU or AUT PAGE) allows the separation of histone types into multiple bands due to post-translational modifications and differences at specific amino acids in the polypeptides (*Zweidler*, 1977). These TAU PAGE separations gave rise to subtype designations H2A.1, H2A.2, H3.1, H3.2, and H3.3.

Functional perspective

Canonical core histone expression is significantly elevated during S phase to provide chromatin packaging for DNA duplicated during

Table 3.3: **Terminology describing histone variation.**

Allelic variants

Copies of canonical histone type genes, possibly with different sequences. Not located at same exact chromosomal locus as expected for alleles. Not histone variants. See also “isoforms”.

Canonical core histones

Core histones with properties described in Table 3.1 that contribute the majority of core histones in chromatin encompassing multiple protein isoforms.

Core histones

Histones that form part of the nucleosome core particle wrapping 147 bp of DNA, comprising types H2A, H2B, H3, and H4. Encompasses both canonical and variant histones.

Heteromorphous variants

Core histone variants with distinct function and localisation that are readily separated by gel electrophoresis.

Homomorphous variants

Canonical core histone subtypes requiring high resolution separation methods such as TAU PAGE. Synonym for “subtypes”.

Families

Synonym for “types”.

Isoforms

Proteins with high sequence identity and largely equivalent function. Functional equivalence has not been demonstrated for canonical histone isoforms.

Linker histones

Histones binding to linker DNA adjacent to the nucleosome core particle. The two linker histone types are H1 and H5.

Non-allelic variants

Synonym for “variant histones”. See also “allelic variants”.

Replacement histones

Synonym for "variant histones". Named because they can replace canonical core histones assembled in S phase.

Replication-dependent histones

Synonym for "canonical core histones". Named because expression occurs primarily in S phase.

Replication-independent histones

Synonym for "variant histones". Named because expression is not predominantly in S phase.

Subtypes

Canonical core histone type isoforms separable by TAU PAGE (e.g. H2A.1 and H2A.2). There is not necessarily functional evidence for differences between subtypes.

Types

Histone proteins sharing sequence homology that participate in specific combinations to define the repeating nucleosome structure. The 5 histone types are H1, H2A, H2B, H3, and H4.

Variant histones

Core histones with properties described in Table 3.1. Contribute a minor proportion of histones in chromatin and perform specialised functions.

Wild type histones

Synonym of "canonical core histones".

replication (*Wu and Bonner*, 1981). This led to their description as "replication dependent", although a supply of canonical histones is inevitably required to partner variants throughout the cell cycle. Metazoan canonical core histone genes are distinctive because they lack introns and give rise to non-polyadenylated protein coding transcripts. Transcript turnover is regulated via a highly conserved 3' stem-loop (*Marzluff et al.*, 2008) (Table 3.1).

In contrast, variant histones such as H2A.Z, TH2B, H3.3, and CENP-A have reduced sequence identity and lower abundance (*Talbert and Henikoff*, 2010). They play functionally specific roles and are mostly expressed outside S phase, so are described as “replication independent”. Since histone variants are interpreted as taking the place of equivalent canonical core histone types, they are referred to as “replacement” histones.

Genomic perspective

Canonical core histone genes are found in 4 clusters. The multiple gene copies in these clustered arrays are sometimes confusingly referred to as “allelic” and the resulting combined protein isoforms are often considered to be “wild type” although both genes and protein products display variation in primary sequence and relative abundance, and their functional equivalence has not been tested. In contrast, almost all variant histones are encoded by single genes dispersed in the genome with typical properties including introns, alternative splicing, and polyadenylated transcripts (Table 3.1).

Cataloguing of canonical core histone diversity

Despite the importance of histones for chromatin organisation and extensive interest in their role in epigenetics and regulation, the curation and classification of human histone gene and protein sequences has not been systematically revisited since the landmark survey by *Marzluff et al.* (2002). The Histone Database has for many years provided an online database of histone sequences from across eukarya using sequence homology searching (*Baxevanis and Landsman*, 1996, 1997, 1998; *Makalowska et al.*, 1999; *Sullivan et al.*, 2000, 2002; *Sullivan and Landsman*, 2004; *Mariño-Ramérez et al.*, 2006, 2011; *Draizen et al.*, 2016), while some manually collated listings including human histones have also been undertaken (*Ederveen et al.*, 2011; *Khare et al.*, 2012; *Kennani et al.*, 2017). This reflects continual revisions in genome sequence databases and the ongoing need for information about human histones. In fact, differences in 38 canonical core histone gene details (Table A.1) have so far accumulated since the original 2002 survey due to rich annotations continuing to propagate into reference sequence databases.

In this manuscript we provide a comprehensive catalogue of canonical core histone genes, encoded proteins, and pseudogenes using reference genome annotations as the originating source. This reveals a surprising and seldom recognised variation in encoded histone proteins that exceeds commonly used subtype designations but whose functional implications have not been investigated.

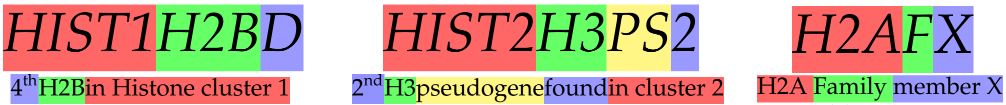
Since curation and annotation of reference databases are dynamic and evolving, we have implemented the manuscript so that it can be regenerated at any time from the most current data in the NCBI RefSeq database in order to maintain its ongoing value as a reference source in an accessible format. All figures and tables were automatically generated using NCBI RefSeq data from 13th March 2017. All dynamically generated values in this copy of the text are displayed with a light grey background. The manuscript generation process has remained stable in our laboratory for several years and represents an example of "reproducible research" (*Schwab et al.*, 2000) that provides a novel model for cataloguing gene families.

3.2 Histone genes

Canonical core histone gene nomenclature

Canonical core histone genes adhere to a Human Genome Organisation (HUGO) Gene Nomenclature Committee (HGNC) endorsed system derived from the cluster number and position relative to other histones (*Marzluff et al.*, 2002). This superseded an earlier arbitrary scheme with backslashes (e.g. H2b/b) that preceded genome sequencing (*Albig et al.*, 1997; *Albig and Doenecke*, 1997).

The canonical core histone gene symbols are divided into 3 parts: HIST cluster, histone type, and identifier letter for the order relative to other histone genes of the same type in the same cluster (Figure 3.1(a)). For example, *HIST1H2BD* is nominally the fourth H2B coding gene in the HIST1 cluster. Identifiers are ordered by their genomic coordinates starting at the telomere of the short arm (*Marzluff et al.*, 2002).


HIST1H2BD
4th H2B in Histone cluster 1
HIST2H3PS2
2nd H3 pseudogene found in cluster 2
H2AFX
H2A Family member X

(a) canonical coding gene (b) canonical pseudogene (c) variant

Figure 3.1: **Histone gene nomenclature.** (a) Canonical core histone gene names encode relative genomic order by cluster. (b) Pseudogenes named since 2002 include cluster, PS label, and discovery order identifier. (c) Most variant core histone genes are identified by type then F for family and identifier letter.

Two exceptions were originally applied to these simple naming rules (*Marzluff et al.*, 2002). Firstly, the positional identifier is omitted if there are no other histones of the same type in the cluster, so *HIST3H2A* is the sole H2A gene in HIST3. Secondly, the human and mouse histone clusters are largely syntenic so positional identifier letters for missing orthologs were skipped to maintain the equivalence of gene symbols. Consequently there is no human *HIST1H2AF* to accommodate *Hist1h2af* in mouse while keeping both –E and –G identifiers consistent for mouse and human orthologs.

Several new histone genes have been uncovered since the original naming (Table A.1) and this required additional nomenclature exceptions. For example, new H2A encoding genes were identified in human and mouse HIST2 clusters preceding *HIST2H2AA*/*Hist2h2aa*. This led to the renaming of the gene to *HIST2H2AA3*/*Hist2h2aa3* and the addition of a new human gene as *HIST2H2AA4*. There are no human orthologs of the new mouse *Hist2h2aa1* and *Hist2h2aa2*.

Furthermore, no distinction was originally made between pseudogenes and functional coding genes, so *HIST3H2BA* is a pseudogene whereas neighbouring *HIST3H2BB* is the only functional H2B coding gene in HIST3. The HGNC definition of a pseudogene is a sequence that is generally untranscribed and untranslated but which has at least 50% predicted amino acid identity across 50% of the open reading frame to a named gene (*Gray et al.*, 2013). Newly uncovered histone pseudogenes are now suffixed with PS and a number in order of discovery (Figure 3.1(b)), such as *HIST1H2APS5* as the fifth H2A pseudogene discovered in HIST1. However, the pseudogenes named in the original classification retain their symbols without PS. This means that the absence of a PS suffix does not indicate a functional gene, and that there is no positional information in the gene symbols of most pseudogenes. Recently, some pseudogenes have also been shown to be functional histone variants (*Taguchi et al.*, 2017).

Histone gene clustering

The human canonical core histone genes are located in clusters HIST1 to HIST4, named in order of decreasing histone gene count (Table 3.2). HIST1 is the major histone gene cluster at locus 6p22.1–6p22.2 with 49 functional core histone genes plus all canonical H1 histones, representing 77 % of all canonical core histone genes. HIST2 at locus 1q21.2 contains 11 coding genes, HIST3 at locus 1q42.13 contains 3 coding genes, and HIST4 at locus 12p12.3 contains 1 coding gene.

HIST1 and HIST2 are both contiguous high density arrays of histone genes. HIST1 spans 15.0 Mbp and is the second most gene dense region in the human genome at megabase scale after the MHC class III region (*Xie et al.*, 2003). The only non-histone protein coding gene located within the principal region of this cluster is *HFE*, encoding the hemochromatosis protein (*Albig et al.*, 1998), although a number of other genes are located proximal to the outlying *HIST1H2AA* and *HIST1H2BA* pair.

It has also been argued that histone clustering does not contribute to gene conversion (*Nei and Rooney*, 2005). HIST1 is located towards the distal end of the major histocompatibility complex (MHC) in the extended class I region (*Gruen et al.*, 1996; *The MHC sequencing consortium*, 1999) and it has even been suggested that this proximity may suppress recombination due to the observed local linkage disequilibrium (*Gruen and Weissman*, 1997). In contrast, HIST2 may be prone to deletions and frequent rearrangements (*Sharp et al.*, 2006; *Brunetti-Pierri et al.*, 2008)

The functional significance of histone gene clustering remains to be demonstrated. It has been suggested that clustering may facilitate coordinate regulation (*Eirin-Lopez et al.*, 2009; *Osley*, 1991), so interpreting such a functional relationship with genome organisation requires an accurate catalogue of the histone genes and their individual roles.

Conversely, progress in understanding histone gene function suggests there is a need to update the definitions of the canonical histone gene clusters. For example, the *HIST1H2AA* and *HIST1H2BA* genes are located 300 kbp upstream of the rest of the HIST1 cluster and separated from other cluster members a number of non-histone genes. Although the two genes have been assigned canonical histone gene names, the *HIST1H2BA* gene in fact encodes the most divergent canonical H2B protein which is also known as TH2/TH2B/hTSH2B and considered to be a testes-specific histone protein (*Zalensky et al.*, 2002; *Li et al.*, 2005; *Shinagawa et al.*, 2014a). Immediately adjacent is *HIST1H2AA* encoding a H2A protein isoform of similarly high variation. The syntenic rat orthologues of *HIST1H2AA* and *HIST1H2BA* are divergently transcribed specifically in testes (*Huh et al.*, 1991), and the mouse orthologues H2AL1/TH2a and TH2B have been shown to participate in gametogenesis (*Govin et al.*, 2007) as well as to enhance stem cell reprogramming (*Shinagawa et al.*, 2014b; *Padavattana et al.*, 2015). It is therefore possible that *HIST1H2AA* and *HIST1H2BA* are undergoing sub-functionalisation and could be reclassified as variant genes encoding histone proteins H2A.L and H2B.1 (*Talbert et al.*, 2012). This would in turn require gene renaming and a recalculation of the length of the HIST1 cluster.

A similar case of updates in definition may apply for the small HIST3 cluster 80 Mbp downstream of HIST2. HIST3 contains protein coding genes *HIST3H2A*, *HIST3H2BB*, and *HIST3H3*. *HIST3H3* encodes testes-specific H3T/H3.4 that has distinctive biochemical properties and appears to be a histone variant (*Witt et al.*, 1996; *Kurumizaka et al.*, 2013), while the proteins encoded by *HIST3H2A* and *HIST3H2BB* are amongst the most divergent canonical histones of their types.

The examples of the *HIST1H2AA* and *HIST1H2BA* pair and the HIST3 cluster illustrate the evolving nature of the human canonical histone complement. This demonstrates the need for a dynamic approach to classification and presentation.

Histone gene sequences

Despite the high conservation of canonical core histone protein sequences, the coding regions of these genes exhibit considerable variation (Figure A.1, A.2, A.3, and A.4).

These differences are largely located in the third base position of codons. The mean number of synonymous substitutions per site (d_S) in the sets of histone gene type isoforms is 3.8 for H2A, 1.7 for H2B, 2.5

for H3, and 2.8 for H4 (Table A.4). This is consistent with observations that synonymous codon divergence far exceeds non-synonymous variation for histone genes across eukaryotes (*Piontkivska et al.*, 2002; *Rooney et al.*, 2002). It supports a hypothesis that histone protein sequence conservation results from birth and death evolution through strong selective pressure at the protein level (*Nei and Rooney*, 2005). Despite the level of synonymous substitution, codon usage is strongly biased towards the most frequently used human codons for most amino acids (Table A.5). This suggests that histone translation may be sensitive to tRNA abundance.

Histone variant genes

The 32 annotated human histone variant genes are listed in Table A.6 for completeness. Histone variant gene families comprise only one or a few copies dispersed across the genome outside the canonical histone clusters. For example, the three H3.3 variant encoding genes are located on chromosomes 1, 12, and 17, far removed from other histone genes.

Most variant gene symbols have a separate 3 part nomenclature consisting of histone type, F for family, and an identifier letter (Figure 3.1(c)). Increasing interest in histone variant function (*Maze et al.*, 2014) coupled with a variety of usages and conflicts between species has led to guidelines for improved consistency in histone variant nomenclature (*Talbert et al.*, 2012).

3.3 Histone transcripts

Canonical core histone transcripts are among the only protein coding messages that do not undergo polyadenylation, instead carrying a unique stem-loop structure (Table 3.1). They do have a 5′ 7–methyl–guanosine cap (*Marzluff et al.*, 2008).

Canonical core histone gene transcription is regulated by cell-cycle dependent phosphorylation of the histone-specific Nuclear Protein of the Ataxia-Telangiectasia locus (NPAT) coactivator and interaction with the accessory protein FADD-Like interleukin-1β-converting enzyme/caspase-8-ASsociated Huge protein (FLASH), resulting in assembly of histone locus bodies coordinating factors responsible for transcription and processing (*Marzluff et al.*, 2008; *Rattray and Müller*, 2012; *Hoefig and Heissmeyer*, 2014). Variability is observed in canonical core histone isoform gene transcription, both by analysis of non-polyadenylated transcripts (*Yang et al.*, 2011) and RNA polymerase II promoter occupancy (*Ederveen et al.*, 2011).

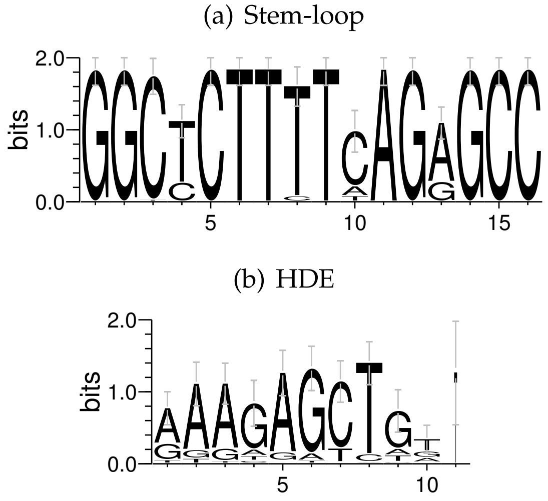
Overall there is estimated to be a 35 fold increase in mammalian canonical histone transcripts during S phase, principally as a result of a 10 fold increase in mRNA stabilisation via stem-loop dependent mechanisms, and a 3–5 fold up-regulation in canonical core histone gene transcription (*Harris et al.*, 1991).

This post-transcriptional regulation is achieved by a stem-loop encoded in the mRNA 3′ untranslated region that is recognised by spliceosome-related RNA processing and stabilisation complexes (*Tan et al.*, 2013). The start location of the annotated stem-loops in human canonical histone transcripts ranges from 22 to 67 bp after the stop codon with a modal value of 35 bp. The sequence logo of aligned stem-loops confirms that the stem-loop is highly conserved (Figure 3.2(a)).

The RNA stem-loop structure is bound by the Stem-Loop Binding Protein (SLBP) which is up-regulated 10–20 fold during S phase to stabilise histone mRNAs (*Whitfield et al.*, 2000). Immediately downstream of the stem-loop a purine-rich Histone Downstream Element (HDE) interacts with U7 snRNA to direct efficient 3′ end processing. Although this is not an annotated feature of histone genes, alignment of the canonical HDE (*Georgiev and Birnstiel*, 1985) to all canonical histone genes shows the modal location of the HDE is 16 bp downstream of the stem-loop. The sequence logo confirms that this feature is also highly conserved (Figure 3.2(b)).

Stabilisation and processing extend the half life of canonical core histone mRNAs during S phase and contribute to increased histone translation efficiency, enabling rapid production of histones to package the newly duplicated genomes.

Although the vast bulk of canonical core histone transcripts appear to be regulated by this mechanism, 1-5% of transcripts are found to be 3′ polyadenylated (*Yang et al.*, 2011) and some genes have annotations indicating two transcripts differing in whether they have stem-loop or polyadenylation signals (Table A.7).


(a) Stem-loop

(b) HDE

Figure 3.2: **Stem-loop and HDE alignment.** Sequence logos for (a) annotated stem-loops and (b) Histone Downstream Elements (HDEs) identified by homology for all canonical core histone gene 3′ untranslated regions (UTRs).

Core histone variant transcripts lack a 3′ stem-loop and are polyadenylated in the same way as most protein coding genes. The exception is H2AX, which has alternatively processed transcripts exhibiting both the stem-loop characteristic of a canonical core histone and a poly(A) tail found on variant core histones (*Mannironi et al.*, 1989; *Pinto and Flaus*, 2010).

3.4 Histone proteins

Many depictions of chromatin imply that canonical core histone protein types behave as a single "wild type" protein. This common assumption is based on the relatively high identity of histone sequences between isoforms and a focus of interest in functional roles of histone variants.

However, the very strong selection pressure on amino acid sequences encoded by canonical core histone genes (*Nei and Rooney*, 2005), the roles of specific amino acid differences in observed proteins (*Maze et al.*, 2014), and the consequences of small variations in proteins on the structure of nucleosomes (*Kurumizaka et al.*, 2013), all suggest that the minor differences in the encoded canonical core histone protein isoforms could have functional implications.

Encoded isoform variation in canonical histone types H2A and H2B is pronounced. 11 H2A and 15 H2B distinct polypeptides are encoded by 17 H2A and 18 H2B coding genes, repectively. In contrast, only 3 H3 and 2 H4 distinct polypeptides are encoded by 14 H3 and 15 H24 coding genes repectively (Table 3.2). Although most variation is a result of amino acid substitutions, some H2A and H2B isoforms also show length differences.

The nomenclature for histone protein isoforms follows directly from the gene names (*Marzluff et al.*, 2002), and supersedes the earlier nomenclature with forward slash which used different isoform letters (*Albig et al.*, 1997; *Albig and Doenecke*, 1997). For example, HIST1H2BB was known as H2A/f in the earlier nomenclature. The encoded proteins described below are products of genes and transcripts listed in table Table A.2. Human canonical core histone polypeptides have typically been numbered with omission of the N-terminal methionine since this is likely to be removed because most sequences have a small hydrophilic amino acid as the second residue (*Xiao et al.*, 2010). This convention predates the Human Genome Variation Society (HGVS) recommendation to include the initial methionine as residue 1. We have omitted the N-terminal methionine on figures, alignments, and amino-acid numbering for consistency.

Canonical H2A isoforms

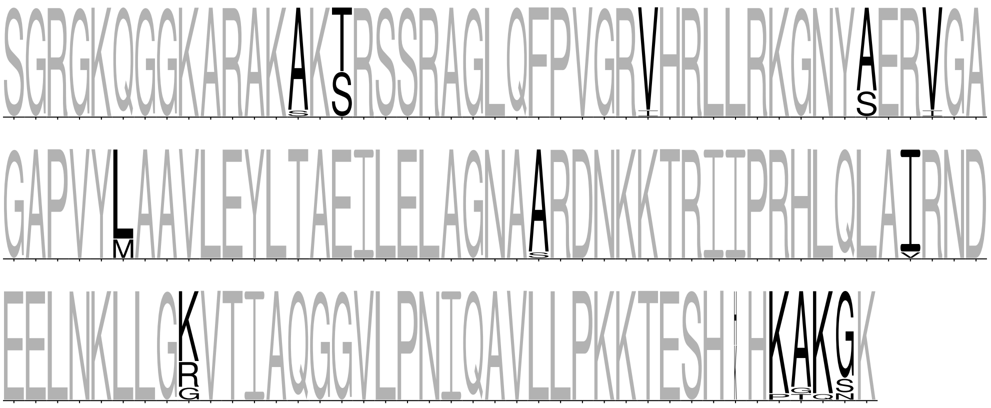
Canonical H2A genes encode 11 different protein isoforms with pairwise identity down to 88 % (Table 3.4). These are separable by TAU PAGE into two bands identified as H2A.1 and H2A.2. TAU PAGE distinguishes leucine from methionine at residue 51 (*Franklin and Zweidler*, 1977; *Zweidler*, 1977), implying there are up to 9 different protein sequences in H2A.1 and 2 in H2A.2. There is no concordance between the bands H2A.1 and H2A.2 and the location of isoform-encoding genes in HIST1 and HIST2 histone gene clusters. For example, the HIST2 cluster contains genes encoding isoforms with both Leu51 and Met51 while HIST1, HIST2, and HIST3 clusters all contain genes encoding H2A isoforms with Leu51. No functional distinction between H2A.1 and H2A.2 has been reported.

Table 3.4: **Canonical H2A encoded protein isoforms.** Upper panel shows isoform variations relative to most common isoform using HGVS recommended nomenclature (*den Dunnen and Antonarakis*, 2003). Lower panel shows sequence logo of all isoforms aligned with invariant residues in grey.

Most common isoform (129 amino acids; HIST1H2A –G, –I, –K, –L, –M)

```
SGRGKQGGKARAKAKTRSSRAGLQFPVGRVHRLLRKGNYAERVGAGAPVYL
AAVLEYLTAEILELAGNAARDNKKTRIIPRHLQLAIRNDEELNKLLGKVTI
AQGGVLPNIQAVLLPKKTESHHKAKGK
```

HIST1H2AA	A14S T16S V30I V43I A70S K99G H123_H124insH K127_G128QS
HIST1H2AB	A40S K99R
HIST1H2AC	T16S K99R
HIST1H2AD	A40S
HIST1H2AE	A40S K99R
HIST1H2AH	G128_K129del
HIST1H2AJ	A126T G128_K129del
HIST2H2AA3	T16S L51M
HIST2H2AA4	T16S L51M
HIST2H2AB	T16S I87V K99G H124_A126KPG G128N
HIST2H2AC	T16S L51M H124del G128S
HIST3H2A	T16S A40S K99R


SGRGKQGGKARAKAKTRSSRAGLQFPVGRVHRLLRKGNYAERVGA
S S I S I
GAPVYLAAVLEYLTAEILELAGNAARDNKKTRIIPRHLQLAIRND
M S V
EELNKLLGKVTIAQGGVLPNIQAVLLPKKTESHHKAKGK
R G S

Excluding *HIST1H2AA* discussed above, the sites of difference in two or more isoforms of canonical H2A are serine or threonine at residue 16, alanine or serine at residue 40, lysine, arginine, or glycine at residue 99, and the C-terminal residues from 124 onwards (Table 3.4). All these sites have implications for post-translational modifications.

Canonical H2B isoforms

Canonical H2B has more isoforms than the other histone types (Table 3.5) with 15 unique proteins diverging down to 77% pairwise identity. Nevertheless, all isoforms appear to migrate together in TAU PAGE.

There is significant variability between isoforms in the N-terminal region and this is one of the most variable sites between canonical core histones of different species. Human H2B has a unique and distinctive N-terminal proline-acidic-proline motif (PEP/PDP) followed by a very variable residue that can be alanine, serine, threonine, or valine (Table 3.5).

Excluding *HIST1H2BA* discussed above, the remaining isoform differences in two or more isoform are mainly the chemically similar valine or isoleucine at residue 39, and serine to glycine and alanine at residues 75 and 124 respectively, which have have post-translational modification implications. A number of H2B genes are annotated to have multiple transcripts, although in most cases these transcripts encode identical protein isoforms. Variation in H2B transcript isoform levels between multiple cancer cell lines has been observed (*Molden et al.*, 2015), although the functional implications are unknown.

Canonical H3 isoforms

Canonical H3 genes encode only 3 different protein isoforms (Table 3.6). The majority of H3 genes are in the HIST1 cluster and encode a single polypeptide sequence (*Ederveen et al.*, 2011) whereas the canonical H3 genes in HIST2 encode a distinct isoform with the interesting difference of serine instead of cysteine at residue 96 that is separable by TAU PAGE (*Franklin and Zweidler*, 1977). By apparent coincidence this means HIST1-encoded canonical H3 isoforms are identified as H3.1 and the HIST2-encoded copies are H3.2. As discussed above, *HIST3H3* with four amino acid differences appears to be the largely testes specific variant H3T that has been assigned as H3.4 in variant nomenclature (*Talbert et al.*, 2012) even though it would be predicted to migrate with H3.1 in TAU PAGE (*Franklin and Zweidler*, 1977).

Table 3.5: **Canonical H2B encoded protein isoforms.** Upper panel shows isoform variations relative to most common isoform using HGVS recommended nomenclature (*den Dunnen et al.*, 2016). For clarity, isoforms encoded by multiple transcripts of a single gene are distingushed by a numerical suffix (Table A.2). Lower panel shows sequence logo of all isoforms aligned with invariant residues in grey.

Most common isoform (125 amino acids; HIST1H2B –C, –E, –F, –G, –I)

```
PEPAKSAPAPKKGSKKAVTKAQKKDGKKRKRSRKESYSVYVYKVLKQVHPD
TGISSKAMGIMNSFVNDIFERIAGEASRLAHYNKRSTITSREIQTAVRLLL
PGELAKHAVSEGTKAVTKYTSSK
```

HIST1H2BA	P3_A4VSS S6G P8_P10TIS S14F T19V A21T D25E S32T V39I V41I G60S N67T G75S N84S T90S
HIST1H2BB	A4S V18I V39I
HIST1H2BD.1	A4T
HIST1H2BD.2	A4T
HIST1H2BH	E2D
HIST1H2BJ	V39I S124A
HIST1H2BK.1	S124A
HIST1H2BK.2	S124A
HIST1H2BL	P3L G75S
HIST1H2BM	A4V A9V V18_T19IN
HIST1H2BN	A4S
HIST1H2BO	E2D V39I
HIST2H2BE	V39I
HIST2H2BF.1	E2D A21V
HIST2H2BF.2	E2D A21V *126Lext*7
HIST3H2BB	E2D A4S S32G V39I G75S I94V

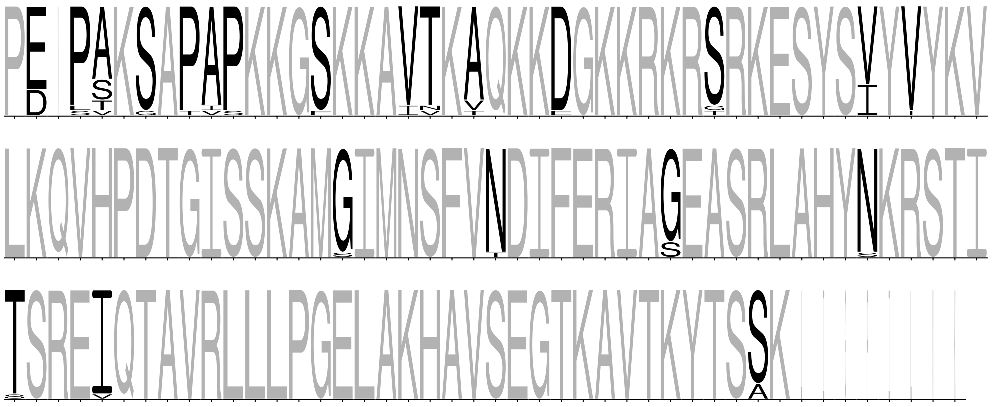
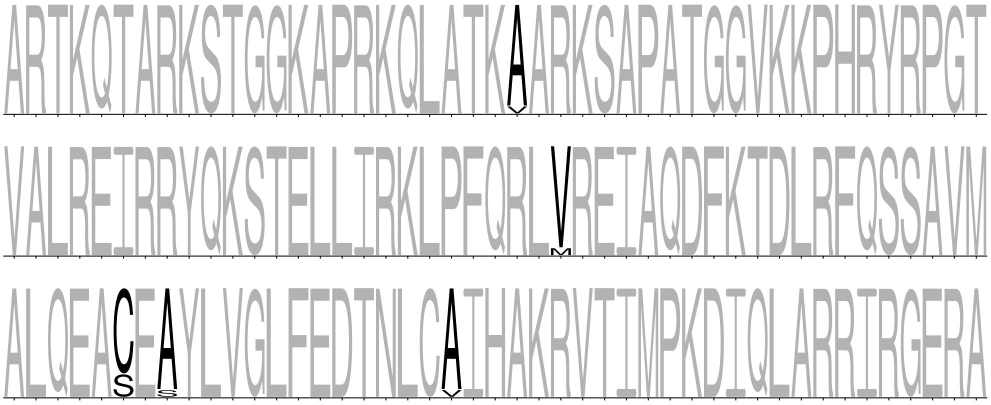


Table 3.6: **Canonical H3 encoded protein isoforms.** Upper panel shows isoform variations relative to most common isoform using HGVS recommended nomenclature (*den Dunnen and Antonarakis*, 2003). Lower panel shows sequence logo of all isoforms aligned with invariant residues in grey.

Most common isoform (135 amino acids; HIST1H3 –A, –B, –C, –D, –E, –F, –G, –H, –I, –J) ARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREI RRYQKSTELLIRKLPFQRLVREIAQDFKTDLRFQSSAVMALQEACEAYLVG LFEDTNLCAIHAKRVTIMPKDIQLARRIRGERA	
HIST2H3A	C96S
HIST2H3C	C96S
HIST2H3D	C96S
HIST3H3	A24V V71M A98S A111V



Canonical H4 isoforms

H4 is the most homogeneous of all canonical core histones, with all but one of the 15 genes encoding an identical protein sequence (Table 3.7). These genes are located across HIST1, HIST2, and the isolated *HIST4H4*.

The divergent *HIST1H4G* gene in the middle of the HIST1 cluster encodes an isoform with 15 amino acid differences and a deletion of the C-terminal 5 residues. This gene is annotated as being transcribed and merits further investigation.

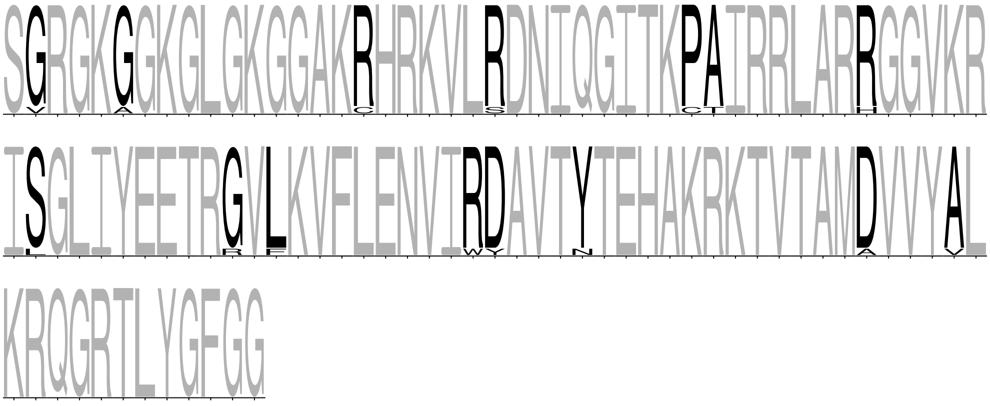
3.5 Reproducible research

A considerable number of builds and updates of the human genome sequence have been made since the last major survey of human canoni-

Table 3.7: **Canonical H4 encoded protein isoforms.** Upper panel shows isoform variations relative to most common isoform using HGVS recommended nomenclature (*den Dunnen and Antonarakis*, 2003). Lower panel shows sequence logo of all isoforms aligned with invariant residues in grey.

Most common isoform (102 amino acids; HIST1H4 –A, –B, –C, –D, –E, –F, –H, –I, –J, –K, –L; HIST2H4 –A, –B; HIST4H4)
`SGRGKGGKGLGKGGAKRHRKVLRDNIQGITKPAIRRLARRGGVKRISGLIY`
`EETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYALKRQGRTLYGFGG`	

HIST1H4G	G2V G6A R17C R23S P32_A33CT R40H S47L G56R L58F R67_D68WY Y72N D85A A89V Y98_G102del


SGRGKGGKGLGKGGAKRHRKVLRDNIQGITKPAIRRLARRGGVKR
ISGLIYEETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYAL
KRQGRTLYGFGG

cal histone genes in 2002 (*Marzluff et al.*, 2002). This has resulted in 20 canonical core histone genes added, 3 removed, and 15 sequences being updated for the current RefSeq release (Table A.1). These changes reflect the ongoing nature and challenges of genome curation and annotation in reference databases (*Bork and Koonin*, 1998).

It is inevitable that sequences and annotations will continue to be revised based on continuous curation and improved experimental insights. This in turn prompts reevaluation of assumptions about their biological contributions, illustrated by the recognition that some genes annotated as canonical core histone isoforms are testes-specific variants (*Talbert et al.*, 2012), and by new observations of cell type-specific expression (*Molden et al.*, 2015).

It is important for communities of researchers to contribute to ongoing formal curation of genomics resources by feeding back information to database maintainers (*Stein*, 2001), although many are unaware of this opportunity (*Holliday et al.*, 2015). In the course of this work we have suggested a significant number of proposals for improvements to RefSeq curators. These were identified automatically by scripted tests for consistency and uniformity between the database annotations and expected properties of histones. Table A.7 lists the current apparent anomalies from these tests and is the basis for ongoing discussions with curators.

Dynamic data can be presented in specialist online database resources such as the Histone Database (*Draizen et al.*, 2016) and HIstome (*Khare et al.*, 2012). These database interfaces provide comprehensive access to data but have limited curation and are difficult to cite. They can also cease to be updated or become inaccessible. In contrast, manuscripts are the established method of scientific communication because they provide descriptive context and accessible formatting for readers, can be peer reviewed, and have well-established methods for citation. There are established mechanisms for permanent archival of published manuscripts. However, static catalogues in manuscripts such as earlier surveys of histone genes (*Albig and Doenecke*, 1997; *Marzluff et al.*, 2002) inevitably become supersed by improvements in source data and cannot be updated to remove errors.

A self-updating manuscript bridges the features of dynamic database and static manuscript presentation styles by providing convenient access to the most current information. It encourages communities of researchers to directly feed into the formal curation process and enables rapid leveraging of the most current data for relatively stable gene families.

Implementating a dynamic manuscript is also an example of "reproducible research" (*Gentleman*, 2005; *Yale Law School Roundtable on Data and Code Sharing*, 2010) that can address the topical challenge of irreproducibility in biological data and interpretation (Nature Editorial, 2012; *Ince et al.*, 2012).

This manuscript is generated directly from sequences and annotations in the core NCBI RefSeq resource (*O'Leary et al.*, 2016). The processing system for the manuscript does not cache intermediate information, so all changes contributed by the community of histone researchers and curated by professional database maintainers are directly and automatically reflected at each manuscript refresh. All scripts including instructions for automatic builds are transparently available in a public repository. Core processing is based on a BioPerl program contributed to the Bio-EUtilities distribution and publicly available since 2013. Dependencies on raw data sources, alignment algorithms, or display output libraries can also be upgraded since the processing is automatic and script-based.

One potential challenge for researchers is the referencing of dynamic data within such a manuscript. This can be simply addressed by users citing the publication as a traditional static manuscript then stating the build date in Materials and Methods, as they would for a database.

Although the dynamic data will remain current, major revisions in understanding of a gene family or accumulation of small insights can in time render the explanatory static manuscript text obsolete. In this case the manuscript can simply be refreshed and republished independently in the same way as a traditional manuscript would be. The underlying scripts generating the dynamic data do not need to be rewritten, although new functions can be added to reflect new insights.

The core scripts underlying this manuscript have been written to facilitate generating equivalent catalogues for other organisms via simple build options. We have successfully trialled this for *Mus musculus* (Appendix B) to demonstrate that an equivalent dynamic manuscript for other histone gene sets only requires drafting appropriate surrounding static explanatory text. The approach is also applicable to larger and more diverse gene families such as our previous cataloguing of the Snf2 family (*Flaus et al.*, 2006).

3.6 Conclusion

Comprehensive tabulation and analysis of the canonical core histone gene family reveals significant diversity of isoforms, and challenges assumptions of bulk chromatin homogeneity. It has led to improvements in consistency of genome annotations and provides a reference for interpreting genomic and proteomic datasets. The differences between canonical core histone protein isoforms uncovers novel questions about their functional significance that suggest future experimental investigations. As a dynamic manuscript, this catalogue can automatically generate an up-to-date overview of the human canonical core histone gene family. It also demonstrates the potential for integrating reproducible research approaches into the scientific literature.

3.7 Materials and methods

The primary manuscript is generated from LaTeX sources, derived from dynamic data, and built using SCons (*Knight*, 2005). Search and download of fresh data is performed by bp_genbank_ref_extractor which was implemented for the Bioperl project (*Stajich et al.*, 2002) and has been included in the Bio-EUtilities module. Analysis of the data relies heavily on BioPerl.

All sources are freely available in a git repository at `https://github.com/af-lab/histone-catalogue.git`, including all sources for figures, manuscript templates, and build system making public all parameters used for processing.

This build of the manuscript was generated using BioPerl 1.006924 and Bio-EUtilities 1.75. Sequence and annotation data was obtained from NCBI RefSeq (*O'Leary et al.*, 2016) on 13th March 2017. Sequence alignments were generated by T-Coffee 11.00.8 (*Notredame et al.*, 2000). Sequence logos were generated using WebLogo version 3.5.0 (*Crooks et al.*, 2004). Description of sequence variants are represented following HGVS recommended nomenclature (*den Dunnen et al.*, 2016).

Chapter 4

Application of FRAP to Histones in Human Cell Nuclei

Abstract. Nucleosomes enable the stable compaction of almost all eukaryotic genomes but also require dynamic properties to enable access to the packaged DNA sequences. The stability of core histones within the nucleosome should be reflected in their capability for dynamic exchange by Fluorescence Recovery After Photobleaching (FRAP). To assay the effect of histone SWI/SNF INdependence (SIN) mutants known to destabilise nucleosomes *in vitro* and in *S. cerevisiae*, we sought to apply FRAP to chromatin in mammalian cell lines. This uncovered a number of challenges resulting from cell motility, nuclear movement within the cell, and chromatin motion within the nucleus with the long time frames required for FRAP of histones. We were able to compensate for the former difficulties by a combination of cell biological and computational techniques, but we were unable to establish an appropriate approach to compensate for motion of the immobile binding sites required for standard FRAP analysis. Visualisation using photoactivated tagged histones demonstrated the extent of this chromatin motion and a further complexity arising from complex non-homogenous channelling of tagged histones during diffusion. This reveals the limitations of FRAP over extremely long time scales, and suggests that this technique is unsuitable for quantitative measurement of histone dynamics in the mammalian nucleus.

4.1 Introduction

The chromatin packaging of eukaryote genomes compacts very large lengths of DNA into the microscopic cell nucleus, facilitates chromosomal movements during cell division, and provides a substrate for molecular mechanisms acting on the genome.

The building block of eukaryotic chromatin is the nucleosome structure, comprising 147 bp of DNA wrapped around an octamer of two copies each of core histones H2A, H2B, H3, and H4 (*Luger et al.*, 1997). In this structure, two H2A/H2B dimers flank a central H3/H4 tetramer. Nucleosome core particles are arranged in a linear chain separated by DNA linkers, and can be further compacted into higher order chromatin structures.

Chromatin functionality at the molecular and cellular levels requires the capability for dynamic rearrangement. Chromatin structure can be modulated through nucleosomes by changing the arrangement of histones and DNA in a process known as remodelling (*Flaus and Owen-Hughes*, 2011), or by altering histone chemical composition through post-translational modification (*Bannister and Kouzarides*, 2011), or exchange of histone variants (*Talbert and Henikoff*, 2010).

A large amount of information has been accumulated about the static structure of the nucleosome at atomic resolution (*McGinty and Tan*, 2014), and about the arrangement of polymeric chromatin (*Kuznetsova and Sheval*, 2016). However, the mechanisms for dynamic rearrangement of chromatin are much less well understood or integrated between the molecular and polymer levels (*Andrews and Luger*, 2011).

Nucleosome dynamics and histone SIN mutants

The archetype of ATP-dependent nucleosome remodelling enzymes is the SWI/SNF complex, which was identified in screens for mating type SWItching (*Stern et al.*, 1984) and Sucrose Non Fermentation (*Carlson et al.*, 1981; *Neigeborn and Carlson*, 1984) in *Saccharomyces cerevisiae*.

Mutations were subsequently identified that compensate for the loss of the SWI/SNF complex and these are collectively known as SIN mutations because they provide SWI/SNF INdependence (*Kruger et al.*, 1995). A subset of SIN mutants are single amino-acid changes in core histones

H3 and H4, providing a direct link between SWI/SNF and chromatin. This also suggests the mutated residues in the histone proteins influence the same pathways for nucleosome dynamics that are leveraged by the remodelling enzyme, and that these residues are significant for nucleosome stability.

The prediction that the stability of SIN mutant containing nucleosomes is affected in chromatin has been tested *in vitro*, where it was observed that SIN mutant nucleosomes display higher thermally driven nucleosome sliding mobility (*Flaus et al.*, 2004) and that the mutated residues affect histone-DNA contacts in crystal structures (*Muthurajan et al.*, 2004).

However, the effect of histone protein SIN mutants in the more complex *in vivo* chromatin environment of mammalian cells has not been demonstrated. Validating the functional significance of these residues is important for understanding nucleosome dynamics and explaining the high degree of conservation of histone protein sequences in eukaryotes.

Fluorescence Recovery After Photobleaching

Fluorescence Recovery After Photobleaching (FRAP) is an optical technique that can be used to report the dynamics of fluorescently tagged molecules within live cells. Tagged molecules inside a small region are irreversibly photobleached by a focused high power laser beam and the recovery rate of fluorescence in the bleached area is measured. The recovery rate is interpreted as unbleached fluorescing molecules from outside the region at the time of photobleaching diffusing into the bleached area. It is assumed that this fluorescence recovery reflects natural protein movement.

A simple chemical equilibrium underlies the model for FRAP for a molecule with a single binding reaction:

$$F + S \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} FS$$

where F represents freely diffusing proteins, S represents immobile vacant binding sites, and FS is the complex between the two when the proteins are bound to the sites. The value of k_{on} and k_{off}, are estimated from the rate at which photobleached F is replaced in the FS complex. However, complexity is added to this simple model if diffusion and space are considered.

Ongoing development of FRAP has led to increasingly complex models with more precision and accuracy than the inverse of an exponential decay. For example, $1 - Ae^{-k_{off}t}$ where t is time, k_{off} is the dissocation constant, and A is the mobile fraction which can be used to estimate the association constant k_{on} (*Mueller et al.*, 2010). Despite the sophistication of current models, FRAP requires assumptions that are difficult to maintain over long experimental observation times. Firstly, equilibrium must be maintained throughout the entire experiment so that both k_{on} and k_{off} remain constant. This also requires that concentrations of both F and S remain constant. Secondly, distribution of the fluorescently tagged molecules must mimic the endogenous protein. And finally, the binding sites must be part of a large, relatively immobile complex on the time and length scale of the recovery. In addition, different FRAP models will have further constraints based on the assumptions involved in its design (*Mueller et al.*, 2010).

FRAP measurements of histones

FRAP has been extensively used to obtain qualitative and quantitative insight into the kinetic properties of chromatin bound proteins (*Phair and Misteli*, 2000; *Essers et al.*, 2005a; *Agresti et al.*, 2005). These rely on the established assumption that chromatin is relatively immobile in the interphase nucleus (*Abney et al.*, 1997) since most proteins show recovery on the scale of seconds to minutes. H2B–GFP (*Kanda et al.*, 1998) has become the standard reference for the immobile fraction in these FRAP experiments (*Dey et al.*, 2000; *Kuipers et al.*, 2011; *Jullien et al.*, 2016).

However, the dynamics of the core histones themselves was measured by FRAP in a seminal and widely cited study by *Kimura and Cook* (2001). Multiple H2B populations were delineated with distinctive exchange rates. Some 3 % of H2B had a rapid recovery within minutes, 40 % had slow recovery with $t_{1/2}$ of 130 min, and over 50 % of H2B molecules had a very slow recovery with $t_{1/2}$ of over 8 hours that was considered to be effectively immobile.

In contrast to H2B which is a histone dimer component, the tetramer histones H3 and H4 were found to have even slower mobility. There were no rapid populations, with only slow and very slow populations of 16 % to 22 % and 62 % to 68 % being identified respectively.

In combination with additional heterokaryon data, Kimura and Cook interpreted the rapid, slow, and very slow by exchanging H2B populations as correlating with transcription units, euchromatin, and heterochromatin respectively. They assigned over 80 % of H3 and H4 as immobile, whereas the remaining ≈20 % was suggested to be mobilised by remodelling. This latter small but significant slowly exchanging fraction of histones provides an opportunity to observe the dynamics of tetramer histones such as SIN point mutants in live mammalian cells.

Aims and objectives

We aimed to develop a technique capable of estimate the *in vivo* effect of histone sequences on nucleosome dynamics and validate previous *in vitro* studies. We choose FRAP due to its previous use on comparing exchange between different histone types (*Kimura and Cook*, 2001). As a positive control, we choose the H4 mutant R45H which has exhibited the highest increase in nucleosome mobility *in vitro* studies (*Flaus et al.*, 2004).

In attempting to achieve quantitative measurements we encountered multiple technical challenges associated with measuring subtle kinetic alterations in nucleosome dynamics over long time periods in live cells. This required us to define the limitations of FRAP for observing molecules such as core histones with extremely slow exchange rates.

4.2 Materials and Methods

Stable HeLa cell lines were created for H2B–EGFP, H2B–EGFP D25G V118I, H3-EYFP, H3–EYFP T45A, H3–EYFP T45E, H4–EYFP, and H4–EYFP R45H (§2.2). Cells were transformed by lipofection (§2.4), split to a low confluence, and treated with 3 µg ml^{-1} Blasticidin-S for one week. Colonies were screened by fluorescence microscopy and positive individual colonies were selected for growth and by Fluorescence Activated Cell Sorting (FACS) to generate homogeneous highly fluorescing cell lines. Transiently expressing cell lines, both HeLa and horse fibroblasts, were transformed by lipofection in the same manner, and imaging performed 48 hours later. Primary horse fibrolasts were a gift from Prof. Elena Giulotto (University of Pavia). HeLa cells were ATCC line CCL-2.

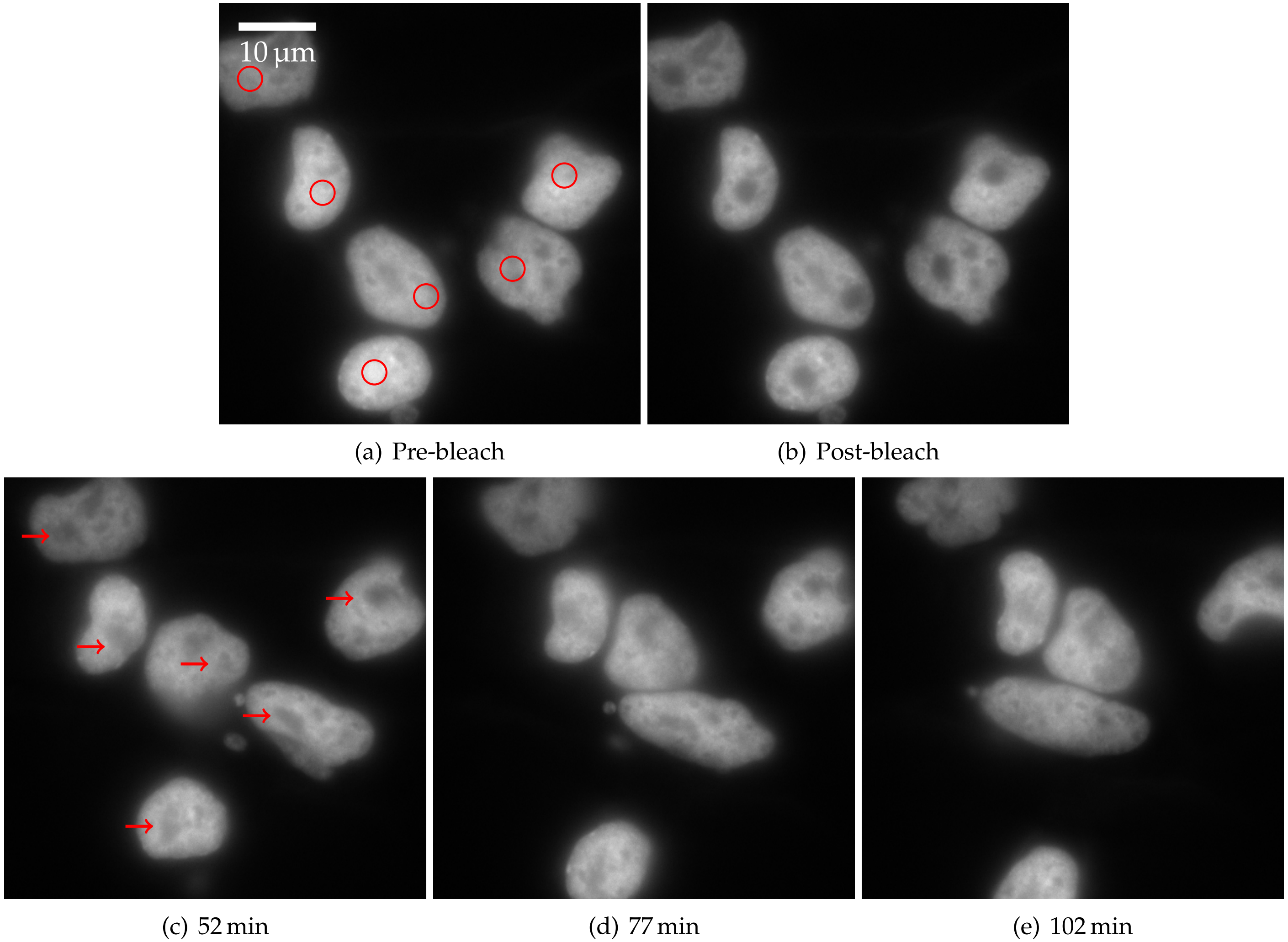
Confocal microscopy was performed on a Zeiss LSM510 Meta microscope using glass bottom LabTek II chambers. Wide-field fluorescence microscopy was performed with an Applied Precision DeltaVision Core system using 35 mm glass bottom MatTek dishes. In both cases, imaging was performed within an acrylic environmental chamber at a temperature of 37 °C and 5 % CO_2. Images were acquired and deconvolved on the DeltaVision system.

Cell movement between time frames was calculated by consecutive template-based registration using normalised cross-correlation. The CropReg script developed for this purpose is available as free open source software in this manuscript repository. Nuclei of interest were identified on the first frame and used as templates on subsequent images. To correct for rotational movement around the z dimension of the optical axis system, registered frames were aligned by rigid body geometric transformation using the ImageJ (*Schneider et al.*, 2012) plugin StackReg (*Thévenaz et al.*, 1998).

Automatic extraction and processing of FRAP recovery curves was performed with the GNU Octave programming language (*Eaton et al.*, 2016) and the Octave Forge Image package. Source code written in Matlab for a previously reported circle FRAP model (*Mueller et al.*, 2008) was kindly gifted by the original authors and ported to GNU Octave. We developed `frapinator`, a new program written in GNU Octave to automate analysis with multiple command line options and released it in the FRAP Octave package as free open source software. `frapinator` includes all individual functions for image pre-processing and FRAP fitting (§5.4).

4.3 Results

To investigate the challenges of performing FRAP in mammalian cells over the several hours required to measure histone mobility, we transfected HeLa cells with H2B-EGFP under the control of an EF-1α promoter and observed recovery in cells for 3.5 hours (Figure 4.1).



(a) Pre-bleach

(b) Post-bleach

(c) 52 min

(d) 77 min

(e) 102 min

Figure 4.1: **Circle FRAP of H2B–EGFP in HeLa.** FRAP experiment performed in a widefield microscope with HeLa stable cell line expressing H2B–EGFP. (a) Last of the acquired images before the bleach event. A total of 20 images with a time interval of 1.7 s were acquired before bleaching. Circles show the location for the bleaching events. (b) First image post the bleach event. (c)–(e) Selected frames from FRAP experiment. Arrows show location of the bleach spot. Total time of FRAP experiment was 3.5 h with an increasing time interval between images, from 15 s to 2.5 min.

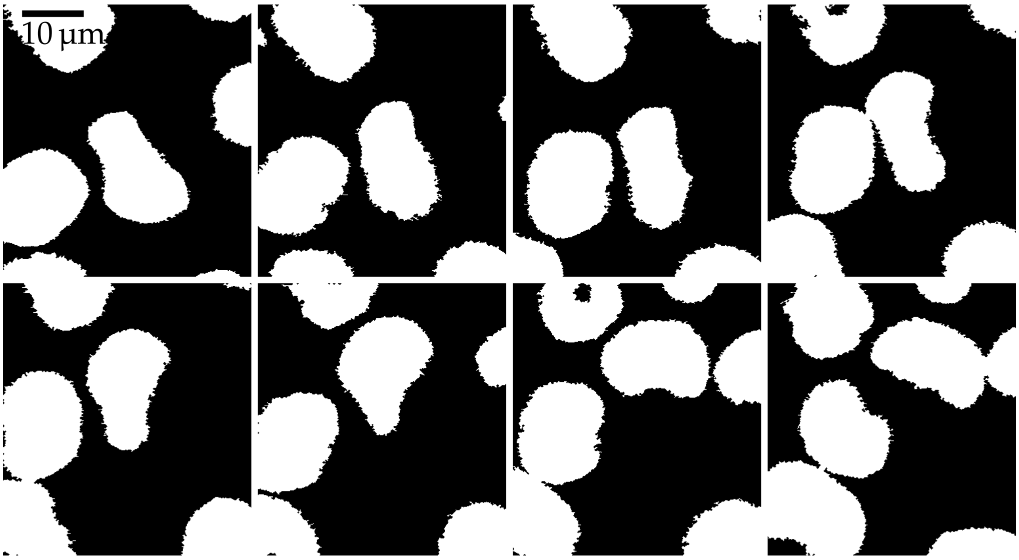


Figure 4.2: **Movement of confluent HeLa cells during a FRAP experiment.** HeLa-derived stable cell line expressing H4 R45H–YFP were observed in a confocal microscope over 8 hours with 1 min interval. Figure shows only the nuclei outline in intervals of 50 min.

The images taken at 25 min intervals revealed considerable changes in position of the nuclei, arrangement of features within each nucleus, and shape of the bleach spot (Figure 4.1). The central requirement of FRAP is to accurately identify and quantitate the signal in the photobleached spot over time. This led us to pursue both cell biological approaches to minimise motility and computational approaches to track imaged regions.

Inhibition of cell motility

We first attempted to reduce motility by restricting the space available using cells at higher confluency for FRAP experiments. This was performed using a HeLa cell line stably expressing H4 R45H to ensure even tagging of histone fluorescence in all cells. This resulted in some decrease in movement of cells but did not achieve complete immobilisation (Figure 4.2). In fact, nuclei frequently underwent considerable reshaping as cells apparently squeezed between their neighbours.

Fibroblast and epithelial cells display the property of contact inhibition of locomotion (*Abercrombie*, 1970). In this cellular growth response, cells attempt to move in an opposite direction after contact with another.

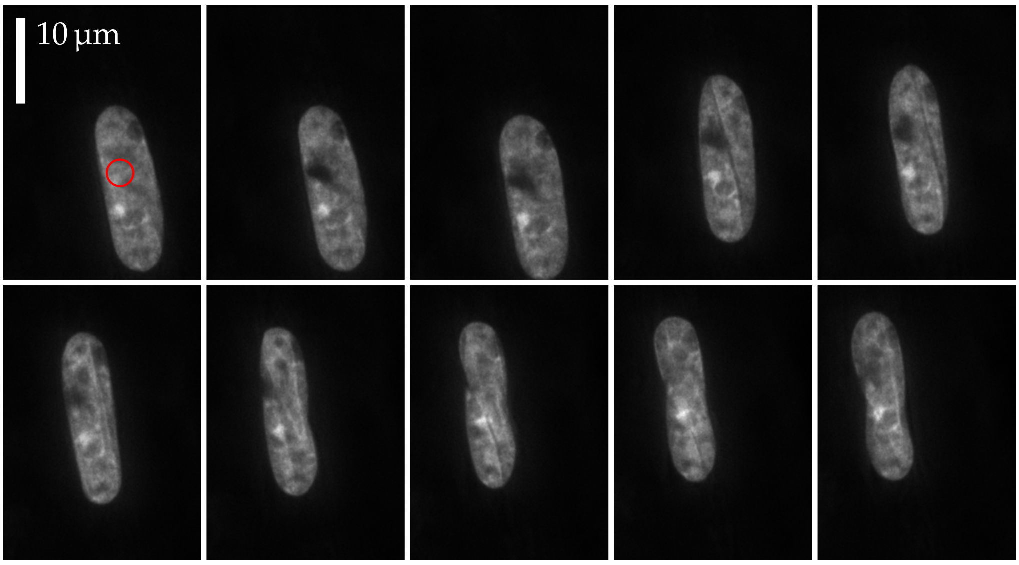


Figure 4.3: **Movement of confluent primary horse cells during a circle FRAP experiment.** Primary horse fibroblast transiently expressing H2B-EGFP imaged in a widefield microscope. Red circle on the first image shows the location of the bleaching event. Following images shows the nuclei movement at 15 min intervals for a total of 2 hours.

As the number of cells increases and they become surrounded by neighbours, the available directions are reduced. Like most cancer cell lines, HeLa cells have lost the ability to activate contact inhibition (*Stephenson*, 1982), so we obtained a primary horse fibroblast cell line. Because primary cell line transfections typically have much lower efficiency than cancer cell lines, we transfected cells at 70 % confluence and performed FRAP after 3 days. This timing enabled us to perform the transfection while cells were actively dividing, which increases the efficiency, and to perform the imaging once they reached confluence for reduced motility.

Despite reaching confluence where contact inhibition of the fibroblasts was expected, we continued to observe movement of the transfected horse cells (Figure 4.3). The characteristics of cell movement also differed dramatically from HeLa cells. The horse cell nuclei exhibit a helical motion about the vector of their movement in x and y axes (Figure 4.3), whereas HeLa nuclear motion was mostly restricted to the z axis relative to the dish (Figure 4.1).

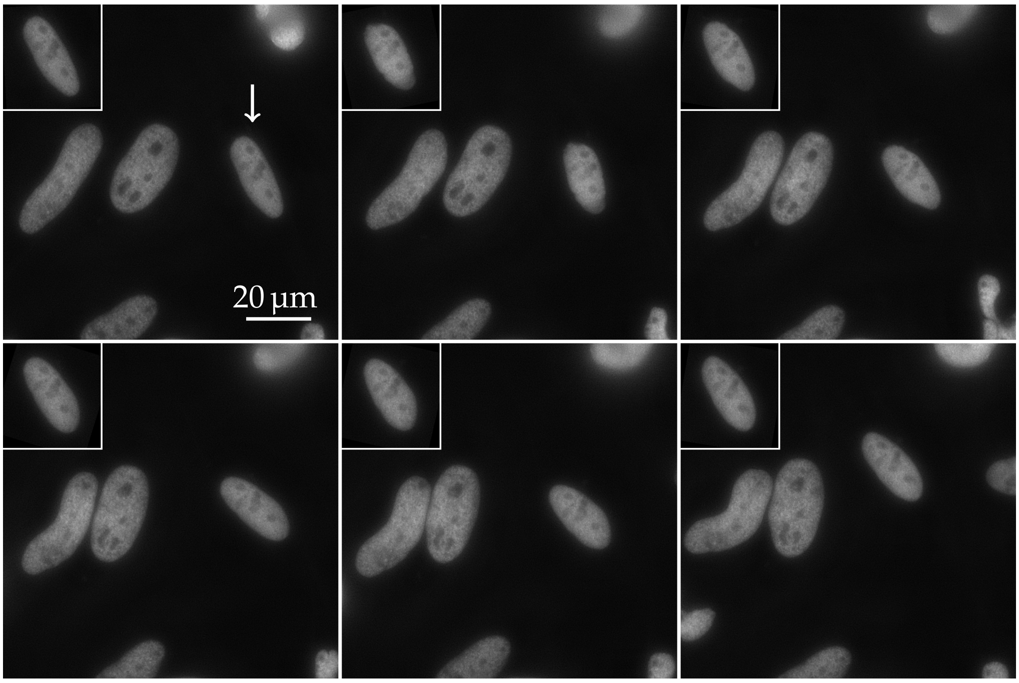


Figure 4.4: **Automatic tracking and alignment of moving cells.** HeLa-derived stable cell line expressing H3–EYFP observed in a widefield microscope. Imaging was performed for a total of 6 hours with 10 min imaging interval. White arrow on the first panel corresponds to the cell that is being automatically tracked and aligned, and is inset on the top left corner of each panel.

Image-based tracking of cell movement

As an alternative strategy we implemented CropReg, a script to automate cell tracking of time series sequences. Using this image processing approach we were able to track individual HeLa cell nuclei throughout an entire sequence of FRAP images provided that nuclei did not overlap (Figure 4.4). Although only a minority of cell image sequences satisfied this requirement throughout the full 8 hour duration of observations, it was possible to collect a sufficient number of cell observations for FRAP calculations.

Chromatin movement within nuclei

While performing the FRAP experiments, we observed movement of fluorescent chromatin features within cell nuclei that was supplementary to the overall motion of the cell itself (Figure 4.5). This could not be

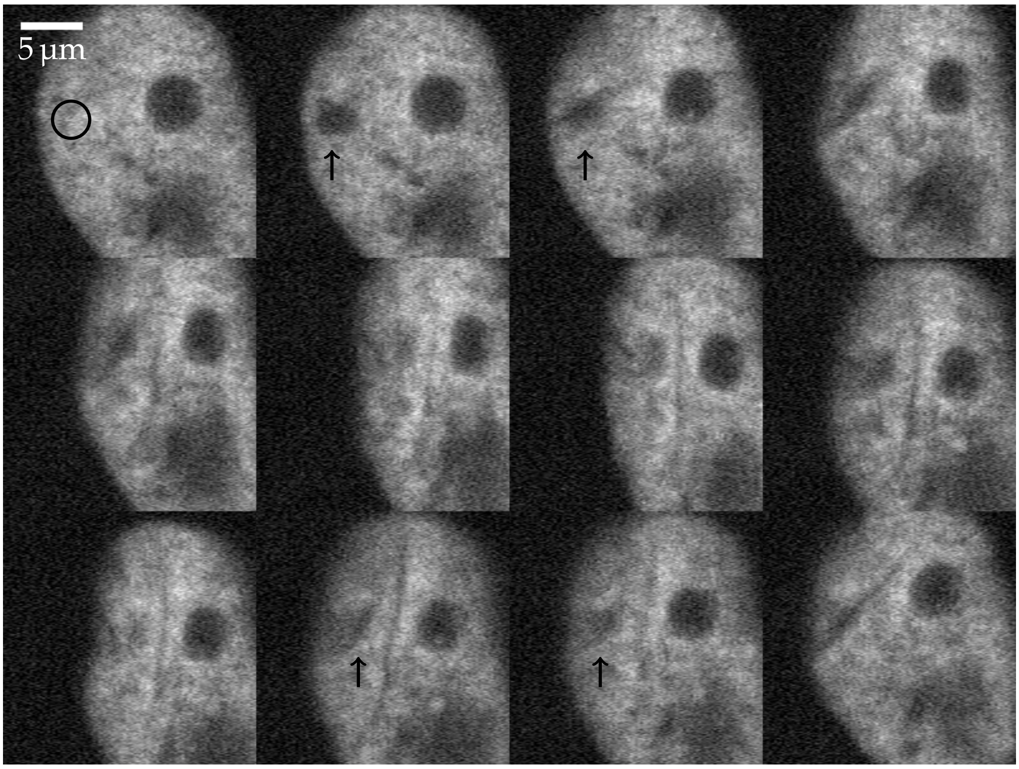


Figure 4.5: **Movement of nuclear features and bleach spot distortion.** FRAP experiment performed in a confocal microscope with HeLa stable cell line expressing H4–YFP. Circle FRAP was performed after 15 pre-bleach images and images acquired with 60 s interval for a total of 4 hours. The first three frames show the bleaching event, followed by the initial 3 hours of recovery with 20 min interval.

accounted for by simple rotational movement of nuclei as rigid bodies around the x or y axis, and instead appeared to involve movement of individual regions within nuclei. Bleach spots also frequently showed elliptical or more complex distortions indicative of structural movements in the chromatin (Figure 4.5).

Selection of G_1 cells

One possible cause for changes in chromatin features that we observed is DNA replication and chromatin repackaging during S phase. Furthermore, the doubling of histone content as a result of S phase breaks a core assumption of FRAP that the system remains in equilibrium throughout the duration of the experiment. Therefore, the requirement to measure for a time period of 8 hours within a single cell cycle phase limits observations to G_1.

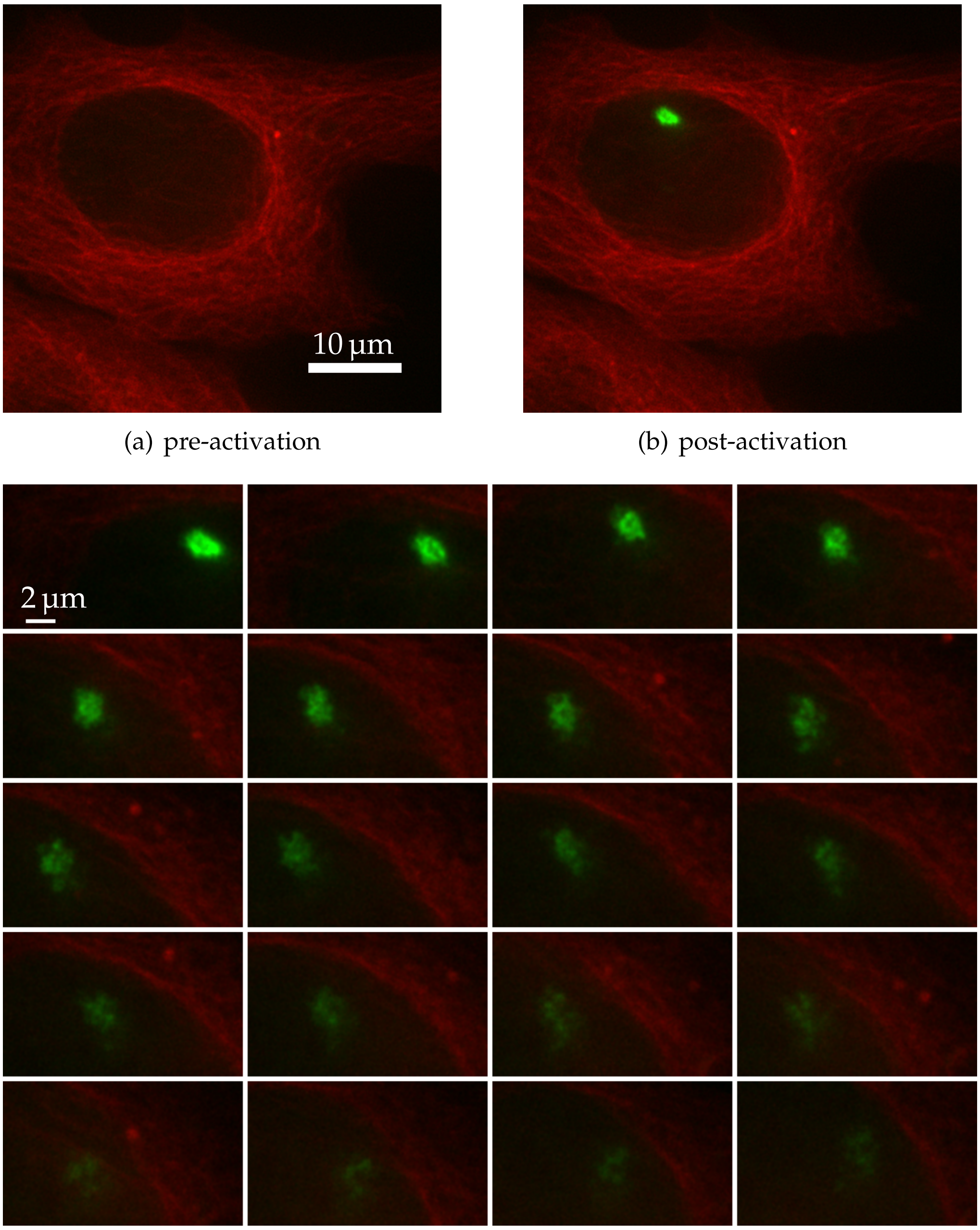
To identify daughter cells that could be confidently assigned to early G_1 because they had sufficient time to complete post-mitotic chromatin unpacking, cells in mitosis were selected and manually tracked for 4 hours. This selection was challenging because HeLa mitotic cells round up as spheres with only a weak connection to the growth surface causing them to exit the field of vision. Low laser power and a 30 min observation interval was used to minimise fluorophore damage. Since our system did not permit simultaneous Z-stack and time lapse imaging, and because cells in mitosis are in a separate focal plane, imaging was performed with a maximal pinhole sized focused between the growing and mitotic cell planes. The resulting blurred images were sufficient to visualise cells during the entire period required for selection. However, even after carefully selecting cells early G_1, structural movements within the bleached region could still be observed (data not shown).

Chromatin movement observed by photoactivation

Although we had surmounted the technical challenges of collecting overlaid images of nuclei for long time periods in G_1 cells containing stably expressing core histones tagged with fluorescent reporters, we were concerned about the non-homogeneity of chromatin behaviour.

To assess the extent of the chromatin motion, we performed photoactivation to track the movement of chromatin in the activated region alone. For this purpose we fused H2B to photoactivatable GFP as H2B–PAGFP. Since PAGFP cannot be easily detected before photoactivation, cells were co-transfected with mCherry–α–tubulin which localises exclusively to the cytoplasm and provides an outline of the nuclear region (Figure 4.6(a)).

Considerable non-homogenous movement of chromatin was clearly evident after activating and following H2B–PAGFP in G_1 cells (Figure 4.6). Instead of homogeneous diffusion of fluorescence, activated spots uncurled over time with individual channels of localised PAGFP appearing in the nuclei (Figure 4.6). This finding suggests that quantitative FRAP is not tractable using a simple FRAP model based on homogenous non-directional diffusion.



(a) pre-activation

(b) post-activation

(c) activated spot over time

Figure 4.6: **Photoactivation experiment demonstrating complex chromatin movement.** HeLa cells co-transfected with H2B–PAGFP and mCherry–α–tubulin. (a) Cell shown before photoactivation of H2B–PAGFP in a widefield microscope, demonstrating mCherry-bounded nuclear region. (b) Same cell shown immediately after photoactivation showning circle photoactivated H2B–PAGFP nuclear spot. (c) Image sequence at 20 min intervals after photoactivation showing complex channelling of H2B–PAGFP diffusion.

4.4 Discussion

We wished to quantitatively determine the effect on human chromatin dynamics of SIN mutations in core histones H3 and H4 known to be destabilising *in vitro* and to affect cell growth in *S. cerevisiae*. We set out to use a previously reported circle FRAP model which accounts for multiple factors in a typical FRAP modelling (*Mueller et al.*, 2008).

However, FRAP recovery is incomplete even after 8 hours for core histones (*Kimura and Cook*, 2001). This led us to address a series of technical challenges in collecting valid quantitative recovery data over extended time periods.

The first problem faced was cell motility, which is an expected property of actively dividing cells. We attempted to reduce motility by taking advantage of the fact that many primary cells display contact inhibition of locomotion and proliferation when they reach high densities. This contact inhibition is a natural mechanism that controls cellular growth in multicellular organisms, and results in a stop in proliferation with the formation of a monolayer of healthy cells in tissue culture.

However, the approach has disadvantages including increased cell handling and reduced transfection efficiency. The potential inability to compare results with published data for immortalised cell lines such as HeLa is also undesirable.

Despite achieving a monolayer of healthy cells that could be maintained stably over 2 weeks, individual transfected primary horse fibroblasts still showed motility despite exhibiting overall characteristics of contact inhibition. Furthermore, nuclei in these cells displayed a helical motion on the direction of cell movement (Figure 4.3).

The possibility of chemically inhibiting cells to reduce motion was considered since previous FRAP experiments with core histones were performed using multiple inhibitors of protein synthesis (*Kimura and Cook*, 2001). However, these studies revealed inhibitor-dependent variations in kinetics and the authors qualified their conclusions about the absolute accuracy of the histone exchange parameters measured.

To better address the problem of cell motility we instead developed a computational approach by writing the program CropReg for cell tracking by normalised cross-correlation template matching. Using automated analysis enabled us to process the large numbers of cell images required to provide statistically valid quantitative measurements of core histone exchange.

The second challenge to measuring core histone exchange by FRAP is that a chemical equilibrium is required between freely diffusing proteins and formation of a complex. Although absolute equilibrium is unlikely in the dynamic cell environment undergoing complex transcriptional and translation responses anyway, DNA replication involving polymerase passage and repackaging of the duplicated genome in S phase will certainly unbalance any equilibrium.

Chromosome compaction in mitosis also generates a chromatin environment that is distinct from interphase. This limits FRAP experiment to either G_1 or G_2 phases. The HeLa cell cycle has a typical G_1 phase of 11.7 hours and a G_2 phase of 3 hours (*Bravo and Celis*, 1980) so the extended time periods needed for FRAP of core histones requires starting FRAP early in G_1 (Figure 4.7).

Post-mitotic chromosomes take approximately 2 hours to migrate within the nucleus and rebuild the interphase nuclear architecture during early G_1 (*Belmont and Bruce*, 1994; *Thomson et al.*, 2004; *Essers et al.*, 2005b). This defines the window for extended FRAP experiments from approximately 3 to 11 hours after mitosis in HeLa cells, although cells lines with even longer G_1 phase could also be used (*Sipos et al.*, 2003).

We wished to avoid the use of drugs for cell cycle arrest since this has been shown to influence FRAP results (*Kimura and Cook*, 2001). We also discounted serum starvation to move cells into the quiescent G_0 phase since this could affect the relevance of measuring core histone exchange (*Pirkmajer and Chibalin*, 2011).

Instead, we developed a procedure to track progression of cells manually during mitosis where visual identification of the cell cycle is possible. This allowed us to minimise the variations of normal cell growth and to identify individual cells exactly 3 hours after start of G_1. This has the added advantage of allowing time for maturation of GFP expressed during the establishment of interphase. The time interval between images during manual selection was increased and both resolution and laser intensity were reduced to minimise phototoxicity or bleaching. This resulted in a set of selected early G_1 cells suitable for FRAP experiments.

The fluorescently tagged histone proteins are constitutively expressed under the control of an EF-1α promoter, so they lack the 3′ reg-

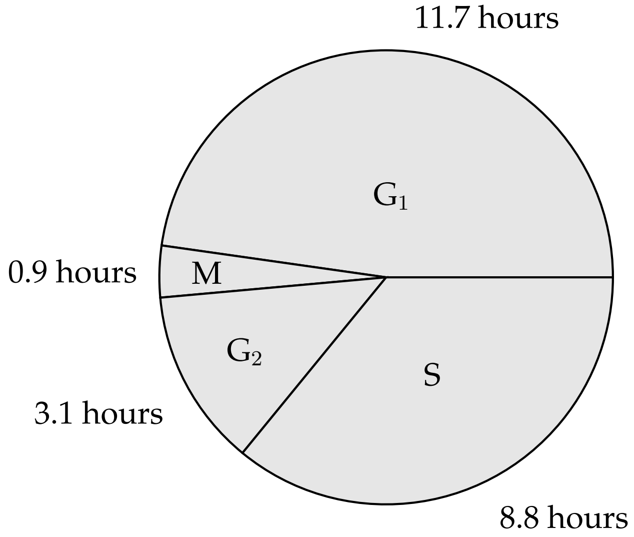


Figure 4.7: **HeLa cell cycle phases and timing.** Under optimal growth conditions the HeLa cell has a median doubling time of 24 hours, with G_1 and S phases of 11.7 and 8.8 hours respectively (*Bravo and Celis*, 1980).

ulatory features of native histone genes. This regulation does not follow the normal expression program of a histone gene and could affect the distribution of the histone in chromatin. Constant expression of tagged histones by a strong constitutive promoter will enrich them in the G_1 and early S phase pools making subsequent incorporation in euchromatin more likely, relative to mid-late S phase where heterochromatic sequences are replicated and packaged (*Rhind and Gilbert*, 2013).

A more realistic tagged histone expression profile could be achieved using flanking regulatory regions from native histone genes, as demonstrated for H3 and CENP–A (*Taylor et al.*, 1986; *Shelby et al.*, 1997). Another potential solution is to insert GFP in-frame into the native gene locus by genome engineering, although the redundancy between the multiple canonical histone genes means that identifying the most appropriate isoform to target could introduce complexities.

Protein synthesis inhibitors were used by *Kimura and Cook* (2001) to address this issue, but this has the disadvantage of potentially affecting many other processes as discussed above.

The final challenge to measuring core histone exchange by FRAP that we identified was non-homogenous regional movement of chromatin itself.

One possible cause for this movement is chromatin repackaging during DNA replication which we addressed by selecting cells at early G_1 phase. Another possible cause is chromatin remodelling as part of a DNA damage response caused by the FRAP photobleaching event itself. The phototoxicity effects of a FRAP experiment are often dismissed on the basis that the photobleaching event of a typical FRAP experiment does not affect cell viability (*Kruhlak et al.*, 2000; *Kimura and Cook*, 2001; *Carrero et al.*, 2003) but the DNA damage that such an experiment could introduce, and the effect that the repair response of such damage may have on chromatin reconfiguration, has not yet been addressed. Lasers of both shorter and longer wavelengths, and longer exposure times, are often used to introduce single and double DNA strand breaks in live cells (*Stixová et al.*, 2014; *Mari et al.*, 2006; *Kim et al.*, 2002) and one recent study has demonstrated activation of base excision and single strand break repair pathways using a 488 nm laser in conditions similar to FRAP (*Muster et al.*, 2017).

Independently of the cause for the chromatin movement, our observations undermine the assumption of FRAP analysis that binding sites remain immobile throughout the FRAP experiment. This assumption is required to interpret recovery as the rate of movement of freely diffusing unbleached molecules into the bleached area which allows the kinetic rates k_{on} and k_{off} to be estimated. If chromatin binding sites also move then the recovery curve becomes a much more complex function of both binding site movement and free diffusion.

Chromatin movement is recognisable by changes in the intra-nuclear features of the fluorescent chromatin and by changes in the circular bleach spot. Although some of these effects are subtle when observed by photobleaching, the photoactivation of an equal circular spot demonstrates clear non-homogenous reshaping of chromatin. Equivalent chromatin movement has also been reported for H4–PAGFP in strip photoactivation *Wiesmeijer et al.* (2008).

The movements we observed were in the range of 4 µm, which is double the size of the bleach spot, and exhibited complicated shapes reminiscent of channelling. This is consistent with chromosome distribution in nuclei that is territorial on the scale of 5 µm (*Sun et al.*, 2000) separated by interchromosomal channels of 10 nm to 100 nm (*Görisch et al.*, 2005).

The clarity of H2B–GFP imaging by photoactivation suggests the opportunity to analyse the paths taken by diffusing core histones. For example, simultaneous use of combined photoactivation and photobleaching of complementary dimer and tetramer histones could enable relative diffusion rates and paths to be determined. Alternatively, an enzymatic mechanism to incorporate a complementary photo-differented label into DNA would facilitate masking for the original location at the same time as tracking the histone diffusion and enable quantitative FRAP. Nevertheless, it is important to recognise that such experiments would test the resolution and sensitivity of microscopes.

4.5 Conclusion

Since its inception over 30 years ago, FRAP has been continuously improved through technical capabilities of light microscopy and sensitive kinetic models that are now able to take into account an increasing number of biophysical features such as container size, non-homogeneous distribution of fluorescence, and profile of bleach spot.

Despite these advances, the ability to perform FRAP over extended time periods of several hours for highly stable complexes such as core histones is limited by the dynamic nature of the cell.

We overcame the challenges of cell motility and selection of cells in G_1 phase, but were not able to develop a method to adjust for changes in chromatin structure within the cell nucleus. While a photobleached spot appears stable and can be tracked over several hours, small natural disturbances and non-homogeneous diffusion impact on photorecovery and estimation of kinetic parameters. We find that FRAP is suitable for semi-quantitative estimates of slowly diffusing molecules but not for the precise quantitative comparisons required to compare core histone mutations.

Ultimately, the issue of long observation times stems from the requirements of FRAP models to achieve full recovery of the mobile populations being measured.

Single particle tracking is a microscopy technique where, as the name suggests, the motion of isolated molecules is observed. In this technique, a small number of particles, small enough that they can be resolved and their individual movements tracked, is observed over a short amount of time, typically on the scale of seconds. The short time is a limitation and not a requirement of the technique, and is caused by observational photobleaching. This would remove the requirement of hours long observation and overcome the issue of chromatin movement. In addition, it would provide an overview of different types of protein motion where FRAP would only provide an average of the movement, and as a super resolution microscopy technique, would also be able to identify confined movement that FRAP would otherwise classify as immobile as been previously the case of MHC class I proteins (*Smith et al.*, 1999). However, the high concentration of histones in the nucleus makes it challenging to observe the required individual molecules. This could possibly be overcame with the use of photo-activatable FPs such as PAGPF which would allow for the activation of a small population and the use of a Selective Plane Illumination Microscopy (SPIM) which allows the observation and excitation of fluorescent molecules in a single focal plane.

Fluorescence Correlation Spectroscopy (FCS) is a fluorescent microscopy technique providing estimates of dynamic parameters by using fluorescence fluctuations in a femtolitre volume. More spatiotemporal dynamics can be obtained by observing the fluorescence fluctuations while moving the measurement volume across the sample, a technique named Spatio-Temporal Image Correlation Spectroscopy (STICS) (*Hebert et al.*, 2005). One other variation of FCS is Raster Scan Image Correlation Spectroscopy (RICS) which is similar to STICS but can be performed on a standard confocal microscope (*Digman et al.*, 2005). In both cases, image regions that may span the entire cell nuclei can be observed and populations with different dynamics localised within it. Such techniques could potentially provide not only a method to compare core histone mutations *in vivo* but also their effect in different chromatin domains. However, the implementation of STICS and RICS is likely to be challenging due to the optimisation of scanning parameters and complex data processing. There are relatively few reports of the use of these approaches outside the laboratory that developed them.

Overall, the current period of rapid technological advances in cell biology research means that new techniques such as RICS and single particle tracking offer hope to address specific challenges such as histone mobility for which existing approaches such as FRAP are poorly suited.

CHAPTER 5

Software Tools for Image and Sequence Analysis

Abstract. The ability to interrogate data, to reproduce methods, and to improve techniques are cornerstones of scientific research. Free software is an important resource in the increasingly data driven quantitative revolution in contemporary molecular cell biology research. We wished to implement a fully transparent and reproducible vision of a build process referred to as "reproducible research" that directly links primary data resources such as sequence databases and microscopy images with the standard expression of scientific research insights as manuscripts and theses. To achieve this I contributed to a number of free software projects including the Octave programming language, the BioPerl suite of scientific data tools, and the Debian GNU/Linux operating system. This included implementing new algorithms, refactoring code for efficiency and consistency, creating maintenance support tools, and packaging software for ease of installation by non-expert users. The software developed and contributed to public repositories should enable other researchers to reproduce our work, and to implement a diverse variety of their own solutions in computational biology.

5.1 Introduction

Today's biology research is mainly digital. Data being generated is so complicated and extensive that analysis is done by software, with computer data acquisition and analysis as an integral part of a researcher's work. Such work is then published, but it is only useful to the scientific community if it can be reproduced by colleagues and further research can be performed to extend from it. Even if the software used is carefully described in the methods, modern programs are so complicated, and analysis is so dependent on so many variables, that there is limited hope for other scientists to recreate the implementation from brief descriptions. This makes access to the original software used in research a prerequesite for reproducibility.

Sharing the code is not enough. Programs that undertake the direct analysis tasks sit on top of several other components. Specific versions of libraries are used, on top of a specific programming language, running in a specific operating system, on hardware with a particular specification. Reproducing this runtime environment is not trivial.

Following from free software came the ideals of open and collaborative software development. The seminal essay "The Cathedral and the Bazaar" by *Raymond* (1999) compares the methods of free software development by an exclusive group, in a "Cathedral" where the source code is only made public for releases, with a method where development happens in the public domain, a "bazaar", where anyone can see and contribute to it. In free software, the source code for a program is available on the internet for anyone. The "bazaar" takes this further by having every single change open on the internet in real time, even before a new version of the software is formally released. This invites users to contribute through discussion about development in public forums. The ideal is that any user can become a developer, simply by sending their contributions to the appropriate public forum. Users are motivated to do this because they have a personal interest in enabling the program to support a specific feature that matters to them. And they are incentivised to contribute to the project because it is less work to participate in the collective than to write their own project just to have the feature they lack. Free software projects (*Schindelin et al.*, 2012; *Stajich et al.*, 2002) are now now almost all developed using the "bazaar" approach.

I faced many computational challenges during our research presented in Chapter 3 and Chapter 4. Using only free software I was able to contribute my work to several free software projects in collaborations that go beyond the boundaries of the typical research group. While the advances allowed me to perform the acquisition and analysis of our own data, I expect that these open contributions will also enable other researchers to advance their own research.

5.2 GNU Octave

The GNU Octave programming language was used extensively for the analysis of microscope images in Chapter 4. Several features of the language such as its handling of multi-dimensional arrays, interactive user interface, and extensive image processing functions, as well as its supportive community, make it an attractive choice for quantitative microscopy.

GNU Octave is a high level array programming language where operations are generalised to multidimensional arrays and Octave is primarily intended for numerical computations (*Eaton et al.*, 2016). All values are multidimensional arrays and this provides an abstraction layer that is useful when writing code for an arbitrary number of dimensions.

Octave also has a Read–Eval–Print Loop (REPL), or interactive top level user interface, similar to the IPython and Unix shells. This reduces the feedback time and provides an efficient environment for exploratory data analysis.

The project has an active community mainly composed of scientists and engineers who provide a large support group of specialists in numerical computations.

The separate Octave Forge project hosts a large number of packages that extend Octave into specific applications such as control systems (*Reichlin*, 2013), time-frequency analysis (*Průša et al.*, 2014), or level sets (*Kraft*, 2015). It provides a collaborative environment for development of Octave packages and another nexus of the Octave community.

Both GNU Octave and Octave Forge packages are free software which allows the study, modification, and distribution of modified source. I made extensive use of this feature and have contributed to improving Octave and its packages for the needs of quantitative microscopy.

Finally, Octave comes with basic support for image processing and the Octave Forge Image package extends this with a number of functions for image processing such as geometric transformations, mathematical morphology, image registration, and noise reduction.

While Octave is well suited to image processing, I identified several problems related to large image size or number of dimensions.

Our microscopy images were several megabytes in size. For example, the single cell FRAP experiments for histone dynamics generating TIFF files were 489 MiB in size. These images had a field of view of 300 by 300 pixels and 2500 time frames on acquisition by an 8-bit camera. Such files sizes are typical in the microscopy field.

Biological microscope images have a varying number of dimensions. In addition to the 2 dimensional plane of a standard image, microscope outputs typically include any combination of z dimension along the optical axis for a volume image; time dimension for time-lapse experiments; and wavelength for multi-channel experiments. Recent microscopy techniques generate images with additional dimensions such as angle, phase, and lifetime. This varying dimensionality adds complexity and makes it challenging to write generalised subroutines for microscope image analysis.

Reading and Writing of Image Files

Octave uses the GraphicsMagick C++ library for the reading and writing of image files which provides a common interface to a varied collection of image format specific libraries covering almost 90 major image formats. In Octave, this functionality is provided via the functions `imfinfo`, `imread`, and `imwrite`.

I rewrote these three functions with the aim of achieving reduced memory usage, increased performance, an improved interface for multidimensional images, and new image types. For this purpose, new options were added to read and write a series of image planes in a single function call, to read specific regions of interest in individual planes, to read and write images with floating point precision, and to append additional planes to existing image files.

Table 5.1: **New or improved functions in GNU Octave.**

`bitcmp`
`bzip2`
`cubehelix`
`fliplr`
`flipud`
`flip`
`frame2im`
`gallery`
`gray2ind`
`gzip`
`hsv2rgb`
`im2double`
`im2frame`
`imfinfo`
`imformats`
`imread`
`imwrite`
`ind2gray`
`ind2rgb`
`inputParser`
`ntsc2rgb`
`pkg`
`psi`
`rectint`
`rgb2hsv`
`rgb2ind`
`rgb2ntsc`
`rot90`
`validateattributes`

As part of this rewrite, new features were introduced such as support for the CMYK colour model, EXIF and GPS metadata, reading and writing of animations in GIF images, and control over the image compression type, while existing features were improved such as support for transparency and indexed images.

To support this development, I created a system that allows changing how the `imread`, `imwrite`, and `imfinfo` functions behave for each file format. This enables Octave extensions to add support for new image file formats or to improve access to the existing formats without addition of new format specific functions. For example, it is now possible for packages to enable reading of microscope specific metadata from `imfinfo`. This system is available via the function `imformats`.

Changes were required in other Octave functions related to images including those involved in conversion between color models, grayscale images, and indexed images to support integer and floating point data types, and multiple dimensions (Table 5.1).

All changes were released with Octave version 3.8.0.

Bio-Formats and Octave Java interface

While GraphicsMagick provides support to read many image formats, its support for scientific microscope image formats is limited. For example, Zeiss confocal microscopes saves images in the LSM file format which is a proprietary extension of TIFF. Although GraphicsMagick reads LSM pixel data, the file metadata such as pixel size, time interval, or region of bleaching event are not retrieved.

Bio-Formats is a free software library for reading and writing image data with a strong focus on microscopy image file formats (*Linkert et al.*, 2010). It is written in the Java programming language and used by other programs in the field of microscope image analysis such as CellProfiler (*Carpenter et al.*, 2006), ImageJ (*Schindelin et al.*, 2015), and OMERO (*Allan et al.*, 2012).

Octave has a native interface to the Java programming language that should enable easy integration with Bio-Formats, and Bio-Formats has a Matlab toolbox that should be compatible with Octave. I identified a series of problems in the Octave interface to Java that could be solved either by improving Octave or Bio-Formats.

In Octave, I rewrote the handling of values returned from Java. Of special importance for BioFormats integration was the automatic conversion of Java arrays which are used to return pixel data. Support for conversion of multidimensional Java arrays was not implemented since I did not require it and this remains open as a future project.

In Bio-Formats, I modified the Matlab toolbox so that it is both Octave and Matlab compatible. This change simplifies the packaging of Bio-Formats for Matlab and Octave making them effectively the same code. Finally, I automated the creation of Octave packages in Bio-Formats so that they can be made available as part of standard Bio-Formats releases.

All changes were released with Octave since version 4.0.0 and Bio-Formats since version 5.1.2

5.3 Octave Forge Image package

The Octave programming language is primarily intended for numerical computations and provides a syntax and set of functions that is particularly convenient for solving linear algebra and differential equations. It also includes a collection of functions for the handling of image data but these are focused on the reading and writing of image files, conversion between colour models, and graphical display.

The Octave Forge Image package supplements Octave with a wide range of specialised functions image analysis. While Octave syntax is identical for any number of dimensions and data type, I found that many of the Image package functions were either limited to two dimensional images or inefficient for the large images such as those used in microscopy. I began an effort to eliminate any limitation on the number of dimensions in the Image package (Table 5.2). This has seen multiple releases, starting in version 2.0.0 and continuing to the current version 2.6.1.

Image Thresholding Algorithms

Image thresholding is a method for image segmentation where the image pixels are separated into classes based on their intensity values. The simplest of thresholding methods uses a single fixed value and separates an image into background and foreground which are then represented in a binary black or white image. This is particularly useful for fluorescence microscopy images where objects of interest show bright high intensity values against a dark background of low intensity values.

Multiple threshold algorithms exist to automate the choice of a threshold value, the most common of which is based on the analysis of an image histogram (*Glasbey*, 1993). Otsu's method (*Otsu*, 1979) is among the most widely used approach and is available in the function `graythresh` in the Image package. I rewrote this function to handle histograms of arbitrary length and vectorised it for performance with additional computation of a "goodness" measure of threshold value optimality. In addition, an unpublished collection of histogram threshold algorithms by Antti Niemistö (personal communication) was made available as an option for `graythresh`. Finally, the rewrite of `graythresh` added the option of using a histogram as input instead of an image, enabling the possibility of histogram processing as a preparatory step.

All these changes were released in the Octave Forge Image package version 2.0.0.

Table 5.2: **New or improved functions in the Octave Forge Image package.**

```
bestblk
bwareafilt
bwareaopen
bwconncomp
bwdist
bwlabel
bwlabeln
bwmorph
bwperim
bwpropfilt
checkerboard
col2im
colfilt
conndef
edgetaper
fftconv2
fftconvn
grayslice
graythresh
im2col
imabsdiff
imadjust
imattributes
imbothat
imclearborder
imclose
imcomplement
imcrop
imdilate
imerode
imfill
imgetfile
imhist
imlincomb
immse
imopen
impixel
impyramid
imquantize
imreconstruct
imregionalmax
imregionalmin
imresize
imrotate
imtophat
intlut
iptcheckconn
label2rgb
labelmatrix
mat2gray
mmgradm
montage
nlfilter
normxcorr2
ordfiltn
otf2psf
padarray
psf2otf
psnr
regionprops
rgb2ycbcr
strel
stretchlim
subimage
tiff_tag_read
watershed
wavelength2rgb
ycbcr2rgb
```

Mathematical Morphology

Mathematical morphology is the analysis of spatial structures, and this provides a powerful image analysis technique based on the shape and form of objects. It is achieved by probing an image with a known shape called the Structuring Element (SE), and filtering the image based on whether the SE fits at each location within the image. Mathematical morphology is a very relevant tool for identifying cell locations and features in microscopy images.

The fundamental operations of mathematical morphology are named dilation and erosion. These are available in the Image package through the `imdilate` and `imerode` functions. More complex morphological operations are built on top of these two operations. For example, the morphological top-hat transform corresponds to the difference between an image and its morphological opening, which in turn corresponds to an erosion followed by a dilation.

As the two fundamental operations, dilation and erosion are ideal targets for improvement since any performance increase or support for new image types will be propagated to higher level morphology functions.

`imdilate` and `imerode` used the general purpose `__spatial_filtering__` function of the Image package. I rewrote the functions aimed at morphology operations with specialisations for different data types and SE. A new `strel` class was created for the SE specialisation which supported both flat and non-flat SE.

These changes were released with the Octave Forge Image package version 2.2.0.

Image Regions of Interest

The function `regioprops` is used to measure different properties of regions of interest in an image such as its centroid, area, or eccentricity. However, the function was inappropriate for the measurement of multiple regions and properties because the whole image was analysed independently for each region and property. I rewrote this function for improved efficiency so that computation of area and intensity weighted centroid of 2000 regions which previously took over 12 hours on a standard desktop computer was finish in under 3 minutes. This improvement was released with the Octave Forge Image package version 2.6.0.

I also wrote the function `bwconncomp` to perform the identification of image regions with an arbitrary number of dimensions. `bwconncomp` is used internally by `regionprops` and creates an array of indices for the image regions as an alternative to labelled images which reduces memory usage. This was released with the Octave Forge Image package version 2.4.0.

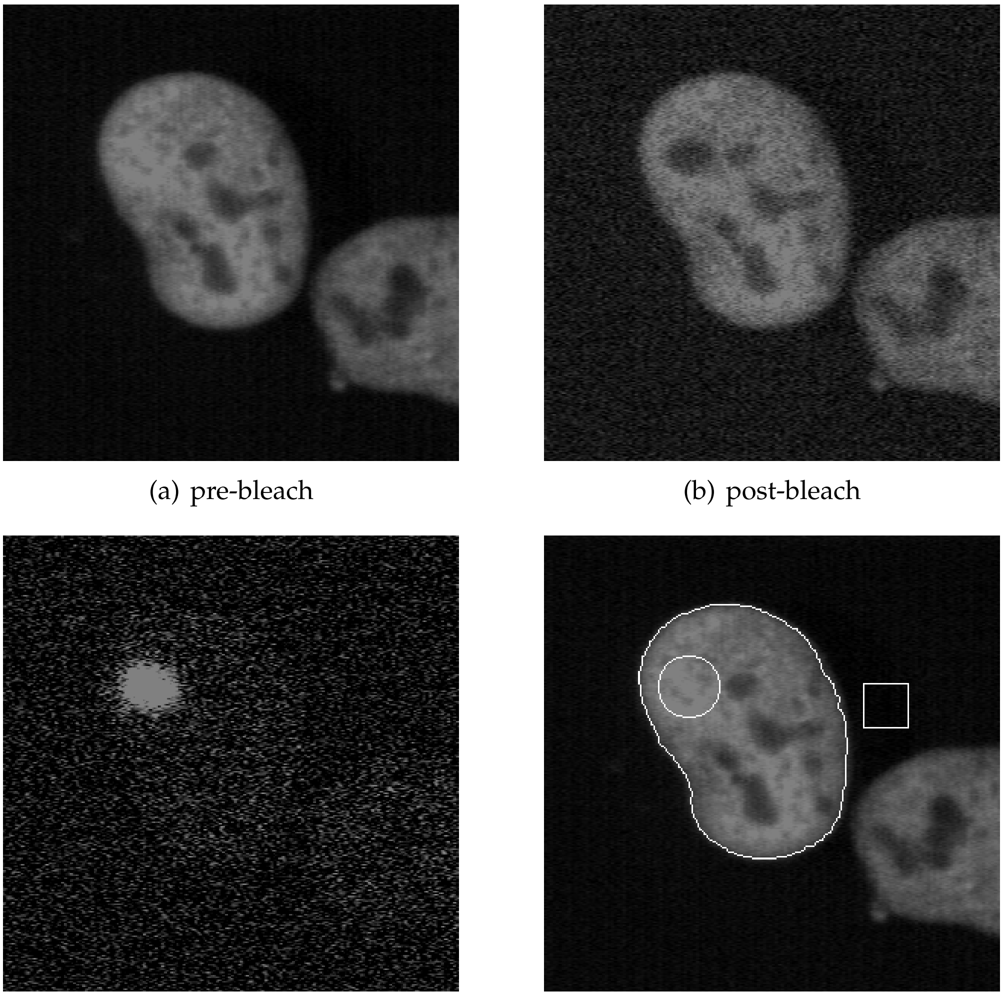
5.4 Octave FRAP package

For the FRAP analysis of histones in Chapter 4 we required tools to estimate binding constants from FRAP recovery data. We obtained the code for a previously reported circle FRAP model (*Mueller et al.*, 2008) by personal communication with the authors (under the GNU General Public Licence (GPL) version 3 or later). This model includes multiple biophysical parameters including the profile of the photobleach, correction for observational photobleaching, finite size of the nucleus, and fitting to both a pure-diffusion model and a full model with binding states.

This code was written in the Matlab programming language so was easily ported to Octave. The main difference was the replacement of the nonlinear fitting from `nlinfit` with `leasqr` from the Octave Forge Optim package since both perform the same Levenberg–Marquardt nonlinear least squares algorithm. To validate my port to Octave, I compared the results obtained with my Octave port against the results obtained by the original authors in Matlab. I picked three of our datasets and both implementations returned the same results.

Identification of the bleach spot, nucleus, and background regions are required for the circle FRAP model. The bleach spot analysis measures intensity recovery and also models the photobleach profile since it takes into account a non-uniform spatial distribution of the bleached spot. The nucleus region analysis defines the finite sized nucleus and takes into account the fluorescence loss due to observational photobleaching. A small region outside the nucleus is used for background correction.

I automated the identification of all these regions to facilitate batch processing. The bleach spot was identified from the difference between



(a) pre-bleach

(b) post-bleach

(c) pre-bleach − post-bleach

(d) Identified ROIs

Figure 5.1: **Automatic selection of regions for FRAP analysis.** HeLa stable cell line expressing the H4 R45H mutant tagged with YFP were imaged every 30 ms in a confocal microscope. A circular shape is used for photobleaching after 100 frames. (a) averaging of 50 pre-bleach images removes most of the noise, allowing for a better refined ROI; (b) average of 5 post-bleach images; (c) subtraction of the post-bleach to the pre-bleach image, gives a clear indication of the bleach spot, as well as faint signal for the nuclear region due to unintentional photobleaching; (a) perimeter of the automatically identified ROIs superimposed on the pre-bleach image: Cell nuclei, bleach spot, and background region.

the post and pre-bleach frames. Individual nuclei were then segmented after automatic thresholding with Otsu's method. A background region was identified as the rectangle of a fixed size with lowest average intensity in the image.

With all the steps automated, I created a single program written in Octave that we named `frapinator` to perform all the analysis for any number of images. All options were made available as command line options. To quickly filter out any faulty analyses, two multi-panel images are created that provide a visual log. One image shows the automatically identified regions (Figure 5.1) and another shows the recovery curves, intermediary analysis, and best fits (Figure 5.2).

I packaged the `frapinator` program and all the FRAP analysis functions into an Octave FRAP package and made it available online at `https://github.com/carandraug/octave-frap`.

5.5 BioPerl

The BioPerl project is an international association of developers of free Perl software for bioinformatics, genomics, and life science (*Stajich et al.*, 2002). It has created the BioPerl distribution of Perl modules which contains almost 800 modules for management and manipulation of biological data as well as programmatic access to databases such as GenBank and SwissProt, and to bioinformatics tools such as ClustalW and Blast+.

The large number of modules in BioPerl became a maintenance problem so in 2011, a new project was initiated to split BioPerl into more manageable module distributions such as Bio-Biblio, Bio-FeatureIO, and Bio-Coordinate.

Dist::Zilla and Pod::Weaver

To reduce existing code, prevent duplication, and to make new releases a easier, Dist::Zilla and Pod::Weaver were adopted for the new distributions.

Dist::Zilla is a program facilitating the writing, packaging, management, and release of free software for libraries that are written in the Perl programming language and released to the Comprehensive Perl Archive Network (CPAN) repository. It comes with a series of plugins to automate the release process such as the addition of copyright notices, discovery of dependencies, and uploading to CPAN.

Pod::Weaver is a program to create documents in Plain Old Documentation (POD) format, a format mainly used to write documenta-

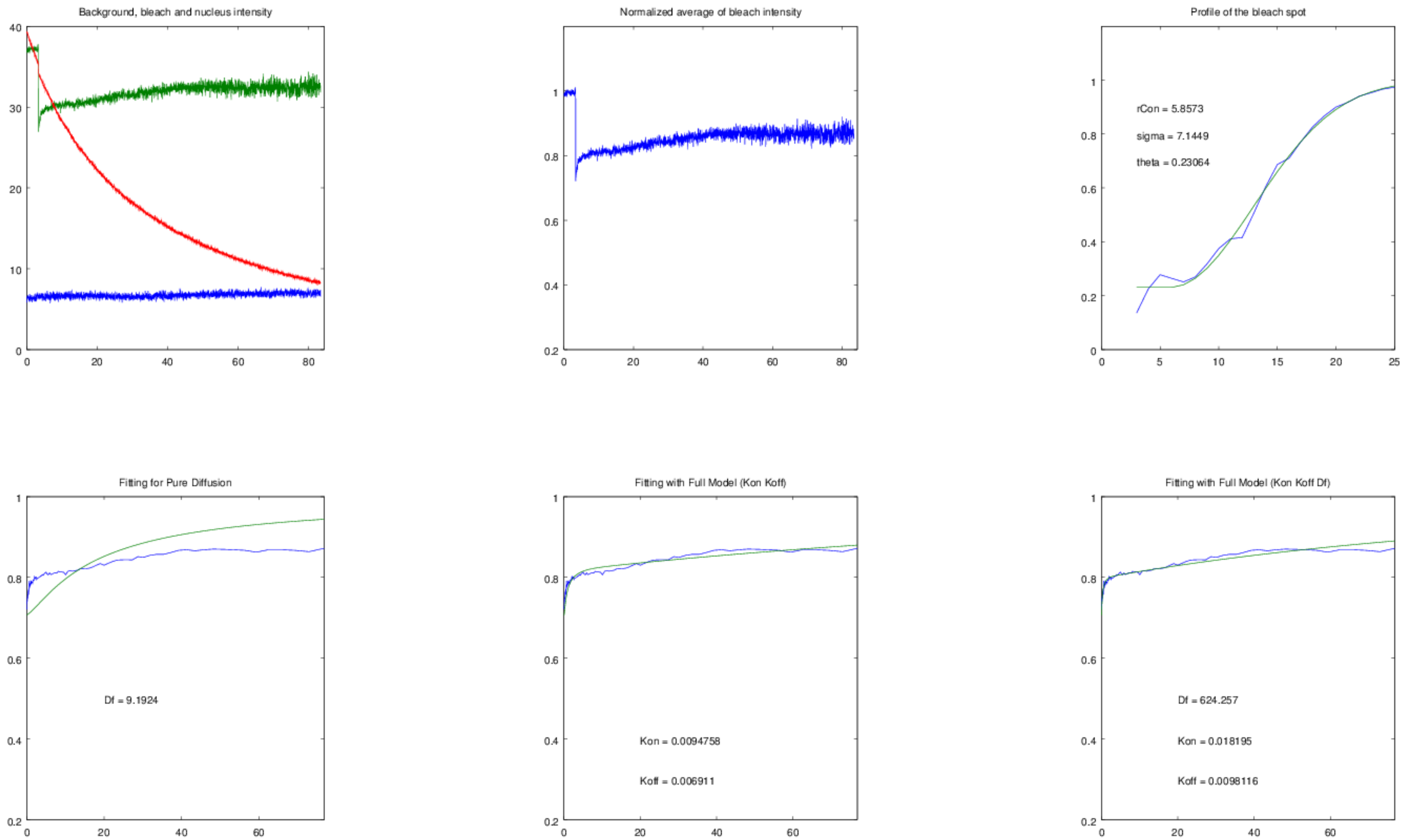


Figure 5.2: **Frapinator visual log files for batch processing.** Each FRAP experiment generates a log file with 6 different plots displaying the analysed values and the fitting to different models. In conjunction with the images in Figure 5.1 this provides an overview of the entire analysis process. The top left plot displays the raw intensity for the background, bleach spot, and nucleus intensity over the duration of the FRAP experiment. This is followed by the normalised intensity for the bleach spot which is then used for the fitting. The top right displays the intensity profile for the bleach spot, and its fit to a radial profile model. The three bottom panels display the data fitted to three different models: Pure diffusion which has no terms for binding constants, full model with a fixed diffusion rate, and full model with all the 3 terms.

tion incorporated inline within Perl modules. Pod::Weaver includes a Dist::Zilla plugin so that most standard generic POD content is generated automatically as part of the release process.

As part of the restructure of the BioPerl distribution, I configured the BioPerl Dist::Zilla plugin bundle, and created two new Pod::Weaver section plugins, GenerateSection and Legal::Complicated.

GenerateSection creates POD sections based on templates. It is used in BioPerl to generate the support section of documentation of individual modules with distribution specific details such as links to the online repository.

Legal::Complicated creates a POD section for copyright details based on comments in individual modules. This is useful because while BioPerl is free software, individual modules may be released under different free software licenses, and each module has its own author who may differ from the copyright holder.

Both new Pod::Weaver plugins and the BioPerl PluginBundle are available on CPAN since 2013. They were used for my contributions to the Bio-EUtilities package.

Bio-EUtilities

For the analysis of the canonical histone gene family (Chapter 3) we required a tool to automate the search of histone genes and download associated sequences. I used the NCBI Gene database for searches and created a new program `bp_genbank_ref_extractor` within the Bio-EUtilities package of BioPerl.

Bio-EUtilities is part of the BioPerl project and provides a Perl interface to NCBI's Entrez Programming Utilities (E-Utilities). Entrez is a federated search engine for multiple databases of biomedical data including Gene. Entrez has an interactive interface at `https://www.ncbi.nlm.nih.gov/` while E-Utilities provides an equivalent programming interface for queries using a fixed URL syntax.

Gene is a public database hosted at the National Center for Biotechnology Information (NCBI) which maps known or predicted genes to other entries in the NCBI Reference Sequence (RefSeq). Therefore, Gene links to the Genome, Nucleotide, and Protein databases (*Brown et al.*, 2015).

The program `bp_genbank_ref_extractor` was created to take a query to the Entrez Gene database as input and to downloads all genomic, transcript, and protein sequences as well as a CSV file with chromosome coordinates, names, and identifiers of returned genes as output. It has several options such as download of flanking sequences, different output formats, choice of genome assembly, and skipping of non coding genes.

`bp_genbank_ref_extractor` is provided with an extensive manual covering all options and examples (Appendix D).

`bp_genbank_ref_extractor` was released with Bio-EUtilities version 1.73.

Debian packaging

Debian is a computer operating system composed entirely of free software and is one of the earliest GNU/Linux distributions.

Debian is package based like all modern free operating systems. This means that it is made out of multiple components known as packages. For example, in Debian there are packages for the Linux kernel, the Perl programming language, and Dist::Zilla. Packages are managed by a package management system which handles their installation, configuration, and removal to simplify a multiplicity of small steps that would otherwise have to be handled manually by the user.

Debian is a widely used GNU/Linux distribution with more than 21000 packages and a large number of derivative distributions. These new distributions inherit the base of their packages from Debian, and some like Ubuntu and Knoppix are in turn the parents of their own derivative distributions. By packaging for Debian, a packager effectively prepares packages for the whole family of Debian based distributions.

While Debian had a package for the main BioPerl distribution, it did not have one for Bio-EUtilities. I packaged Bio-EUtilities for Debian with the aim of making it easier for others to reproduce our results. Similarly, to make it easier for prospective BioPerl developers, I packaged all the Dist::Zilla plugins required to produce new BioPerl releases as well as all the module distributions required by them (Table 5.3).

Table 5.3: **Perl module distributions packaged for Debian.**

Bio-EUtilities
Webagent which interacts with and retrieves data from NCBI's E-Utils.

Config-MVP-Slicer
Module to extract embedded plugin config from parent config.

Dist-Zilla-Config-Slicer
Config::MVP::Slicer customized for Dist::Zilla.

Dist-Zilla-Plugin-AutoMetaResources
Dist::Zilla plugin to ease filling `resources` metadata.

Dist-Zilla-Plugin-MojibakeTests
Dist::Zilla plugin that provides author tests for source encoding.

Dist-Zilla-Plugin-ReadmeFromPod
Dist::Zilla plugin to generate a README from POD.

Dist-Zilla-Plugin-Test-Compile
Common tests to check syntax of Perl modules, using only core modules.

Dist-Zilla-Role-PluginBundle-PluginRemover
Dist::Zilla plugin to add `-remove` functionality to a bundle.

MooseX-Types-Email
Email address validation type constraints for Moose.

Pod-Weaver-Plugin-EnsureUniqueSections
Pod::Weaver plugin to check for duplicate POD section headers.

Pod-Weaver-Section-Contributors
Pod::Weaver plugin for a section listing contributors.

Pod-Weaver-Section-GenerateSection
Pod::Weaver plugin to add POD sections from a template text.

Pod-Weaver-Section-Legal-Complicated
Pod::Weaver plugin for per module authors, copyright holders, and license.

Test-Mojibake
Module to check source for encoding misbehaviour.

5.6 Build systems for reproducible research

Even if the original data is available for a computational biology investigation such as microscopy image or genome sequence analysis and the runtime environment can be duplicated, reproducing results is dependent on invoking the same commands and same options in the same order as the original analysis. The necessary information to achieve this is often undocumented and difficult to reconstruct.

A build system is a software tool that automates the process of performing a complex series of steps for the generation of an artefact. It is mainly used in software engineering to automate software compilation and packaging but the process of maintaining software has many parallels with maintaining reproducible research projects in computational biology so the same tools can be used. For example, in a software compilation project, object code is built from the source code whereas in a research project, figures, tables, and values are built from raw data. Likewise, in a software project, an executable program is built from multiple object code whereas in a research project a manuscript is built from the figures, tables, and values.

Claerbout and colleagues (*Schwab et al.*, 2000) proposed this parallel between software engineering and research projects, and coined the term "reproducible research". They created a standard build system for the generation of figures and manuscript from author data and analysis software as an extension to GNU Make, one the most common build system. Following on from this, Madagascar (*Fomel and Hennenfent*, 2007) was developed as a software solution specialised for reproducible computational experiments, based on the software build system SCons.

The SCons (from Software Construction) build system (*Knight*, 2005) was designed to be a replacement of Make and resembles Make in concept. However, it has the advantage of being configured using Python which is a modern programming language often praised for its readability.

For the work in Chapter 3 I made extensive use of the Perl programming language which is not supported by default in SCons but is commonly used in bioinformatics. To simplify the use of SCons for our project I created a SCons perl tool which adds automatic prerequesite scanning, perl configuration options, and multiple functions for using Perl scripts.

The SCons perl tool is available at `https://bitbucket.org/carandraug/scons-perl5` under a free software licence.

5.7 Discussion

Reproducible research has two principal motivations. One reason is to allow other researchers to reliably perform the same analysis. This can be on the same dataset as the original research in order to identify possible errors in the methodology and reflects the essential replicability requirement of all science. When associated with the freedom to inspect the source code, reproducible research enables anyone to check for errors directly in the code. All software used in this thesis is free and open source so any part of it can be interrogated and replicated.

The second reason, and the one of most important concerns of this chapter, is to enable others to use the tools and insights developed for new research directions. This necessitates a particular approach to the development of scientific software, requiring more insight in the software design and engineering to ensure its usefulness.

It is not sufficient to develop software that solves our problem within the constraints of our data because the software must also be configurable and generalised so that it can fit the needs of other users. For example, the Pod::Weaver plugins I created to solve the problem of making BioPerl releases self-documenting was implemented with extra capabilities that were not needed for the project. The code could also have been implemented directly in the BioPerl PluginBundle but having them as separate individual plugins enables other Perl software projects to use them.

As free software, there is no sales or reliable download data, so I do not know exactly how many people use my work since I mainly hear when parts do not work. I do know people are using it because they have reported issues and even suggested improvements to me. This happened for the Pod::Weaver plugins when users with newer versions of Perl reported issues before I had tested them ourselves.

Users reporting issues are an important signal of interest so the more users I have the more issues are reported to me. My work with Octave is likely to have had the most impact since issues are reported or discussed for the Octave Forge Image package several times a month. However, my work there has become positively entangled with that of other contributors so I cannot measure the exact impact of single contributors alone. This means that even if I had the information on how many people use Octave, and how many use the Image package, it would be impossible to measure how much use is being made of the specific functions I improved.

We recognise that full fledged programs impose limitations even though they are popular with more users who want "turn key" solutions. A program is easier to make use of by novice users but the availability of the same functionality as a software library allows reuse by other software developers, and the two are not mutually exclusive. This can be addressed by coding with the potential to be useful for others by writing with modularity in mind so that any part can be used in other programs. For example, when writing `frapinator` to perform our analysis of FRAP data, I wrote in such a way that the program is a shell around our Octave FRAP package. `frapinator` simply reads the options and then calls a series of a functions from the FRAP package. This allows other researchers to reuse the parts they need. For example, image analysts can utilise the functions to identify the bleach spot but implement alternative FRAP models, possibly from another such library.

Maintaining the balance of flexibility and usability requires additional effort. For example, my choice to work with the Debian base to package software we needed for our projects stemmed from the difficulty members of our group had with installing some of the required dependencies. I could have installed it on their behalf but that wouldn't benefit other researchers in the same position. The situation could also reoccur each time a user got a new machine or update.

Another alternative would be to include dependencies within our software. However, this has its own risks by making the dependencies opaque. For example, part way through our project NCBI moved their E-Utilities services to an encrypted network protocol which had been quickly addressed in Bio-Eutilities but would have necessitated updating our entire histone catalogue (Chapter 3) if we had included an outdated version of Bio-EUtilities directly with it.

5.8 Conclusion

The various software tools reported in this chapter were developed because they were needed for our research. Reports from users suggest that they have been useful for others. Furthermore, the software has been developed in a modular way to facilitate reuse, but also made available with packaging to enhance general uptake.

CHAPTER 6

Discussion and Future Perspectives

Metaresearch, or Research on research, is a very topical area and has become widely discussed after a series of recent studies showing low reproduciblity rates for published research (*Ioannidis et al.*, 2015). The most extensive of these reproducibility studies have been in the field of psychology (*Open Science Collaboration*, 2015), computational science (*Collberg and Proebsting*, 2016), and cancer biology (*Prinz et al.*, 2011; *Ioannidis et al.*, 2009). They identified the lack of data availability and details on methodology in analysis steps as the main causes of difficulties in reproducibility.

In this thesis we attempted to follow the best practices for reproducible computational analysis of biological data. Central to our approach was the automation of the data analysis using software build utilities. These tools require explicitly declared instructions on how to generate a certain file or reach a specific goal such as performing analysis on a dataset to generate a figure for a manuscript. Having such a system in place ensures reproducibility, makes the analysis transparent, and enables other researchers to build upon the work more easily.

We used SCons, a software build system, to automate the analysis steps of all our data. Even so, this still requires a researcher wishing to reproduce our results to install all the required software and we recognise that this is not a trivial task. To support this challenge, we included configure scripts which are run at the start of the build system and identify missing tools or features. GNU/Linux distributions in turn ease the task of installing the multiple tools by providing a package manager, so we have also contributed to the packaging of some of these required software components for the widely used Debian distribution.

Finally, once the methods are available, access to the data is required to reproduce the analysis. In our case, we applied our approach to both data from public sequence databases (Chapter 3) and to our own experimental microscopy image data (Chapter 4). For the sequences in Chapter 3 the usefulness is enhanced because our analysis can be rerun after improvements in the source data as well as for repeatability of results. Nevertheless, we made available a snapshot of the data used to build this thesis for historical validations. For Chapter 4, we made the data publicly available in a research data repository.

This setup has enabled us to generate thesis and manuscript documents where each figure and table can be traced directly back to the raw data and the code used to generate them is freely available. In Chapter 3, references to values are inserted in the text as part of the analysis to avoid loss of validity as the data changes and the analysis improves. In Chapter 4, the individual images, regions, and scaling options used to create the inset figures are all explicit in code. Even in the Introduction and Discussion for both chapters, we automated the generation of all figures. For example, Figure 1.4 can be regenerated from the crystal structure, as we include PyMOL scripts for their creation, as is the Ti*k*Z source code for LaTeX to generate Figure 1.5 so that others can improve it.

Only one exception to complete transparency exists in this thesis. Figures in Chapter 4 from the DeltaVision microscope were deconvolved automatically after acquisition by SoftWoRx which is non-free software. As this was the very first step after image acquisition, we trace all the steps from these deconvolved images.

6.1 Data in Flux

We found that our Histone Catalogue (Chapter 3) was particularly well suited to this approach to transparent generation of a manuscript from original data. Not only are all the steps of a computational nature without any wet-lab work, but the transitory nature of sequence data makes the automated implementation of reproducible research useful beyond the intellectual of achievement making it reproducible.

When we started the Histone Catalogue project, the last publication of an equivalent catalogue was from 2002 and was becoming outdated. We initiated this project with the aim of making a catalogue that would be constantly up to date. One year later in 2011, the Histone Database `http://research.nhgri.nih.gov/histones/` website with an up-to-date database specialised for histones was released (*Mariño-Ramírez et al.*, 2011). However, no update was made to that database until 2016 when a version 2.0 was released (*Draizen et al.*, 2016). We assume that further updates will be infrequent and dependent on funding. The Histone Database is a manually curated database of histone genes of all species so it holds the same human sequences as we have catalogued. However, it does not have the extent of analysis of our catalogue. While the Histone Database reflects a fixed point in time, its manual curation could give a higher quality database.

Our automated approach is inferior in principle because it is dependent on the annotation of the existing genes in public repositories while the Histone Database uses searches for sequences with specific predicted structural motifs followed by manual curation. However, as long as the logic behind the structural motif prediction can be encoded, our Histone Catalogue could be adapted to perform the same search as part of the build. Alternatively, the Histone Catalogue could continue being built using annotations as a representation of what is currently recognised. We created supplementary Table A.7 to list differences between the current model and annotation with the purpose of identifying anomalies and either improving the model or the annotation. Anomalies to the annotation can be fed back to RefSeq, as we have done many times since we started our Histone Catalogue project.

The main point of our approach is that readers will have access to the code. They will not be limited to downloading the sequences we found at a certain time, since they can get the sequences at their point in time and can modify the code to perform alternative analyses or use alternative data sources. The design of our catalogue is modular such that the choice of source data is just a block on which the rest of the analysis sits. Similarly, all tables and figures are individual blocks, so new tables and figures could be added to a new type of catalogue. By making the whole code base public we are providing the full opportunity to expand it.

The initial purpose of the Histone Catalogue was to tabulate all human histone genes and the differences between their coding sequences. With time, we added other figures and became curious about the unique regulatory elements of histones which led to us adding an option to download downstream sequences in the Bio-EUtilities code and then to the identification of multiple regulatory elements that were not annotated in the RefSeq sequences. We were happy to find NCBI responsive to such corrections and they made our submissions live for all RefSeq users.

We also tested the advantage of modifying the code ourselves for new data sources. The original catalogue was limited to human genes but it failed when applied to the analysis of the mouse genome. We changed the software catalogue to make it organism independent and have demonstrated its functionality by building a complete set of figures for mouse canonical core histones (Appendix B). We envision that a similar catalogue could be generated for other gene families by adapting the code base.

6.2 Data not in Flux

By using sequence data already available in public sources for the Histone Catalogue, we avoided the issue of making our own primary research results public. Public repositories dedicated to scientific data for multiple types of information have already been created such as GenBank for nucleotide sequences, Protein Data Bank (PDB) for the structure of molecules, and FlowRepository for flow cytometry. However, databases specialised for light microscopy are still not available (*Lemberger*, 2015).

We faced this issue for the FRAP data we acquired in Chapter 4 and originally solved it by deploying a public OMERO server, a system designed for storage and sharing of biological data, specialised for microscopy (*Allan et al.*, 2012). The infrastructure was provided by e-INIS, the Irish National e-Infrastructure project to support the Irish research community. This was planned to provide our Centre for Chromosome Biology (CCB) with a centralised location for sharing data between researchers and possibly with the outside community. However, few CCB users trialled the system and it did not achieve the uptake we would like. Ultimately, the e-INIS project was closed, and left without the required infrastructure we shut down the server. This left us again without a platform for sharing our microscopy data, and reflects an unexpected challenge for the reproducible research approach.

While there is no repository for microscopy data, there are general purpose unstructured repositories for research data. BioStudies is a database hosted by EMBL–EBI to collect data associated with scientific publications (*McEntyre et al.*, 2015) while Zenodo hosted at CERN provides a "catch-all repository" for research data. In the end, we made all the data used to build this thesis available for download on Zenodo and DOI `10.5281/zenodo.377035`. In addition to the microscopy data, and in the interest of replicability, we also included the snapshot of histone sequences used to build this thesis.

This experience also exposed the issue of long-term stability of resources. If the e-INIS project had terminated after completion of the thesis, the dataset would no longer be available for future readers. Even if the e-INIS project had not come to an end, maintenance of the service for the rest of the user group was not trivial so there would have been the need for a specialised system administrator when the project ended. This means that preserving data for posterity may require a party with large and long term resources such as NCBI, EMBL–EBI, or CERN.

6.3 Runtime Environment

In computational research, a major challenge to reproducibility is runtime environments. At the most simple level, it may seem that computation is a matter of installing the required software but this step alone can already be quite complicated for scientific software. However, the interdependency between software at all levels can also have an impact on the results. Even hardware can be an issue, although higher levels languages such as Python, R, and Octave provide an abstraction from hardware that reduces this.

One proposal has been the distribution of Virtual Machines (VM) with the entire stack of tools required to reproduce an analysis (*Hurley et al.*, 2015; *Angiuoli et al.*, 2011). However, this is effectively redundant with the repositories of the GNU/Linux distributions they are based on, so VMs are very wasteful in storage space. It also misses an important point of the reproducibility in transparency of how the system was built.

In free software distributions such as Debian, it is possible to specify the build environment in a machine readable format that can be reused later to reproduce the environment. In the case of Debian, the `.buildinfo` control file encodes all packages, with version details, used to perform a build, as well as hardware architecture, and a file checksum of the result for validation. This could greatly simplify the management of VMs, as only a single copy of the distribution repositories would have to be stored. However, this requires all software to be packaged.

We have packaged the dependencies for our Histone Catalogue in Debian. We have also added configure scripts which run before the build and check for installed dependencies. In addition, we also created a test suite for our Histone Catalogue to confirm that the environment works as expected while at the same time supporting our own development process by preventing regressions.

6.4 Continuous Delivery

While researchers can, and we hope will, continue to develop and use our Histone Catalogue approach, the level of skill required to reproduce it is not common in the biochemistry and cell biology community. Without a local bioinformatician, the usefulfness of this power can be lost. This means that training opportunities and encouragement needs to be provided to increase the relevant skills of typical molecular cell biologists if the vast resources of data available are to be fully used.

For situations where data is in flux as is the case of unversioned public databases such as RefSeq used in our Histone Catalogue, the software approach of continuous delivery could be applied to sequence data analysis. In this approach, software is developed and is always in a state that can be released. Instantaneous releases are automated and anyone can download the release with the latest change. In the case of the Histone Catalogue, changes to the data could be configured to trigger an updated build of the sequence catalogue. This would allow a biochemist without bioinformatics aspirations to download the latest analysis as the manuscript PDF.

Several tools already exist to perform such automatic builds such as Buildbot or Jenkins. While our Histone Catalogue is not currently triggered by data updates, distributing it via a public build server that is triggered by a new RefSeq release, roughly every two months, is an interesting future project that would make it more useful to the wider research community.

The same approach and tools could be used for an ongoing project as new data is being acquired. In our project for estimating the effect of point mutations on the dynamics of histone proteins by FRAP, new images of FRAP experiments could trigger their analysis to refine the estimates obtained from the previous data. Similarly, if we change the FRAP model, for example, to account for an artefact discovered as part of the research, this would trigger re-analysis of all previous data. Even automatic comparison between new mutations could be automated in the same manner. This approach can be useful as the project advances and provide a live indication of progress as the results become more accurate.

6.5 Version Control

The entire development of this thesis is available online as a git repository at `https://github.com/carandraug/phd-thesis`. The use of a version control system provides access not only to the thesis at the point of submission, but also to its entire history. This is of less use in a PhD thesis which was written only at the end of the project. However, both Chapter 3 describing the Histone Catalogue, and Chapter 4 describing our approach to histone dynamics by FRAP, have been written as an individual manuscript with the same ideals.

The goal of version control is to track changes to files, which is a distinct task from the goal of reproducing the complete analyses. However, version control provides extra information about a project to anyone interested in continuing it by exposing the reasons for the choice of methods implementation. For example, the core representation of an histone gene, which is central to the Histone Catalogue project, has gone through three major revisions and individual analysis scripts were constantly modified. A version control system associates each change with a message so that each line of the analysis can be traced back in history to an explanation about why it was done. We used this feature multiple times while developing the Histone Catalogue when a piece of existing code raised questions about the analysis or interpretations. This is quite common and version control systems will have a command that annotates each line with the version that introduced it.

As we used LaTeX for writing our chapters, we were also able to use version control for the writing of the manuscripts. Overall, this collaborative authoring process was very successful. One of the main reasons for this is the support version control provides for merging of changes between multiple authors editing the same document in parallel. However, the tools for handling the merging are specialised for source code and work on lines, which is very different from text where differences happen in sentences. The creation of merge tools for text could simplify this greatly by making changes easier to track and commit messages more meaningful to the authoring process.

Even if such tools were available, the concept of a manuscript as a text file that is compiled into a document, the syntax of LaTeX or even alternative markup formats are all alien concepts in biology. In a recently submitted paper describing "Good Enough Practices in Scientific Computing" (*Wilson et al.*, 2016), the authors also recognised this issue, and so provide an alternative recommendation of using an online collaborative word processor. While this satisfies having a central master document accessible to all authors, it would prevent the automatic insertion of up to date figures and values as we do in our Histone Catalogue.

The availability of the history of a project through version control could also be used for meta-research because it would provide data on the method of writing scientific publications.

6.6 Software Engineering and Biology

We faced many issues with this project due to our lack of knowledge in software engineering. Biologists do not typically receive the training for this type of work, although our observation is that almost all our problems had solutions that were obvious to software engineers. Concepts such as a build system, version control, test units, and reproducible builds are not new in the field of software engineering but biologists are still catching up. This is why new organisations like software carpentry `https://software-carpentry.org/` and data carpentry `http://www.datacarpentry.org/` have been established to teach researchers methods and perspectives that are already practised in software engineering.

6.7 Conclusion

While ultimately we concluded that FRAP is unsuitable for our original research on histone contribution to nucleosome stability, we developed and improved a multitude of tools for other researchers. The Octave FRAP package provides researchers with a flexible environment FRAP analysis, and the work in Octave to support N dimensional images provides support for the new imaging modalities.

In addition, we created a catalogue of all human canonical core histones, and by handling the manuscript as a software project we hope to have created the basis for future projects. For example, our Histone Catalogue could be "forked" for an alternative gene family. Our lab had previously manually generated a catalogue of the Snf2 subfamilies (*Flaus et al.*, 2006) which could be improved in the same manner.

Overall, a major finding is that it is possible to implement manuscripts and entire thesis by following the reproducible research approach for the computational and analysis phases of chromatin research. This openness has a number of important advantages for reproducibility and progress in the quantitative molecular cell biology.

APPENDIX **A**

Human Canonical Core Histone Catalogue

Table A.1: Changes in human canonical core histone gene catalogue annotated in NCBI RefSeq obtained on 2017-03-13 compared to *Marzluff et al.* (2002).

New genes:
HIST1H2AG, HIST1H2APS1, HIST1H2APS2, HIST1H2APS3, HIST1H2APS4, HIST1H2APS5, HIST1H2APS6, HIST1H2BPS1, HIST1H2BPS2, HIST1H2BPS3, HIST1H3PS1, HIST1H4PS1, HIST2H2AA3, HIST2H2AA4, HIST2H2BF, HIST2H3D, HIST2H3DP1, HIST2H3PS2, HIST2H4A, HIST2H4B

Changed sequences:
HIST1H2AA, HIST1H2AB, HIST1H2AD, HIST1H2AE, HIST1H2AH, HIST1H2AJ, HIST1H2BA, HIST1H2BB, HIST1H2BJ, HIST1H4G, HIST2H2AB, HIST2H2AC, HIST3H2A, HIST3H3

Removed genes:
HIST1H2AF, HIST2H2AA, HIST2H4

Now identified as coding genes:
HIST2H3A

Table A.2: Catalogue of annotated human canonical core histone genes, transcripts, and encoded proteins. Data obtained from NCBI RefSeq (*O'Leary et al.*, 2016) on 2017-03-13.

Type	Gene name	Gene UID	Transcript accession	Protein accession
H2A	HIST1H2AA	221613	NM_170745	NP_734466
H2A	HIST1H2AB	8335	NM_003513	NP_003504
H2A	HIST1H2AC	8334	NM_003512	NP_003503
H2A	HIST1H2AD	3013	NM_021065	NP_066409
H2A	HIST1H2AE	3012	NM_021052	NP_066390
H2A	HIST1H2AG	8969	NM_021064	NP_066408
H2A	HIST1H2AH	85235	NM_080596	NP_542163
H2A	HIST1H2AI	8329	NM_003509	NP_003500
H2A	HIST1H2AJ	8331	NM_021066	NP_066544
H2A	HIST1H2AK	8330	NM_003510	NP_003501
H2A	HIST1H2AL	8332	NM_003511	NP_003502
H2A	HIST1H2AM	8336	NM_003514	NP_003505
H2A	HIST1H2APS1	387319	pseudogene	pseudogene
H2A	HIST1H2APS2	85303	pseudogene	pseudogene
H2A	HIST1H2APS3	387323	pseudogene	pseudogene
H2A	HIST1H2APS4	8333	pseudogene	pseudogene
H2A	HIST1H2APS5	10341	pseudogene	pseudogene
H2A	HIST1H2APS6	100509927	pseudogene	pseudogene
H2A	HIST2H2AA3	8337	NM_003516	NP_003507
H2A	HIST2H2AA4	723790	NM_001040874	NP_001035807
H2A	HIST2H2AB	317772	NM_175065	NP_778235
H2A	HIST2H2AC	8338	NM_003517	NP_003508
H2A	HIST3H2A	92815	NM_033445	NP_254280
H2B	HIST1H2BA	255626	NM_170610	NP_733759
H2B	HIST1H2BB	3018	NM_021062	NP_066406
H2B	HIST1H2BC	8347	NM_003526	NP_003517
H2B	HIST1H2BD	3017	NM_021063	NP_066407
H2B	HIST1H2BD	3017	NM_138720	NP_619790
H2B	HIST1H2BE	8344	NM_003523	NP_003514
H2B	HIST1H2BF	8343	NM_003522	NP_003513
H2B	HIST1H2BG	8339	NM_003518	NP_003509

H2B	HIST1H2BH	8345	NM_003524	NP_003515
H2B	HIST1H2BI	8346	NM_003525	NP_003516
H2B	HIST1H2BJ	8970	NM_021058	NP_066402
H2B	HIST1H2BK	85236	NM_001312653	NP_001299582
H2B	HIST1H2BK	85236	NM_080593	NP_542160
H2B	HIST1H2BL	8340	NM_003519	NP_003510
H2B	HIST1H2BM	8342	NM_003521	NP_003512
H2B	HIST1H2BN	8341	NM_003520	NP_003511
H2B	HIST1H2BO	8348	NM_003527	NP_003518
H2B	HIST1H2BPS1	100288742	pseudogene	pseudogene
H2B	HIST1H2BPS2	10340	pseudogene	pseudogene
H2B	HIST1H2BPS3	100820735	pseudogene	pseudogene
H2B	HIST2H2BA	337875	pseudogene	pseudogene
H2B	HIST2H2BB	338391	pseudogene	pseudogene
H2B	HIST2H2BC	337873	pseudogene	pseudogene
H2B	HIST2H2BD	337874	pseudogene	pseudogene
H2B	HIST2H2BE	8349	NM_003528	NP_003519
H2B	HIST2H2BF	440689	NM_001024599	NP_001019770
H2B	HIST2H2BF	440689	NM_001161334	NP_001154806
H2B	HIST3H2BA	337872	pseudogene	pseudogene
H2B	HIST3H2BB	128312	NM_175055	NP_778225
H3	HIST1H3A	8350	NM_003529	NP_003520
H3	HIST1H3B	8358	NM_003537	NP_003528
H3	HIST1H3C	8352	NM_003531	NP_003522
H3	HIST1H3D	8351	NM_003530	NP_003521
H3	HIST1H3E	8353	NM_003532	NP_003523
H3	HIST1H3F	8968	NM_021018	NP_066298
H3	HIST1H3G	8355	NM_003534	NP_003525
H3	HIST1H3H	8357	NM_003536	NP_003527
H3	HIST1H3I	8354	NM_003533	NP_003524
H3	HIST1H3J	8356	NM_003535	NP_003526
H3	HIST1H3PS1	100289545	pseudogene	pseudogene
H3	HIST2H3A	333932	NM_001005464	NP_001005464
H3	HIST2H3C	126961	NM_021059	NP_066403
H3	HIST2H3D	653604	NM_001123375	NP_001116847
H3	HIST2H3DP1	106479023	pseudogene	pseudogene

H3	HIST2H3PS2	440686	pseudogene	pseudogene
H3	HIST3H3	8290	NM_003493	NP_003484
H4	HIST1H4A	8359	NM_003538	NP_003529
H4	HIST1H4B	8366	NM_003544	NP_003535
H4	HIST1H4C	8364	NM_003542	NP_003533
H4	HIST1H4D	8360	NM_003539	NP_003530
H4	HIST1H4E	8367	NM_003545	NP_003536
H4	HIST1H4F	8361	NM_003540	NP_003531
H4	HIST1H4G	8369	NM_003547	NP_003538
H4	HIST1H4H	8365	NM_003543	NP_003534
H4	HIST1H4I	8294	NM_003495	NP_003486
H4	HIST1H4J	8363	NM_021968	NP_068803
H4	HIST1H4K	8362	NM_003541	NP_003532
H4	HIST1H4L	8368	NM_003546	NP_003537
H4	HIST1H4PS1	10337	pseudogene	pseudogene
H4	HIST2H4A	8370	NM_003548	NP_003539
H4	HIST2H4B	554313	NM_001034077	NP_001029249
H4	HIST4H4	121504	NM_175054	NP_778224

Figure A.1: Sequence logo for the human canonical H2A gene coding regions listed in table Table A.2. Initiator codon ATG and stop codon are omitted.

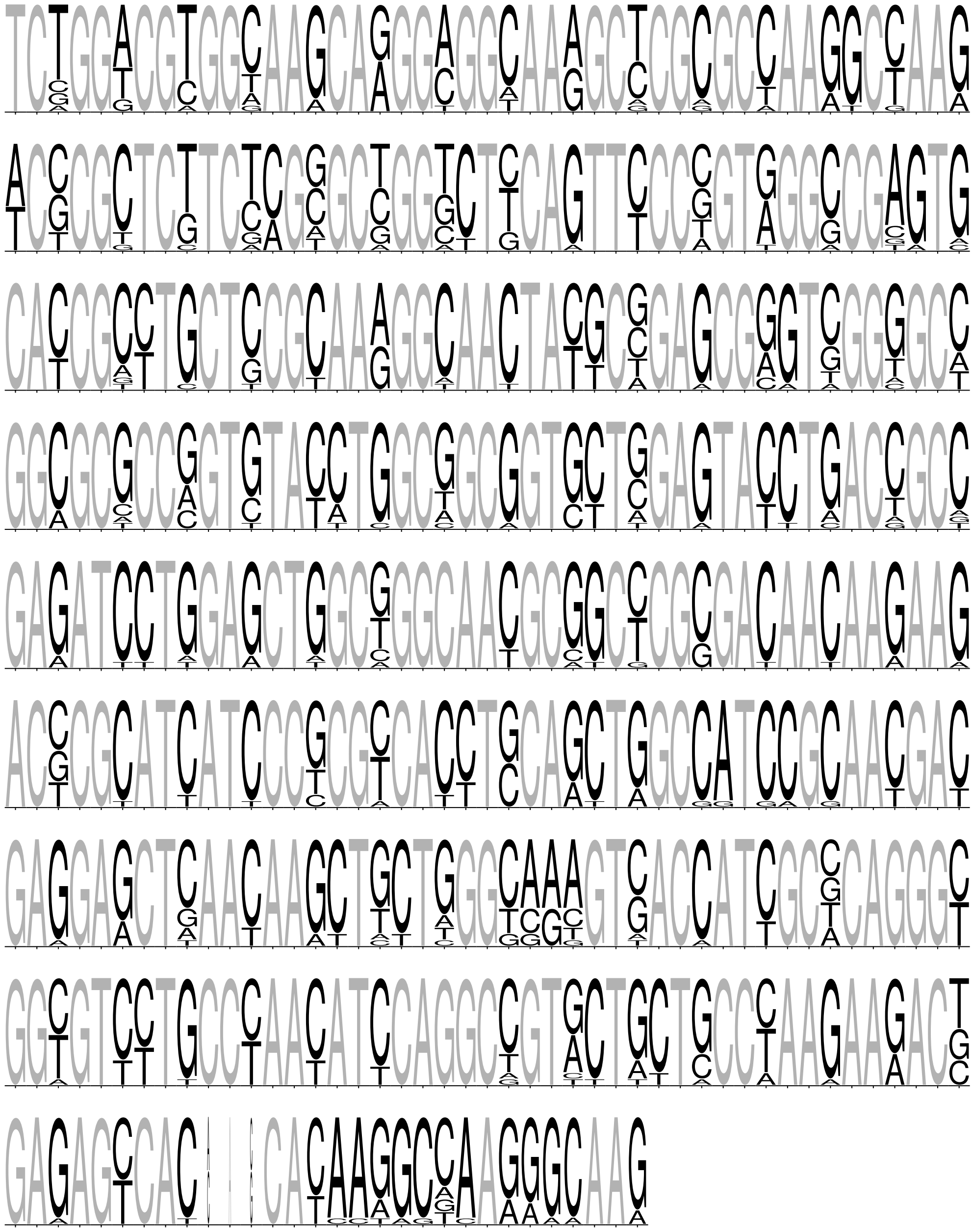


Figure A.2: Sequence logo for the human canonical H2B gene coding regions listed in table Table A.2. Initiator codon ATG and stop codon are omitted.

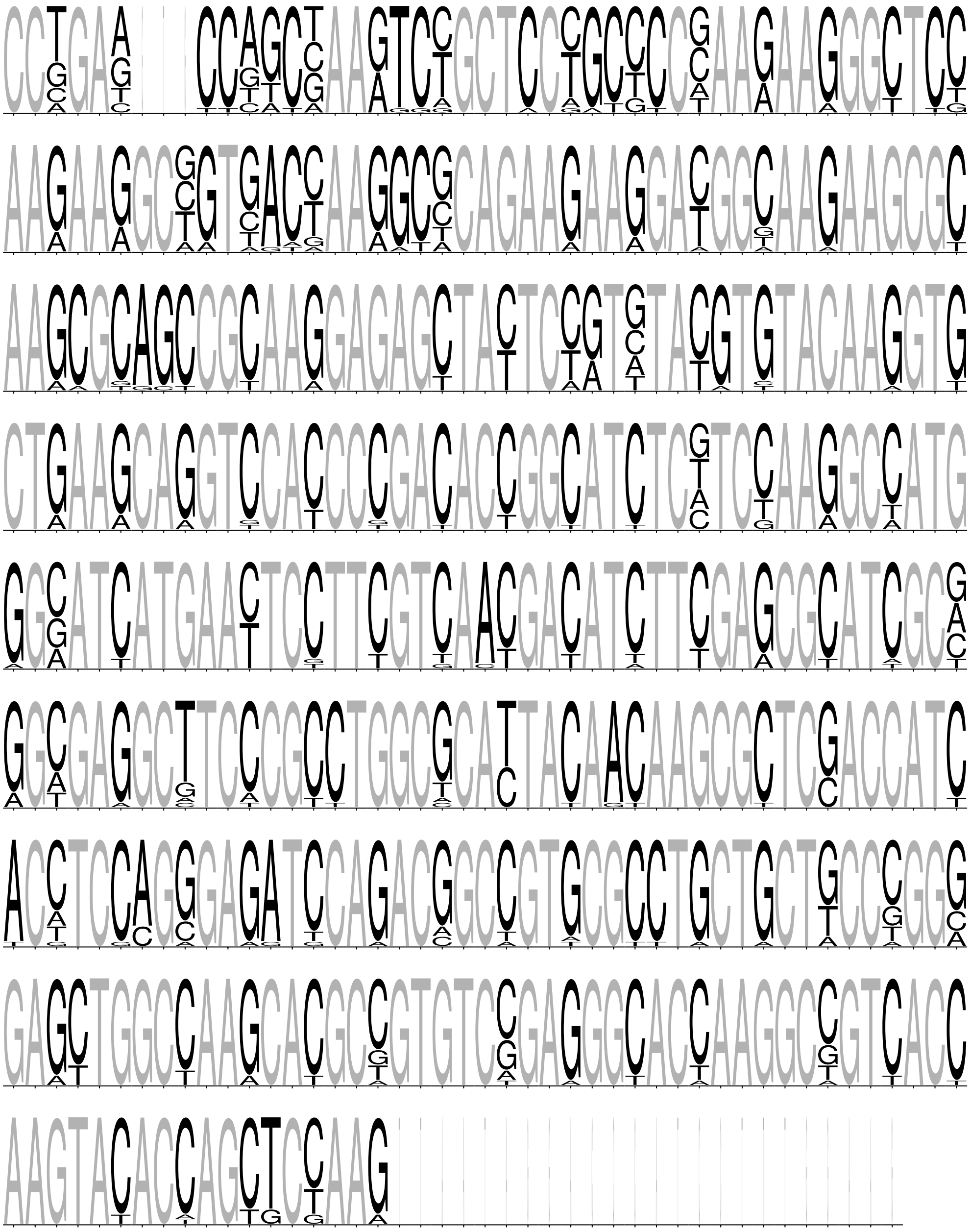


Figure A.3: Sequence logo for the human canonical H3 gene coding regions listed in table Table A.2. Initiator codon ATG and stop codon are omitted.

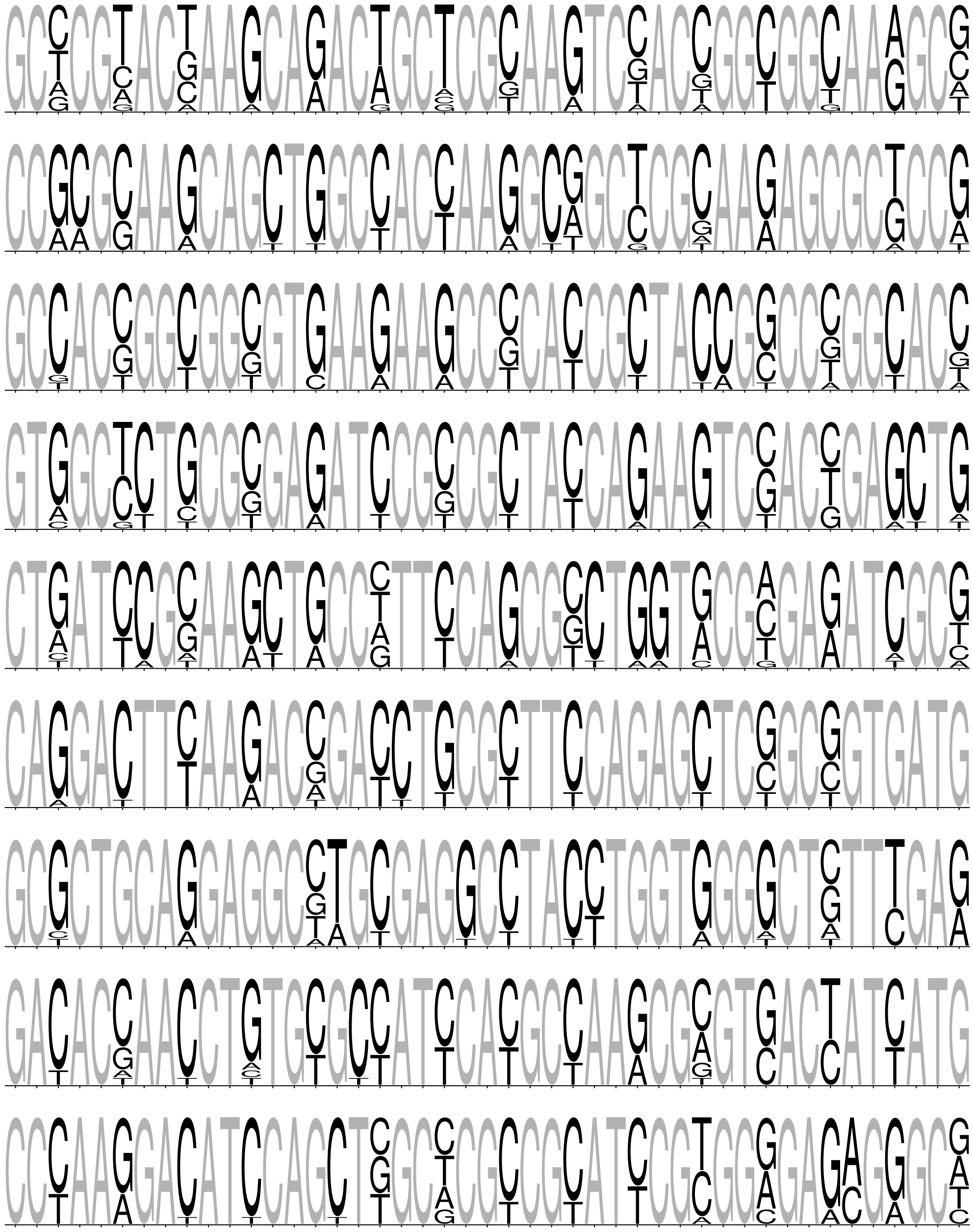


Figure A.4: Sequence logo for the human canonical H4 gene coding regions listed in table Table A.2. Initiator codon ATG and stop codon are omitted.

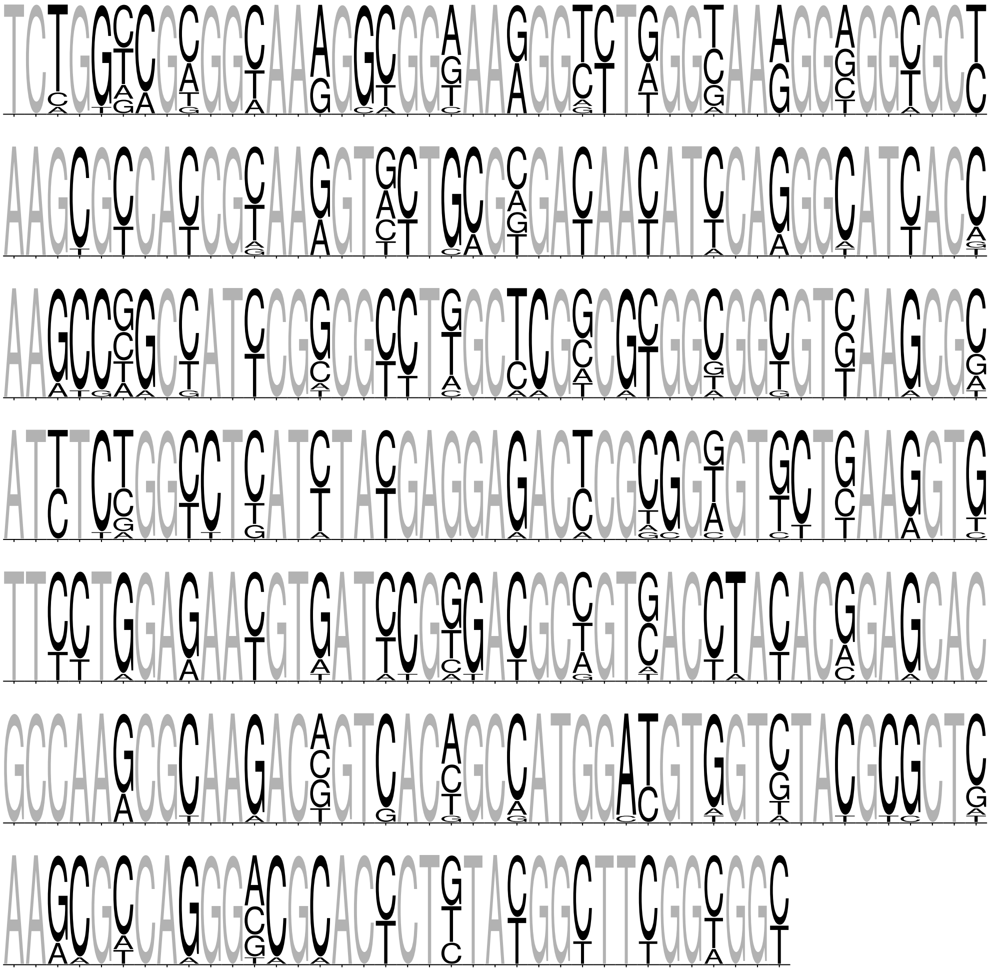


Table A.3: HGNC histone family names (*Gray et al.*, 2015), commonly used protein names, and nomenclature proposal of *Talbert et al.* (2012).

Family	Common name	*Talbert et al.* (2012)	Notes
H2AFB	H2A.Bbd	H2A.B	equivalent to H2A.L
H2AFJ	H2A.J	H2A.J	at HIST4 locus
H2AFV	H2A.Z-2	H2A.Z.2	—
H2AFX	H2AX/H2A.X	H2A.X	—
H2AFY	macroH2A/mH2A	macroH2A	—
H2AFZ	H2A.Z	H2A.Z.1	—
H2BFM	H2B.s	H2B.M	homologous to H2B.W, X-linked
H2BFS	—	—	pseudogene
H2BFWT	—	H2B.W	testes specific, X-linked
H2BFXP	—	—	pseudogene
HIST1H2BA	TH2B/TSH2B	TS H2B.1	testes specific
HIST3H3	H3T	H3.4	testes specific
H3F3	H3.3	H3.3	euchromatin related
CENPA	CENP-A	—	centromere-specific

Table A.4: Mean synonymous and non-synonymous distances between pairs of canonical core histone genes. Mean over all pairwise comparisons of non-synonymous (d_N) and synonymous (d_S) nucleotide substitutions and mean of d_N/d_S ratios for canonical core histone coding regions computed using *Goldman and Yang* (1994) model implemented in `codeml` from the PAML package (*Padavattana et al.*, 2007).

	d_N	d_S	d_N/d_S
H2A	0.0114	3.7823	0.00523
H2B	0.0159	1.726	0.00805
H3	0.0041	2.5328	0.00161
H4	0.0134	2.7556	0.00674

Table A.5: Codon usage frequency for each amino acid and canonical core histone type.

		H2A	H2B	H3	H4
Ala	GCT	0.19	0.29	0.28	0.24
	GCC	0.43	0.42	0.39	0.55
	GCA	0.07	0.07	0.08	0.06
	GCG	0.30	0.22	0.25	0.15
Arg	CGT	0.14	0.07	0.18	0.18
	CGC	0.61	0.83	0.57	0.57
	CGA	0.11	0	0.06	0.05
	CGG	0.12	0	0.13	0.16
	AGA	0.01	0.01	0.01	0.03
	AGG	0.02	0.10	0.05	0.00
Asn	AAT	0.15	0.21	0.07	0.35
	AAC	0.85	0.79	0.93	0.65
Asp	GAT	0.12	0.22	0.14	0.42
	GAC	0.88	0.78	0.86	0.58
Cys	TGT	NA	NA	0.24	0
	TGC	NA	NA	0.76	1.00
Gln	CAA	0.16	0.05	0.08	0.13
	CAG	0.84	0.95	0.92	0.87
Glu	GAA	0.10	0.13	0.15	0.08
	GAG	0.90	0.87	0.85	0.92
Gly	GGT	0.18	0.08	0.14	0.26
	GGC	0.57	0.68	0.56	0.50
	GGA	0.14	0.09	0.06	0.14
	GGG	0.11	0.15	0.23	0.10
His	CAT	0.18	0.30	0.32	0.16
	CAC	0.82	0.70	0.68	0.84
Ile	ATT	0.11	0.09	0.23	0.37
	ATC	0.88	0.88	0.76	0.60
	ATA	0.01	0.02	0.01	0.03
Leu	TTA	0.02	0.01	0	0.03
	TTG	0.10	0.05	0.08	0.14
	CTT	0.07	0.05	0.07	0.17

	CTC	0.19	0	0.09	0.23
	CTA	0.05	0.05	0.07	0.03
	CTG	0.57	0.84	0.70	0.38
Lys	AAA	0.25	0.12	0.20	0.27
	AAG	0.75	0.88	0.80	0.73
Met	ATG	1.00	1.00	1.00	1.00
Phe	TTT	0.35	0.19	0.38	0.23
	TTC	0.65	0.81	0.62	0.77
Pro	CCT	0.26	0.24	0.18	0.21
	CCC	0.35	0.42	0.33	0.21
	CCA	0.09	0.15	0.12	0.14
	CCG	0.30	0.18	0.37	0.43
Ser	TCT	0.42	0.11	0.11	0.67
	TCC	0.14	0.49	0.23	0.20
	TCA	0.02	0.05	0.01	0.07
	TCG	0.18	0.12	0.23	0.03
	AGT	0.08	0.02	0.03	0
	AGC	0.14	0.20	0.39	0.03
Thr	ACT	0.23	0.10	0.29	0.23
	ACC	0.59	0.74	0.48	0.45
	ACA	0.03	0.05	0.08	0.17
	ACG	0.15	0.12	0.16	0.15
Trp	TGG	NA	1.00	NA	1.00
Tyr	TAT	0.31	0.16	0.14	0.31
	TAC	0.69	0.84	0.86	0.69
Val	GTT	0.07	0.07	0	0.14
	GTC	0.30	0.33	0.12	0.28
	GTA	0.11	0.03	0.11	0.07
	GTG	0.51	0.57	0.78	0.52

Table A.6: Catalogue of annotated human core histone variant genes, transcripts, and encoded proteins. Data obtained from NCBI RefSeq (*O'Leary et al.*, 2016) on 2017-03-13.

Type	Gene name	Gene UID	Transcript accession	Protein accession
H2A	H2AFB1	474382	NM_001017990	NP_001017990
H2A	H2AFB2	474381	NM_001017991	NP_001017991
H2A	H2AFB3	83740	NM_080720	NP_542451
H2A	H2AFJ	55766	NM_177925	NP_808760
H2A	H2AFJ	55766	NR_027716	non-coding
H2A	H2AFV	94239	NM_012412	NP_036544
H2A	H2AFV	94239	NM_138635	NP_619541
H2A	H2AFV	94239	NM_201436	NP_958844
H2A	H2AFV	94239	NM_201516	NP_958924
H2A	H2AFV	94239	NM_201517	NP_958925
H2A	H2AFVP1	654500	pseudogene	pseudogene
H2A	H2AFX	3014	NM_002105	NP_002096
H2A	H2AFY	9555	NM_001040158	NP_001035248
H2A	H2AFY	9555	NM_004893	NP_004884
H2A	H2AFY	9555	NM_138609	NP_613075
H2A	H2AFY	9555	NM_138610	NP_613258
H2A	H2AFY2	55506	NM_018649	NP_061119
H2A	H2AFZ	3015	NM_002106	NP_002097
H2A	H2AFZP1	54049	pseudogene	pseudogene
H2A	H2AFZP2	346990	pseudogene	pseudogene
H2A	H2AFZP3	728023	pseudogene	pseudogene
H2A	H2AFZP4	100462795	pseudogene	pseudogene
H2A	H2AFZP5	100288330	pseudogene	pseudogene
H2A	H2AFZP6	100462800	pseudogene	pseudogene
H2B	H2BFM	286436	NM_001164416	NP_001157888
H2B	H2BFS	54145	pseudogene	pseudogene
H2B	H2BFWT	158983	NM_001002916	NP_001002916
H2B	H2BFXP	767811	pseudogene	pseudogene
H3	CENPA	1058	NM_001042426	NP_001035891
H3	CENPA	1058	NM_001809	NP_001800
H3	H3F3A	3020	NM_002107	NP_002098

H3	H3F3AP1	654505	pseudogene	pseudogene
H3	H3F3AP2	664611	pseudogene	pseudogene
H3	H3F3AP3	100689229	pseudogene	pseudogene
H3	H3F3AP4	440926	pseudogene	pseudogene
H3	H3F3AP5	347376	pseudogene	pseudogene
H3	H3F3AP6	644914	pseudogene	pseudogene
H3	H3F3B	3021	NM_005324	NP_005315
H3	H3F3BP1	100287087	pseudogene	pseudogene
H3	H3F3BP2	100420410	pseudogene	pseudogene
H3	H3F3C	440093	NM_001013699	NP_001013721

Table A.7: Variations in canonical core histone gene annotations compared to expectation of single exon, single transcript, and stem-loop. Data obtained from NCBI RefSeq (*O'Leary et al.*, 2016) on 2017-03-13.

- HIST1H2BD has 2 transcripts.
- HIST1H2BD has 2 exons on transcript NM_138720.
- Gene HIST1H2BD has no annotated stem-loop on transcript NM_138720.
- HIST1H2BD has possible stem loop in genomic sequence starting at 38 bp from the end of protein XP_005249096.1 CDS
- HIST1H2BD has possible stem loop in genomic sequence starting at 38 bp from the end of protein NP_066407.1 CDS
- HIST1H2BD has possible stem loop in genomic sequence starting at 38 bp from the end of protein NP_619790.1 CDS
- HIST1H2BK has 2 transcripts.
- HIST1H2BK has 2 exons on transcript NM_080593.
- Gene HIST1H2BK has no annotated stem-loop on transcript NM_080593.
- HIST1H2BK has possible stem loop in genomic sequence starting at 38 bp from the end of protein NP_542160.1 CDS
- HIST1H2BK has possible stem loop in genomic sequence starting at 38 bp from the end of protein NP_001299582.1 CDS
- HIST2H2BF has 2 transcripts.
- HIST2H2BF has 2 exons on transcript NM_001161334.
- Gene HIST2H2BF has no annotated stem-loop on transcript NM_001161334.
- HIST2H2BF has possible stem loop in genomic sequence starting at 44 bp from the end of protein NP_001019770.1 CDS
- HIST1H3D has 2 exons on transcript NM_003530.

Appendix B

Mouse Canonical Core Histone Catalogue

Table B.1: Mouse catalogue build information

Data accession date

13th March 2017

Build date

21st August 2017

BioPerl version

1.006924

Bio-EUtilities version

1.75

T-Coffee version

11.00.8

WebLogo version

3.5.0

Table B.2: Count of histone mouse genes.

	H2A	H2B	H3	H4	Total
HIST1	13 + 2ψ	15 + 0ψ	9 + 0ψ	11 + 0ψ	48 + 2ψ
HIST2	4 + 0ψ	2 + 0ψ	3 + 0ψ	1 + 0ψ	10 + 0ψ
HIST3	1 + 0ψ	1 + 1ψ	0 + 0ψ	0 + 0ψ	2 + 1ψ
HIST4	0 + 0ψ	0 + 0ψ	0 + 0ψ	1 + 0ψ	1 + 0ψ
Total	18 + 2ψ	18 + 1ψ	12 + 0ψ	13 + 0ψ	61 + 3ψ

Table B.3: Catalogue of annotated mouse canonical core histone genes, transcripts, and encoded proteins. Data obtained from NCBI RefSeq (*O'Leary et al.*, 2016) on 2017-03-13.

Type	Gene name	Gene UID	Transcript accession	Protein accession
H2A	Hist1h2aa	319163	NM_175658	NP_783589
H2A	Hist1h2ab	319172	NM_175660	NP_783591
H2A	Hist1h2ac	319164	NM_178189	NP_835496
H2A	Hist1h2ad	319165	NM_178188	NP_835495
H2A	Hist1h2ae	319166	NM_178187	NP_835494
H2A	Hist1h2af	319173	NM_175661	NP_783592
H2A	Hist1h2ag	319167	NM_178186	NP_835493
H2A	Hist1h2ah	319168	NM_175659	NP_783590
H2A	Hist1h2ai	319191	NM_178182	NP_835489
H2A	Hist1h2aj	319174	pseudogene	pseudogene
H2A	Hist1h2ak	319169	NM_178183	NP_835490
H2A	Hist1h2al	667728	pseudogene	pseudogene
H2A	Hist1h2an	319170	NM_178184	NP_835491
H2A	Hist1h2ao	665433	NM_001177544	NP_001171015
H2A	Hist1h2ap	319171	NM_178185	NP_835492
H2A	Hist2h2aa1	15267	NM_013549	NP_038577
H2A	Hist2h2aa2	319192	NM_178212	NP_835584
H2A	Hist2h2ab	621893	NM_178213	NP_835585
H2A	Hist2h2ac	319176	NM_175662	NP_783593
H2A	Hist3h2a	319162	NM_178218	NP_835736
H2B	Hist1h2ba	319177	NM_175663	NP_783594
H2B	Hist1h2bb	319178	NM_175664	NP_783595
H2B	Hist1h2bc	68024	NM_001290380	NP_001277309
H2B	Hist1h2bc	68024	NM_023422	NP_075911
H2B	Hist1h2be	319179	NM_001177653	NP_001171124
H2B	Hist1h2be	319179	NM_001290530	NP_001277459
H2B	Hist1h2be	319179	NM_178194	NP_835501
H2B	Hist1h2bf	319180	NM_178195	NP_835502
H2B	Hist1h2bg	319181	NM_178196	NP_835503
H2B	Hist1h2bh	319182	NM_178197	NP_835504
H2B	Hist1h2bj	319183	NM_178198	NP_835505

H2B	Hist1h2bk	319184	NM_175665	NP_783596
H2B	Hist1h2bl	319185	NM_178199	NP_835506
H2B	Hist1h2bm	319186	NM_178200	NP_835507
H2B	Hist1h2bn	319187	NM_178201	NP_835508
H2B	Hist1h2bp	319188	NM_001290466	NP_001277395
H2B	Hist1h2bp	319188	NM_178202	NP_835509
H2B	Hist1h2bq	665596	NM_001097979	NP_001091448
H2B	Hist1h2bq	665596	NM_001313880	NP_001300809
H2B	Hist1h2br	665622	NM_001110555	NP_001104025
H2B	Hist1h2br	665622	NM_001313878	NP_001300807
H2B	Hist2h2bb	319189	NM_175666	NP_783597
H2B	Hist2h2be	319190	NM_178214	NP_835586
H2B	Hist3h2ba	78303	NM_030082	NP_084358
H2B	Hist3h2bb-ps	382522	pseudogene	pseudogene
H3	Hist1h3a	360198	NM_013550	NP_038578
H3	Hist1h3b	319150	NM_178203	NP_835510
H3	Hist1h3c	319148	NM_175653	NP_783584
H3	Hist1h3d	319149	NM_178204	NP_835511
H3	Hist1h3e	319151	NM_178205	NP_835512
H3	Hist1h3f	260423	NM_013548	NP_038576
H3	Hist1h3g	97908	NM_145073	NP_659539
H3	Hist1h3h	319152	NM_178206	NP_835513
H3	Hist1h3i	319153	NM_178207	NP_835514
H3	Hist2h3b	319154	NM_178215	NP_835587
H3	Hist2h3c1	15077	NM_178216	NP_835734
H3	Hist2h3c2	97114	NM_054045	NP_473386
H4	Hist1h4a	326619	NM_178192	NP_835499
H4	Hist1h4b	326620	NM_178193	NP_835500
H4	Hist1h4c	319155	NM_178208	NP_835515
H4	Hist1h4d	319156	NM_175654	NP_783585
H4	Hist1h4f	319157	NM_175655	NP_783586
H4	Hist1h4h	69386	NM_153173	NP_694813
H4	Hist1h4i	319158	NM_175656	NP_783587
H4	Hist1h4j	319159	NM_178210	NP_835582
H4	Hist1h4k	319160	NM_178211	NP_835583
H4	Hist1h4m	100041230	NM_001195421	NP_001182350

H4	Hist1h4n	319161	NM_175657	NP_783588
H4	Hist2h4	97122	NM_033596	NP_291074
H4	Hist4h4	320332	NM_175652	NP_783583

Table B.4: Canonical mouse H2A encoded protein isoforms.

Most common isoform (129 amino acids; Hist1h2a –b, –c, –d, –e, –g, –i, –n, –o, –p) SGRGKQGGKARAKAKTRSSRAGLQFPVGRVHRLLRKGNYSERVGAGAPVYLAAVLEYLTAEILELAGNAARDNKKTRIIPRHLQLAIRNDEELNKLLGRVTIAQGGVLPNIQAVLLPKKTESHHKAKGK	
Hist1h2aa	R3_G4PT Q6R A14V T16S K36Q S40_E41AQ V43I I62V I79T H124del A126_G128SQT
Hist1h2af	A126P
Hist1h2ah	G128_K129del
Hist1h2ak	S122T
Hist2h2aa1	T16S S40A L51M R99K
Hist2h2aa2	T16S S40A L51M R99K
Hist2h2ab	T16S S40A L51M I87V R99G H124_A126KPG G128N
Hist2h2ac	T16S S40A L51M R99K H124del G128S
Hist3h2a	T16S

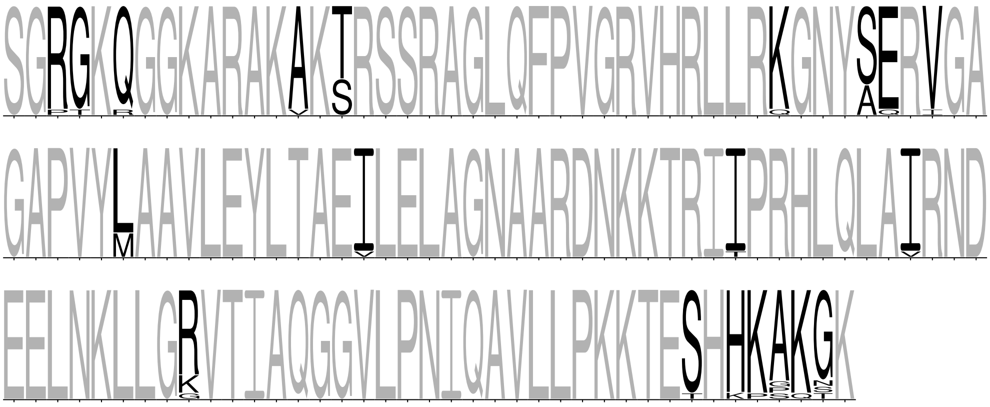


Table B.5: Canonical mouse H2B encoded protein isoforms.

Most common isoform (125 amino acids; Hist1h2b –c.1, –c.2, –e.1, –e.2, –e.3, –g)

PEPAKSAPAPKKGSKKAVTKAQKKDGKKRKRSRKESYSVYVYKVLKQVHPD
TGISSKAMGIMNSFVNDIFERIAGEASRLAHYNKRSTITSREIQTAVRLLL
PGELAKHAVSEGTKAVTKYTSSK

Hist1h2ba	P3V A4_K5insV S6G P8_P10TIS S14F A21T D25E K27R S32C V39I V41I G60S N67T G75S
Hist1h2bb	A4S V18_T19IS G75S
Hist1h2bf	G75S
Hist1h2bh	V18L
Hist1h2bj	G75S
Hist1h2bk	G75S S124A
Hist1h2bl	G75S
Hist1h2bm	A4T
Hist1h2bn	G75S
Hist1h2bp.1	A4V A7V G75S *126Iext*11
Hist1h2bp.2	A4V A7V G75S
Hist1h2bq.1	G75S *125Next*8
Hist1h2bq.2	G75S
Hist1h2br.1	G75S *125Next*8
Hist1h2br.2	G75S
Hist2h2bb	E2D A21V
Hist2h2be	P3L V39I G75N A97S S124A
Hist3h2ba	A4_K5SR A7T V18I S32G V39I G75S I94V

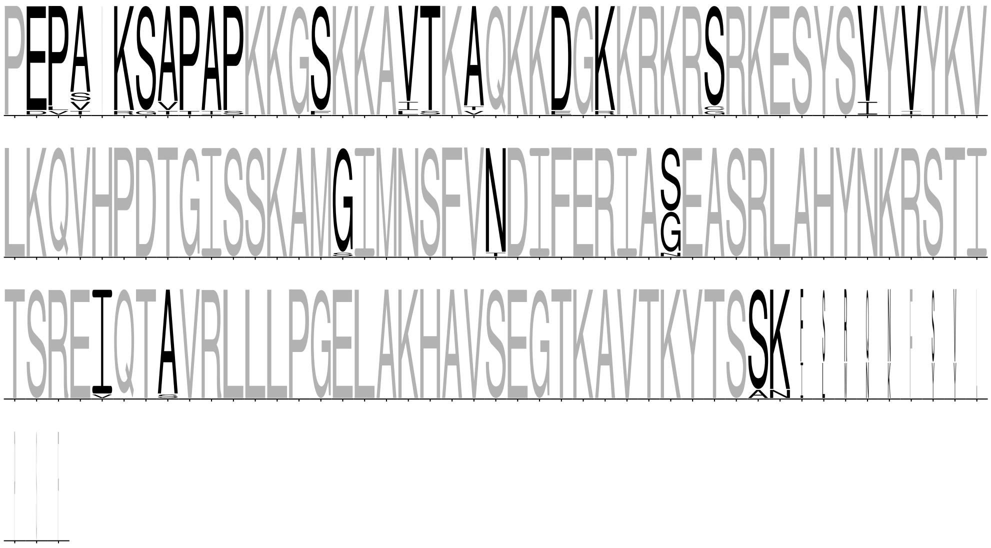


Table B.6: Canonical mouse H3 encoded protein isoforms.

Most common isoform (135 amino acids; Hist1h3 –b, –c, –d, –e, –f; Hist2h3 –b, –c1, –c2) ARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREI RRYQKSTELLIRKLPFQRLVREIAQDFKTDLRFQSSAVMALQEASEAYLVG LFEDTNLCAIHAKRVTIMPKDIQLARRIRGERA	
Hist1h3a	S96C
Hist1h3g	S96C
Hist1h3h	S96C
Hist1h3i	S96C

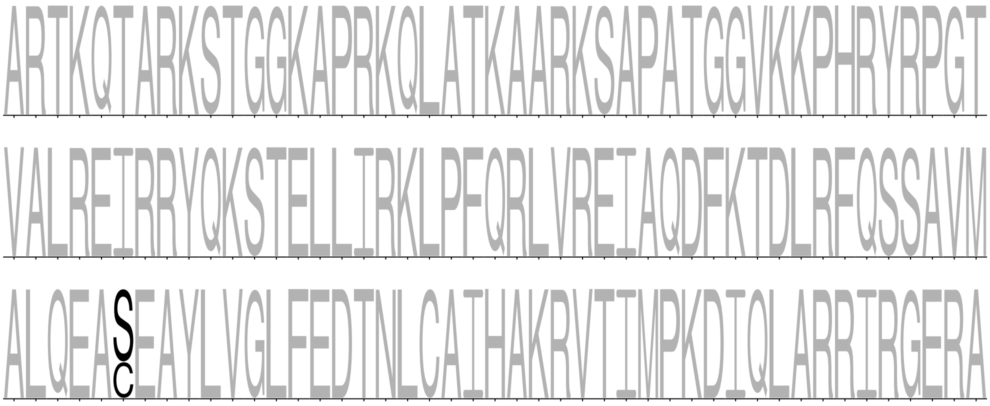

ARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGT
VALREIRRYQKSTELLIRKLPFQRLVREIAQDFKTDLRFQSSAVM
ALQEASEAYLVGLFEDTNLCAIHAKRVTIMPKDIQLARRIRGERA
C

Table B.7: Canonical mouse H4 encoded protein isoforms.

Most common isoform (102 amino acids; Hist1h4 –a, –b, –c, –d, –f, –h, –i, –j, –k, –m, –n; Hist2h4; Hist4h4)
SGRGKGGKGLGKGGAKRHRKVLRDNIQGITKPAIRRLARRGGVKRISGLIY EETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYALKRQGRTLYGFGG

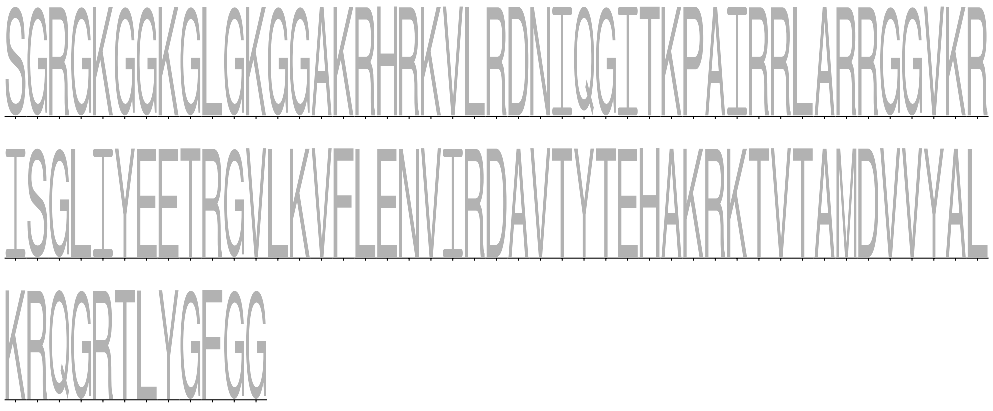


Table B.8: Catalogue of variant core histone genes and products.

Type	Gene name	Gene UID	Transcript accession	Protein accession
H2A	H2af-ps	100040209	pseudogene	pseudogene
H2A	H2afb1	68231	NM_026627	NP_080903
H2A	H2afb2	624153	NM_001281530	NP_001268459
H2A	H2afb3	624957	NM_001281531	NP_001268460
H2A	H2afj	232440	NM_177688	NP_808356
H2A	H2afv	77605	NM_001347064	NP_001333993
H2A	H2afv	77605	NM_029938	NP_084214
H2A	H2afx	15270	NM_010436	NP_034566
H2A	H2afy	26914	NM_001159513	NP_001152985
H2A	H2afy	26914	NM_001159514	NP_001152986
H2A	H2afy	26914	NM_001159515	NP_001152987
H2A	H2afy	26914	NM_012015	NP_036145
H2A	H2afy2	404634	NM_207000	NP_996883
H2A	H2afy3	67552	pseudogene	pseudogene
H2A	H2afz	51788	NM_001316995	NP_001303924
H2A	H2afz	51788	NM_016750	NP_058030
H2B	H2bfm	69389	NM_027067	NP_081343
H3	H3f3a	15078	NM_008210	NP_032236
H3	H3f3a-ps1	15079	pseudogene	pseudogene
H3	H3f3a-ps2	15080	pseudogene	pseudogene
H3	H3f3b	15081	NM_008211	NP_032237

Table B.9: Variations in canonical core histone gene annotations compared to expectation of single exon, single transcript, and stem-loop. Data obtained from NCBI RefSeq (*O'Leary et al.*, 2016) on 2017-03-13.

- Hist1h2an has unmatched stem-loop sequence GGCTCTTTTCAGAGCT on transcript NM_178184.
- Hist1h2bc has 2 transcripts.
- Hist1h2bc has 2 exons on transcript NM_023422.
- Gene Hist1h2bc has no annotated stem-loop on transcript NM_023422.
- Hist1h2bc has possible stem loop in genomic sequence starting at 36 bp from the end of protein NP_001277309.1 CDS
- Hist1h2bc has possible stem loop in genomic sequence starting at 36 bp from the end of protein NP_075911.2 CDS
- Hist1h2be has 3 transcripts.
- Hist1h2be has 2 exons on transcript NM_178194.
- Gene Hist1h2be has no annotated stem-loop on transcript NM_178194.
- Hist1h2be has possible stem loop starting at position 630 of NM_178194
- Hist1h2be has 2 exons on transcript NM_001177653.
- Gene Hist1h2be has no annotated stem-loop on transcript NM_001177653.
- Hist1h2be has possible stem loop starting at position 604 of NM_001177653
- Hist1h2bp has 2 transcripts.
- Hist1h2bp has 2 exons on transcript NM_178202.
- Gene Hist1h2bp has no annotated stem-loop on transcript NM_178202.

- Hist1h2bp has possible stem loop in genomic sequence starting at 29 bp from the end of protein XP_006516753.1 CDS
- Hist1h2bp has possible stem loop in genomic sequence starting at 29 bp from the end of protein NP_001277395.1 CDS
- Hist1h2bq has 2 transcripts.
- Hist1h2bq has 2 exons on transcript NM_001097979.
- Gene Hist1h2bq has no annotated stem-loop on transcript NM_001097979.
- Hist1h2bq has possible stem loop in genomic sequence starting at 31 bp from the end of protein NP_001300809.1 CDS
- Hist1h2br has 2 transcripts.
- Hist1h2br has 2 exons on transcript NM_001110555.
- Gene Hist1h2br has no annotated stem-loop on transcript NM_001110555.
- Hist1h2br has possible stem loop in genomic sequence starting at 31 bp from the end of protein NP_001300807.1 CDS
- Hist2h2be has unmatched stem-loop sequence GGCCC-CTTTGATCTTT on transcript NM_178214.
- Hist3h2ba has stem-loop 19 bp long on transcripts NM_030082.
- Hist3h2ba has unmatched stem-loop sequence GGCTCTTTTCA-GAGCCACC on transcript NM_030082.
- Hist1h3d has stem-loop 19 bp long on transcripts NM_178204.
- Hist1h3d has unmatched stem-loop sequence GGCTCTTTTCA-GAGCCACC on transcript NM_178204.
- Hist1h4b has unmatched stem-loop sequence GGTC-CTTTTCAGGACC on transcript NM_178193.
- Hist1h4j has unmatched stem-loop sequence GGTC-CTTTTCAGGACC on transcript NM_178210.

- Hist4h4 has stem-loop 17 bp long on transcripts NM_175652.
- Hist4h4 has unmatched stem-loop sequence GGCC-CTTTTCAGGGCCC on transcript NM_175652.

Figure B.1: Sequence logo for the mouse canonical H2A gene coding regions listed in table Table B.3. Initiator codon ATG and stop codon are omitted.

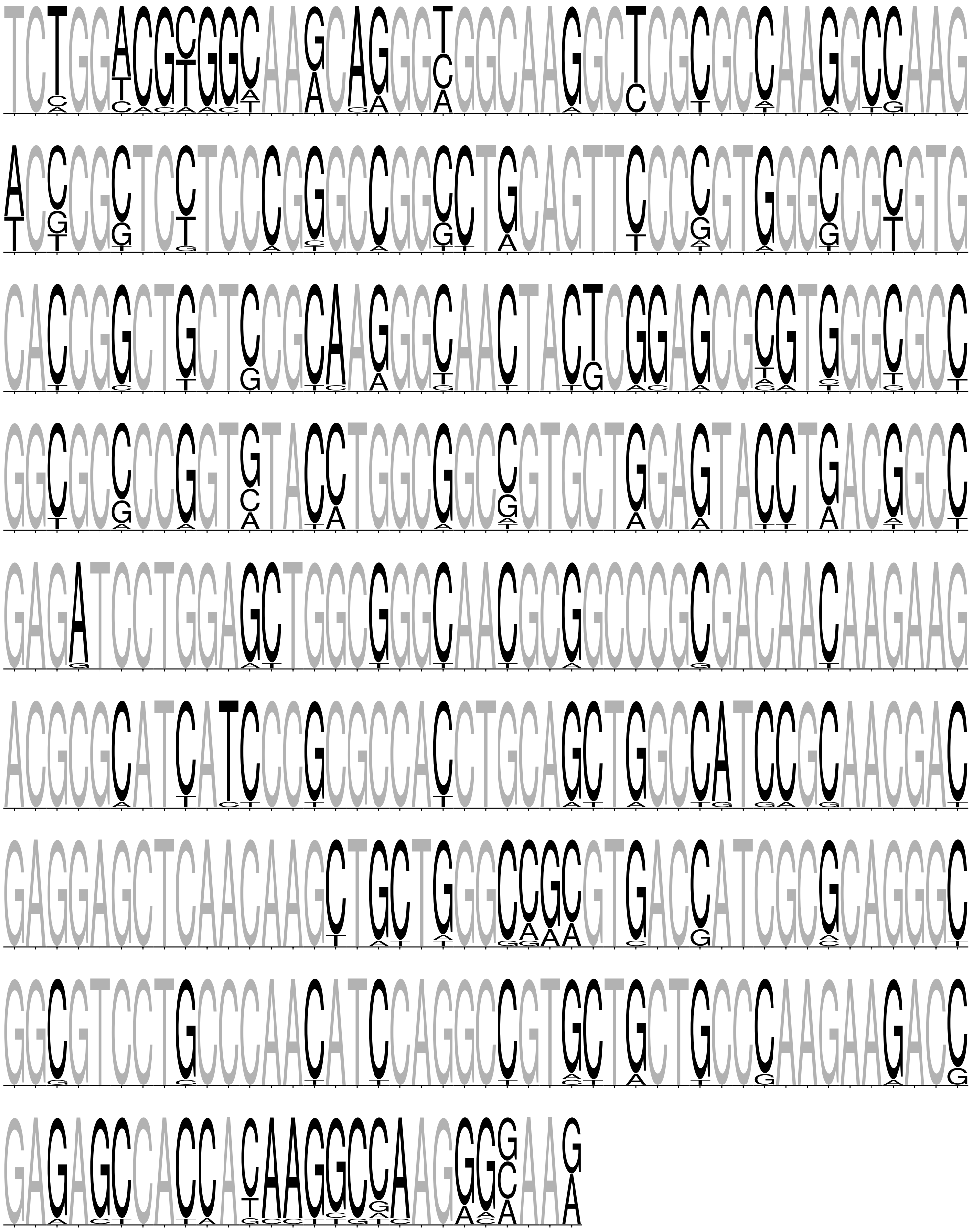


Figure B.2: Sequence logo for the mouse canonical H2B gene coding regions listed in table Table A.2. Initiator codon ATG and stop codon are omitted.

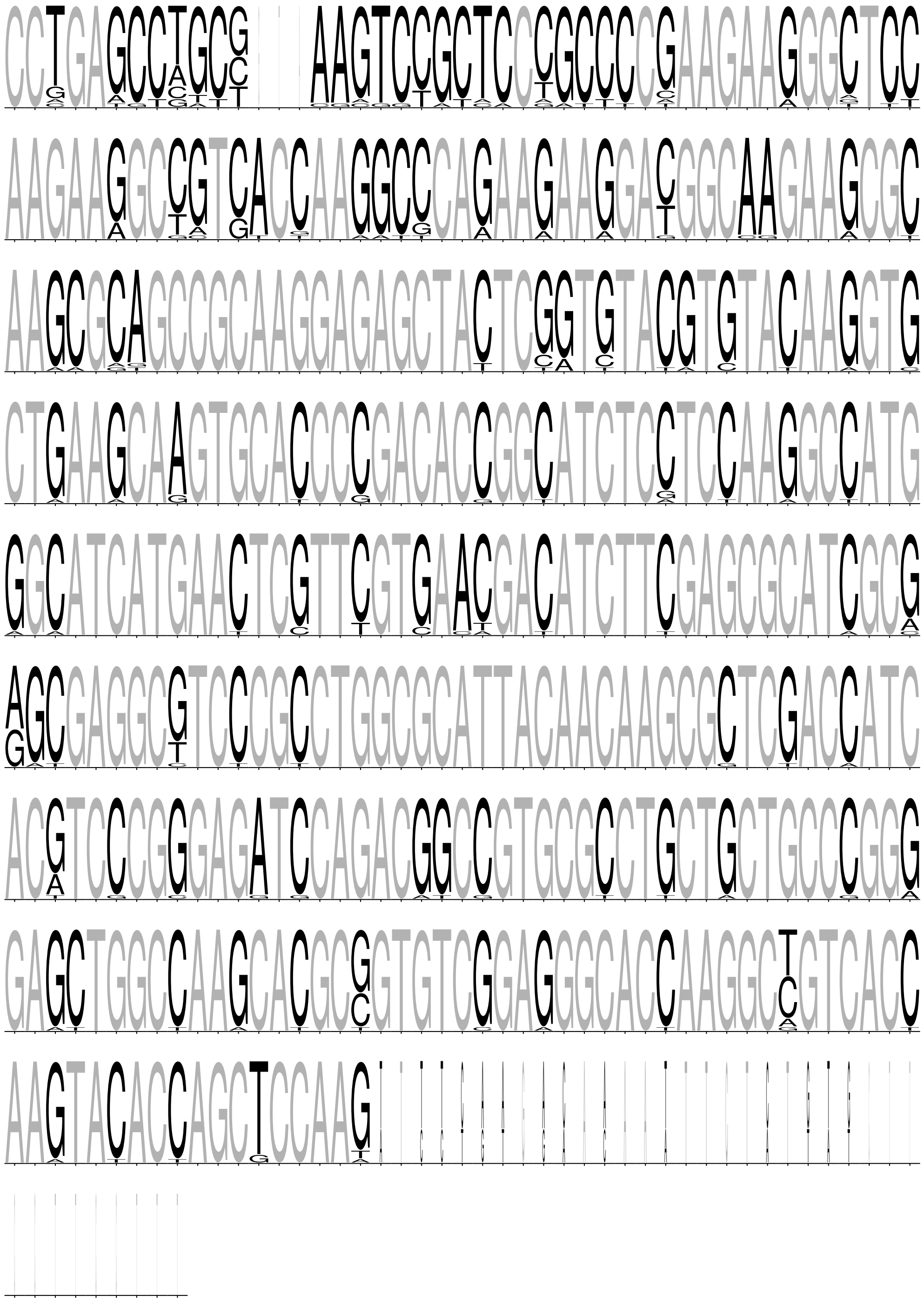


Figure B.3: Sequence logo for the mouse canonical H3 gene coding regions listed in table Table B.3. Initiator codon ATG and stop codon are omitted.

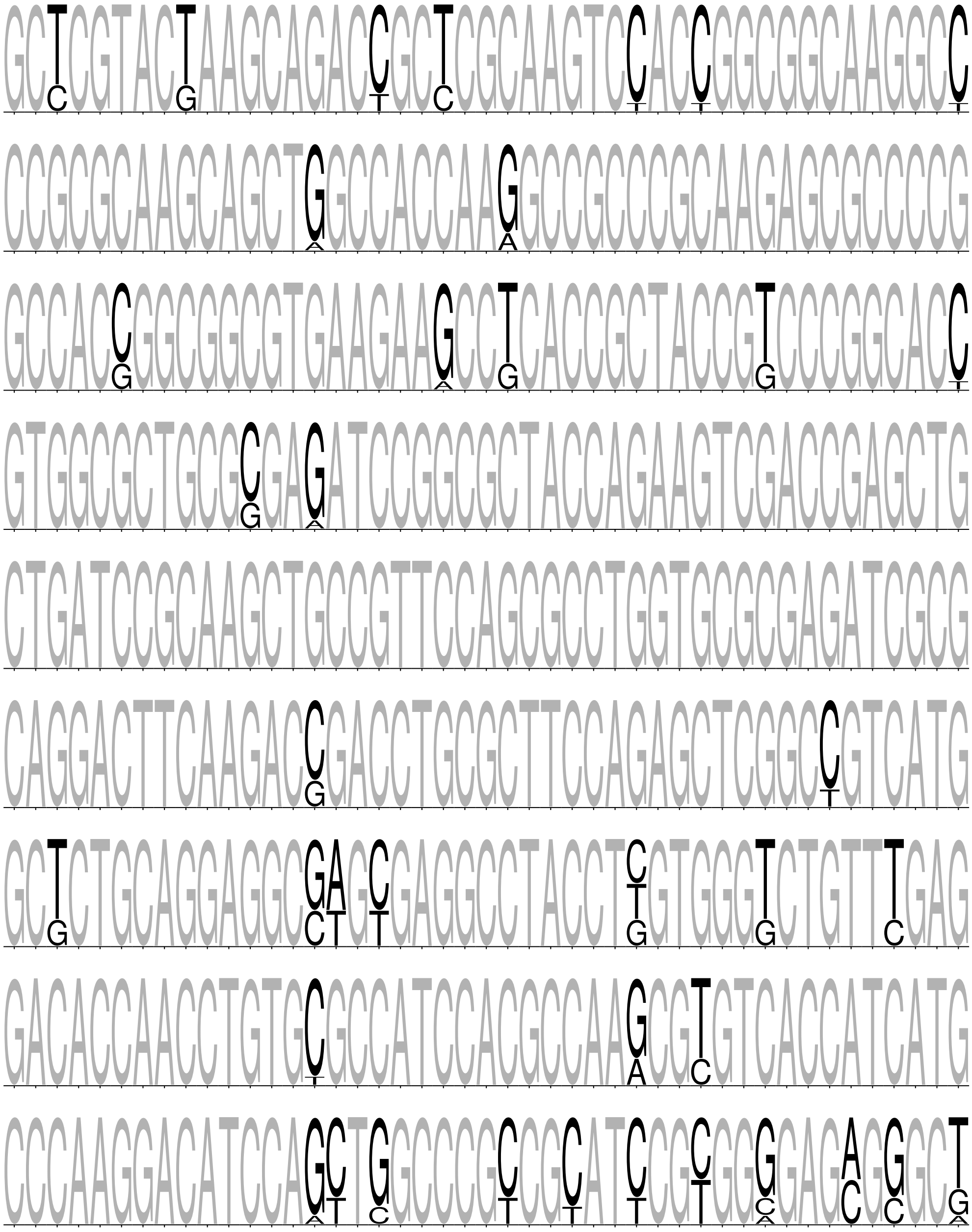


Figure B.4: Sequence logo for the mouse canonical H4 gene coding regions listed in table Table B.3. Initiator codon ATG and stop codon are omitted.

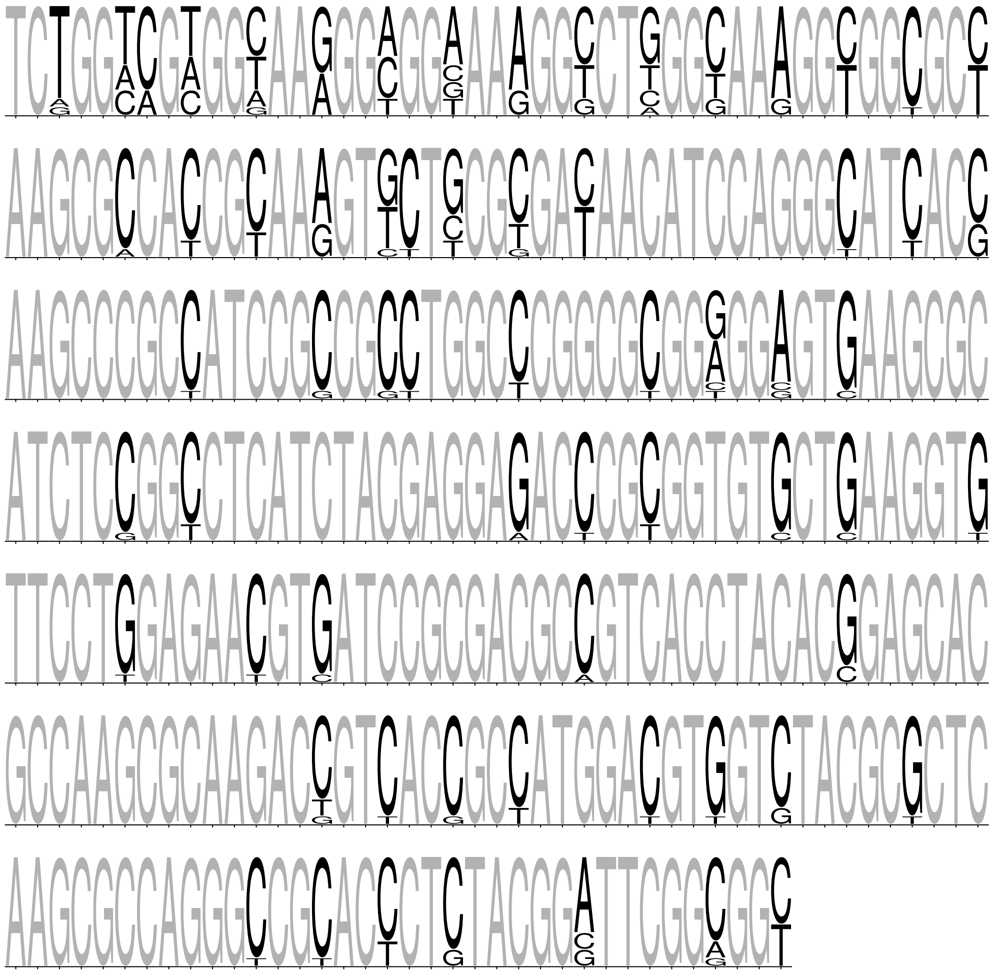


Table B.10: Mean synonymous and non-synonymous distances between pairs of canonical core histone genes. Mean over all pairwise comparisons of non-synonymous (d_N) and synonymous (d_S) nucleotide substitutions and mean of d_N/d_S ratios for canonical core histone coding regions computed using *Goldman and Yang* (1994) model implemented in `codeml` from the PAML package (*Padavattana et al.*, 2007).

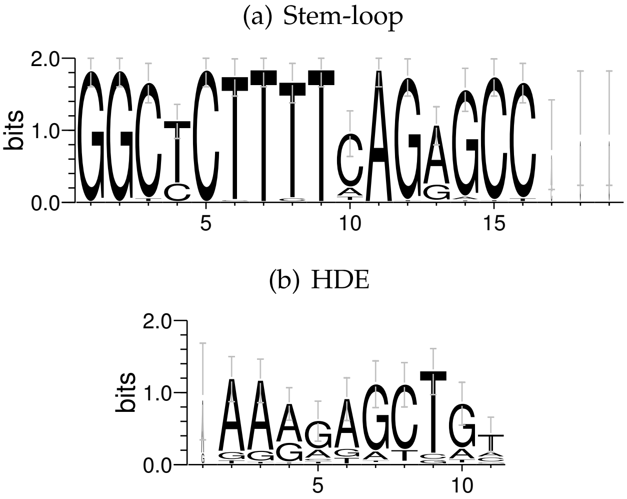
	d_N	d_S	d_N/d_S
H2A	0.0124	1.5788	0.00964
H2B	0.0129	0.6678	1.49093
H3	0.0016	0.4883	0.00602
H4	0.0008	0.7833	0.001

Table B.11: Codon usage frequency for each amino acid and canonical core histone type.

		H2A	H2B	H3	H4
Ala	GCT	0.08	0.16	0.18	0.13
	GCC	0.61	0.52	0.64	0.73
	GCA	0.03	0.02	0.00	0.01
	GCG	0.29	0.30	0.18	0.13
Arg	CGT	0.09	0.02	0.19	0.09
	CGC	0.72	0.84	0.69	0.80
	CGA	0.01	0	0	0.01
	CGG	0.17	0.13	0.09	0.09
	AGA	0.00	0.02	0	0.02
	AGG	0.01	0	0.03	0
Asn	AAT	0.04	0.09	0	0.04
	AAC	0.96	0.91	1.00	0.96
Asp	GAT	0.03	0.12	0	0.18
	GAC	0.97	0.88	1.00	0.82
Cys	TGT	NA	0	0.31	NA
	TGC	NA	1.00	0.69	NA
Gln	CAA	0.07	0.36	0.01	0
	CAG	0.93	0.64	0.99	1.00
Glu	GAA	0.03	0.02	0.01	0.02
	GAG	0.97	0.98	0.99	0.98
Gly	GGT	0.10	0.03	0.11	0.24
	GGC	0.73	0.80	0.74	0.48
	GGA	0.09	0.03	0.01	0.20
	GGG	0.08	0.15	0.14	0.08
His	CAT	0.10	0.36	0	0.08
	CAC	0.90	0.64	1.00	0.92
Ile	ATT	0.04	0.01	0.04	0.03
	ATC	0.96	0.98	0.96	0.97
	ATA	0	0.01	0	0
Leu	TTA	0.01	0	0	0
	TTG	0.02	0.01	0.02	0.02
	CTT	0.01	0.01	0.03	0.06
	CTC	0.12	0	0.05	0.41

	CTA	0.05	0.01	0.01	0.01
	CTG	0.80	0.97	0.90	0.50
Lys	AAA	0.10	0.05	0.04	0.24
	AAG	0.90	0.95	0.96	0.76
Met	ATG	1.00	1.00	1.00	1.00
Phe	TTT	0.17	0.13	0.19	0
	TTC	0.83	0.87	0.81	1.00
Pro	CCT	0.04	0.27	0.12	0
	CCC	0.51	0.48	0.33	1.00
	CCA	0.02	0.06	0	0
	CCG	0.43	0.19	0.54	0
Ser	TCT	0.24	0.04	0.01	0.42
	TCC	0.36	0.46	0.16	0.46
	TCA	0.01	0.01	0	0.04
	TCG	0.20	0.26	0.35	0.08
	AGT	0.01	0	0	0
	AGC	0.18	0.24	0.47	0
Thr	ACT	0.03	0.03	0.11	0.07
	ACC	0.49	0.70	0.82	0.75
	ACA	0.02	0.04	0	0
	ACG	0.45	0.22	0.07	0.19
Trp	TGG	NA	1.00	NA	NA
Tyr	TAT	0.06	0.06	0	0
	TAC	0.94	0.94	1.00	1.00
Val	GTT	0	0.02	0	0.07
	GTC	0.18	0.21	0.33	0.34
	GTA	0.03	0.00	0	0
	GTG	0.78	0.77	0.67	0.59

Figure B.5: Sequence logos for (a) annotated stem-loops and (b) Histone Downstream Elements (HDEs) identified by homology for all canonical core histone gene 3′ untranslated regions (UTRs).


(a) Stem-loop

(b) HDE

Table B.12: Changes in mouse canonical core histone gene catalogue annotated in NCBI RefSeq obtained on 2017-03-13 compared to *Marzluff et al.* (2002).

Removed genes: Hist2h3ca1, Hist2h3ca2, Hist3h2bb
New genes: Hist1h2al, Hist1h2ap, Hist1h2bl, Hist1h2bq, Hist1h2br, Hist1h4n, Hist2h3c1, Hist2h3c2, Hist3h2bb-ps
Changed sequences: Hist1h2ak, Hist2h2ab

APPENDIX **C**

List of primers

AFG112

`TGTCCGAGGGTACTAAGGCCGTCACCAAGTACACCAGCGCT`

HIST1H2BJ V118I correction site-directed mutagenesis (forward)

AFG113

`AGCGCTGGTGTACTTGGTGACGGCCTTAGTACCCTCGGACA`

HIST1H2BJ V118I correction site-directed mutagenesis (reverse)

AFG114

`GACTAAGGCGCAGAAGAAAGACGGCAAGAAGCGCAAGCGCA`

HIST1H2BJ D25G correction site-directed mutagenesis (forward)

AFG115

`TGCGCTTGCGCTTCTTGCCGTCTTTCTTCTGCGCCTTAGTC`

HIST1H2BJ D25G correction site-directed mutagenesis (reverse)

AFG116

`CGCGGTACCGCCACCATGTCTGGTCGCGGCAAACA`

KpnI + HIST1H2AB (forward)

AFG118

`GCGGATCCCTTTCCCTTGGCCTTATGATGG`

HIST1H2AB + BamHI (reverse)

AFG119

`GTGACTAAGGCGCAGAAGAAAGA`

HIST1H2BJ D25G screening (forward) for use with AFG113

AFG120

`AGCGCTGGTGTACTTGGTGAC`

HIST1H2BJ V118I screening (reverse) for use with AFG114

AFG121

`CGCGGTACCGCCACCATGGCTGGCGGTAAGGCTG`

KpnI + H2AFZ (forward)

AFG122

`GCGGATCCGACAGTCTTCTGTTGTCCTTTCTTC`

H2AFZ + BamHI (reverse)

AFG123

`CCGCGGCGGCGTGAAGCACATCTCCGGCCTCATCTACGAG`

HIST1H4J R45H (CGC to CAC) site-directed mutagenesis (forward)

AFG124

`CTCGTAGATGAGGCCGGAGATGTGCTTCACGCCGCCGCGG`

HIST1H4J R45H (CGC to CAC) site-directed mutagenesis (reverse)

AFG130

`CGCGGTACCGCCACCATGTCGGGCCGCGGCAA`

KpnI + H2AFX (forward)

AFG131

`GCGGATCCGTACTCCTGGGAGGCCTG`

H2AFX + BamHI (reverse)

AFG132

`GCGGATCCGTACTCCTGGGCGGCCTGGGTGGCCTTCTT`

H2AFX S139A (TCC to GCC) + BamHI (reverse)

AFG133

`GCGGATCCGTACTCCTGGTCGGCCTGGGTGGCCTTCTT`

H2AFX S139D (TCC to GAC) + BamHI (reverse)

AFG134

`GCGGATCCGTACTCCTGCTCGGCCTGGGTGGCCTTCTT`

H2AFX S139E (TTC to GAG) + BamHI (reverse)

AFG151

`CTCACCGTTACCGCCCGGGCGAGGTGGCTCTGCGCGAGATCCG`

HIST1H3B T45E (GAG) site-directed mutagenesis (forward)

AFG152

`CGGATCTCGCGCAGAGCCACCTCGCCCGGGCGGTAACGGTGAG`

HIST1H3B T45E (GAG) site-directed mutagenesis (reverse)

AFG153

`CTCACCGTTACCGCCCGGGCGCCGTGGCTCTGCGCGAGATCCG`

HIST1H3B T45A (GCC) site-directed mutagenesis (forward)

AFG154

`CGGATCTCGCGCAGAGCCACGGCGCCCGGGCGGTAACGGTGAG`

HIST1H3B T45A (GCC) site-directed mutagenesis (reverse)

AFG155

`CGCGGTACCGCCACCATGTCTGGACGAGGGAAGCAG`

KpnI + HIST1H2AA (forward)

AFG156

`GCGGATCCCTTGCTTTGGGCTTTATGGTGG`

HIST1H2AA + BamHI (reverse)

AFG157

`CGCGGTACCGCCACCATGTCAGGACGCGGAAAGCAG`

KpnI + HIST2H2AB (forward)

AFG158

`GCGGATCCCTTGTTCTTGCCAGGCTTGTG`

HIST2H2AB + BamHI (reverse)

AFG395

`CAGGGGGAGGTGTGGGAGG`

GFP sequencing primer for pBOS. Binds to the 3′ UTR on the pBOS plasmid

AFG396

`GCCATCATAAGGCCAAGGGAAAGGATCCACCGGTCGCCACC`

HIST1H2AB frameshift correction site-directed mutagenesis (forward)

AFG397

GGTGGCGACCGGTGGATCCTTTCCCTTGGCCTTATGATGGC

HIST1H2AB frameshift correction site-directed mutagenesis (reverse)

AFG398

GAAGAAAGGACAACAGAAGACTGTGGATCCACCGGTCGCCACC

H2AFZ frameshift correction site-directed mutagenesis (forward)

AFG399

GGTGGCGACCGGTGGATCCACAGTCTTCTGTTGTCCTTTCTTC

H2AFZ frameshift correction site-directed mutagenesis (reverse)

AFG400

CAGGAGTACGATCCACCGGTCGCCACC

H2AFX frameshift correction site-directed mutagenesis (forward)

AFG401

GGTGGCGACCGGTGGATCGTACTCCTG

H2AFX frameshift correction site-directed mutagenesis (reverse)

AFG417

GCGCCCTGTCTCCCATCC

pMH3.2-614 vector amplification. Primer for sequence after the stop codon (forward)

AFG418

GGCGAAGACGGAAGACGCC

pMH3.2–614 vector amplification. Reverse complement for the sequence before the start codon (reverse)

AFG419

ATGCCAGAGCCAGCGAAGTC

H2B-EGFP amplification for blunt end cloning (forward)

AFG420

TTACTTGTACAGCTCGTCCATGCC

H2B-EGFP amplification for blunt end cloning (reverse)

AFG424

ATGGCTCGTACTAAACAGACAGCTC

H3-EYFP amplification for blunt end cloning (forward)

AFG457

`CGCGCGCATATGGTGAGCAAGGGCGAGGAG`

NdeI + EGFP for cloning from pBOS into pET (forward)

AFG458

`GCGGGATCCATTACTTGTACAGCTCGTCCATGCCG`

EGFP + BamHI for cloning from pBOS into pET (reverse)

AFG478

`CCGGATCCACGGGTCGCCACCATGGTGAGCAAGGGCGAG`

BamHI + PAGFP (forward)

AFG479

`CGCGGCCGCGGCCGCTTTACTTGTACAGCTCGTCCATG`

PAGFP + NotI (reverse)

APPENDIX D

bp_genbank_ref_extractor manual

D.1 NAME

bp_genbank_ref_extractor - Retrieves all related sequences for a list of searches on Entrez gene

D.2 VERSION

version 1.75

D.3 SYNOPSIS

bp_genbank_ref_extractor [options] [Entrez Gene Queries]

D.4 DESCRIPTION

This script searches on *Entrez Gene* database and retrieves not only the gene sequence but also the related transcript and protein sequences.

The gene UIDs of multiple searches are collected before attempting to retrieve them so each gene will only be analyzed once even if appearing as result on more than one search.

Note that *by default no sequences are saved* (see options and examples).

D.5 OPTIONS

Several options can be used to fine tune the script behaviour. It is possible to obtain extra base pairs upstream and downstream of the gene, control the naming of files and genome assembly to use.

See the section bugs for problems when using default values of options.

--assembly

When retrieving the sequence, a specific assemly can be defined. The value expected is a regex that will be case-insensitive. If it matches more than one assembly, it will use the first match. It defauls to `(primary|reference) assembly`.

--debug

If set, even more output will be printed that may help on debugging. Unlike the messages from **--verbose** and **--very-verbose**, these will not appear on the log file unless this option is selected. This option also sets **--very-verbose**.

--downstream, --down

Specifies the number of extra base pairs to be retrieved downstream of the gene. This extra base pairs will only affect the gene sequence, not the transcript or proteins.

--email

A valid email used to connect to the NCBI servers. This may be used by NCBI to contact users in case of problems and before blocking access in case of heavy usage.

--format

Specifies the format that the sequences will be saved. Defaults to *genbank* format. Valid formats are 'genbank' or 'fasta'.

--genes

Specifies the name for gene file. By default, they are not saved. If no value is given defaults to its UID. Possible values are 'uid', 'name', 'symbol' (the official symbol or nomenclature).

--help

Display the documentation (this text).

--limit

When making a query, limit the result to these first specific results. This is to prevent the use of specially unspecific queries and a warning will be given if a query returns more results than the limit. The default value is 200. Note that this limit is for *each* search.

--non-coding, --nonon-coding

Some protein coding genes have transcripts that are non-coding. By default, these sequences are saved as well. **--nonon-coding** can be used to ignore those transcripts.

--proteins

Specifies the name for proteins file. By default, they are not saved. If no value is given defaults to its accession. Possible values are 'accession', 'description', 'gene' (the corresponding gene ID) and 'transcript' (the corresponding transcript accesion).

Note that if not using 'accession' is possible for files to be overwritten. It is possible for the same gene to encode more than one protein or different proteins to have the same description.

--pseudo, --nopseudo

By default, sequences of pseudo genes will be saved. **--nopseudo** can be used to ignore those genes.

--save

Specifies the path for the directory where the sequence and log files will be saved. If the directory does not exist it will be created although the path to it must exist. Files on the directory may be rewritten if necessary. If unspecified, a directory named *extracted sequences* on the current directory will be used.

--save-data

This options saves the data (gene UIDs, description, product accessions, etc) to a file. As an optional value, the file format can be specified. Defaults to CSV.

Currently only CSV is supported.

Saving the data structure as a CSV file, requires the installation of the Text::CSV module.

--transcripts, --mrna

Specifies the name for transcripts file. By default, they are not saved. If no value is given defaults to its accession. Possible values are 'accession', 'description', 'gene' (the corresponding gene ID) and 'protein' (the protein the transcript encodes).

Note that if not using 'accession' is possible for files to be overwritten. It is possible for the same gene to have more than one transcript or different transcripts to have the same description. Also, non-coding transcripts will create problems if using 'protein'.

--upstream, --up

Specifies the number of extra base pairs to be extracted upstream of the gene. This extra base pairs will only affect the gene sequence, not the transcript or proteins.

--verbose, --v

If set, program becomes verbose. For an extremely verbose program, use **--very-verbose** instead.

--very-verbose, --vv

If set, program becomes extremely verbose. Setting this option, automatically sets **--verbose** as well. For help in debugging, consider using **--debug**

D.6 EXAMPLES

```
bp_genbank_ref_extractor \
  --transcripts=accession \
  '"homo sapiens"[organism] AND H2B'
```

Search Entrez Gene with the query `'"homo sapiens"[organism] AND H2B'` and save their transcripts sequences only. Note that default value of **--limit** may only extract some of the hits.

```
bp_genbank_ref_extractor \
  --transcripts=accession --proteins=accession \
   --format=fasta \
   ’"homo sapiens"[organism] AND H2B’ \
   ’"homo sapiens"[organism] AND MCPH1’
```

Save both transcript and protein sequences in the fasta format, for two queries, `’"homo sapiens"[organism] AND H2B’` and `’"homo sapiens"[organism] AND MCPH1’`.

```
bp_genbank_ref_extractor \
  --genes --down=500 --up=100 \
  ’"homo sapiens"[organism] AND H2B’
```

Download genomic sequences, including 500 bp downstream and 100 bp upstream of each gene.

```
bp_genbank_ref_extractor \
  --genes --asembly=’Alternate HuRef’ \
  ’"homo sapiens"[organism] AND H2B’
```

Download genomic sequences from the Alternate HuRef genome assembly.

```
bp_genbank_ref_extractor --save-data=CSV \
  ’"homo sapiens"[organism] AND H2B’
```

Do not save any sequence, only save the results in a CSV file.

```
bp_genbank_ref_extractor --save=’search-results’ \
  --genes=name  downstream=500 --upstream=200 \
  --nopseudo --nonnon-coding --transcripts --proteins \
  --format=fasta --save-data=CSV \
  ’"homo sapiens"[organism] AND H2B’ \
  ’"homo sapiens"[organism] AND MCPH1’
```

Ignoring non-coding and pseudo genes, downloads: genomic sequences with 500 and 200 bp downstream and upstream respectively, using the gene name as filename; transcript and proteins sequences using their accession number as filename; everything in fasta format plus a CSV file with search results; saved in a directory named *search-results*

D.7 NON-BUGS

- When supplying options, it's possible to not supply a value and use their default. However, when the expected value is a string, the next argument may be confused as value for the option. For example, when using the following command:

  ```
  bp_genbank_ref_extractor --transcripts \
    'H2A AND homo sapiens'
  ```

  we mean to search for 'H2A AND homo sapiens' saving only the transcripts and using the default as base for the filename. However, the search terms will be interpreted as the base for the filenames (but since it's not a valid identifier, it will return an error). To prevent this, you can either specify the values:

  ```
  bp_genbank_ref_extractor --transcripts='accession' \
    'H2A AND homo sapiens'
  ```

  or you can use the double hash to stop processing options. Note that this should only be used after the last option. All arguments supplied after the double dash will be interpreted as search terms

  ```
  bp_genbank_ref_extractor --transcripts \
    -- 'H2A AND homo sapiens'
  ```

D.8 NOTES ON USAGE

- Genes that are marked as 'live' and 'protein-coding' should have at least one transcript. However, This is not always true due to mistakes on annotation. Such cases will throw a warning. When faced with this, be nice and write to the entrez RefSeq maintainers http://www.ncbi.nlm.nih.gov/RefSeq/update.cgi.

- When creating the directories to save the files, if the directory already exists it will be used and no error or warning will be issued unless **--debug** as been set. If a non-directory file already exists with that name bp_genbank_ref_extractor exits with an error.

- On the subject of verbosity, all messages are saved on the log file. The options **--verbose** and **--very-verbose** only affect their printing to standard output. Debug messages are different as they will only show up (and be logged) if requested with **--debug**.
- When saving a file, to avoid problems with limited filesystems such as NTFS or FAT, only some characters are allowed. All other characters will be replaced by an underscore. Allowed characters are:

  a-z 0-9 - + . , () {} []'
- **bp_genbank_ref_extractor** tries to use the same file extensions that bioperl would expect when saving the file. If unable it will use the '.seq' extension.

Appendix E

Structure and function of histone H2AX

The following manuscript was authored by David Miguel Susano Pinto and Andrew Flaus as part of this thesis study. It was published in "Genome Stability and Human Diseases", Subcellular Biochemistry series, volume 50, 2010.

Abstract. Histone H2AX is a histone variant found in almost all eukaryotes. It makes a central contribution to genome stability through its role in the signaling of DNA damage events and by acting as a foundation for the assembly of repair foci. The H2AX protein sequence is highly similar and in some cases overlapping with replication-dependent canonical H2A, yet the H2AX gene and protein structures exhibit a number of features specific to the role of this histone in DNA repair. The most well known of these is a specific serine at the extreme C-terminus of H2AX which is phosphorylated by Phosphoinositide–3–Kinase-related protein Kinases (PIKKs) to generate the γH2AX mark. However, recent studies have demonstrated that phosphorylation, ubiquitylation and other post-translational modifications are also crucial for function. H2AX transcript properties suggest a capability to respond to damage events. Furthermore, the biochemical properties of H2AX protein within the nucleosome structure and its distribution within chromatin also point to features linked to its role in the DNA damage response. In particular, the theoretical inter-nucleosomal spacing of H2AX and the potential implications of amino acid residues distinguishing H2AX from canonical H2A in structure and dynamics are considered in detail. This review summarises current understanding of H2AX from a structure–function perspective.

E.1 Introduction

Maintenance of the genome stability is of great importance to all organisms because DNA damage can have serious biological implications including genetic disorders and cancer (*McKinnon and Caldecott*, 2007). One mechanism for maintaining genome stability is to increase DNA repair (*Lengauer et al.*, 1998) and an important paradigm for DNA repair is the mechanism for identifying and facilitating re-ligation of DNA Double Strand Breaks (DSBs). DSBs are one of the most serious forms of DNA damage because they involve loss of genetic continuity. They arise from a variety of causes including not only the action of DNA damaging agents but also normal functions such as meiosis and antibody class switching. An important player in the DNA Damage Response (DDR) for dealing with DSBs is the histone variant H2AX, which is an integral component of the chromatin packaging of eukaryotic genomes.

Chromatin Structure and Genome Stability

Eukaryotic DNA is not dispersed randomly within the cell nucleus. Instead, it is packaged into the chromatin structure which compacts DNA and organises accessibility to the genome. This chromatin packaging is hierarchical, based on the nucleosome as a fundamental building block. The canonical nucleosome comprises two copies of each core histone (H2A, H2B, H3 and H4) around which 147 bp of DNA is wound in a superhelical spiral (*Davey et al.*, 2002). The nucleosomes are connected by short DNA linkers to form repeating units which subsequently arrange in a number of higher-order structural levels up to condensed metaphase chromosomes.

Despite its modular structure, the arrangement of chromatin is not static and must be “remodeled” during nuclear processes including DNA repair. This remodeling is driven by two general mechanisms: Firstly, molecules such as ATP-dependent remodelers and chromatin-binding proteins can directly modify the structure. Secondly, physiochemical properties of chromatin can be modulated by post-translational modification or insertion of histone variants to alter its stability (*Felsenfeld and Groudine*, 2003; *Ausió*, 2006).

H2AX and DNA repair

H2AX and another histone variant, H2A.Z, were both identified in human cells by their different migration compared to canonical H2A isoforms on SDS and acetic acid–urea gels (*West and Bonner*, 1980). In this separation, H2AX and H2A.Z were two of four unidentified species arbitrarily labeled T, W, X and Z. Subsequently, it was found that T and W were the ubiquitylated forms of X and Z (*West and Bonner*, 1980). H2A.Y is an alternative name for macroH2A1. Although originally labeled as H2A.X, the internal period ('.') separating the X has fallen into disuse so the H2AX name is almost universally used in the DNA repair field. In contrast, the internal period has historically been retained in H2A.Z, whose major roles have been subsequently associated with transcription (*Ausió*, 2006).

A distinct function for H2AX was uncovered some 18 years after its initial identification when human and mouse serine 139 was observed to be rapidly phosphorylated in response to treatments that cause DSBs (*Rogakou et al.*, 1998). In structural nomenclature, the phosphorylation occurs on the serine oxygen in the gamma position so the modified form is widely referred to as γH2AX. This γH2AX phosphoprotein is found to be rapidly concentrated around DSBs in centers termed "foci" that can extend for a range of up to 2 Mbp away from the damage site (*Rogakou et al.*, 1999).

The amino acid region surrounding serine 139 matches the consensus recognition sequence for a set of PhosphoInositide–3–Kinase-related protein Kinases (PIKKs) known to be central in the DNA damage response from genetic studies in yeast (*Downs et al.*, 2000). The link between PIKKs and γH2AX formation has been directly demonstrated by biochemical inhibition using mutagenesis in yeast (*Downs et al.*, 2000) and wortmannin in higher cells (*Paull et al.*, 2000).

γH2AX is a widely recognised participant in DSB repair and is one of the earliest markers of damage (*Pilch et al.*, 2003). Other DNA repair-related proteins subsequently congregate at the γH2AX foci during the repair process. Although their recruitment to DSBs is not completely dependent on H2AX phosphorylation, H2AX is an important element in proper damage response foci formation by enhancing the retention of repair factors after their initial recruitment (*Celeste et al.*, 2003a). H2AX$^{-/-}$

mice have moderate defects including radiation sensitivity, growth retardation and immunodeficiency which are consistent with deficiencies in DNA repair (*Celeste et al.*, 2002, 2003b). Importantly, these phenotypes are only moderate and suggest redundancy for the role of H2AX. Nevertheless, karyotypes of H2AX-deleted genomes also reveal a high number of translocations and chromosome rearrangements directly demonstrating increased genomic instability.

E.2 Structural Properties of H2AX

Based on the linkage between the early H2AX phosphorylation event and the DDR, a large number of studies have focussed on γH2AX and its subsequent interactions with the repair mechanism. Less consideration has been given to the biochemical properties of H2AX itself.

H2AX is one of a set of histone H2A proteins encoded in eukaryotic genomes, the human genome holding 21 genes of H2A forms. The canonical human H2A has two biochemically separable isoforms, H2A.1 and H2A.2. No functional difference between those isoforms is known, and the basis of their distinction appears to be dependent on residue 51 encoding, respectively, either a leucine or methionine despite further heterogeneity within each isoform (*Bonenfant et al.*, 2006; *Marzluff et al.*, 2002). Four additional H2A variants with distinct functions are encoded in humans and other higher eukaryotes: H2AX, H2A.Z, macroH2A1, macroH2A2, H2A.F/Z and H2ABbd (*Marzluff et al.*, 2002).

Definition of H2AX

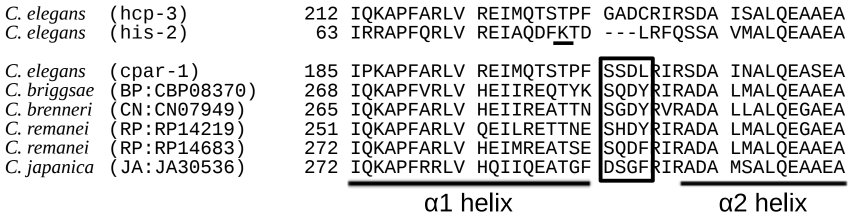
The H2AX variant is principally defined by the capacity to accept phosphorylation on a serine near the C-terminus through the activity of PIKKs such as ATM, ATR and DNA-PK on the consensus motif SQ[E/D]Φ (where Φ is a hydrophobic residue). The number of residues separating this motif from the core histone fold region is variable and has been claimed to correlate with the evolutionary complexity of the organism (*Redon et al.*, 2002). For example, the spacing of 29 residues between the end of the H2AX α3 helix and the phospho serine in *Saccharomyces cerevisiae* and *Giardia lamblia* is 12 residues shorter than in humans and mice.

In higher eukaryotes, H2AX is encoded as a separate histone variant of H2A but in lower organisms such as *S. cerevisiae*, *G. lamblia* and certain protists, the distinguishing H2AX features are merged into the canonical H2A (*Malik and Henikoff*, 2003; *Sullivan Jr. et al.*, 2006) so that the canonical H2A also acts as the H2AX variant. In *Drosophila melanogaster*, the H2AX feature is instead merged with H2A.Z as a single variant, H2AvD, that is distinct from the canonical H2A (*Madigan et al.*, 2002).

Based on phylogenetic analysis, it has been suggested that H2AX appeared multiple times in eukaryotes as an example of parallel evolution (*Malik and Henikoff*, 2003). However, the differences between metazoan H2AX and canonical H2A sequences are few in number and this could be confounding to phylogenetic algorithms. An alternative hypothesis is that the H2AX function is ancestral and canonical H2A evolved from the H2AX when complete phosphorylation became unnecessary or undesirable as genomes expanded. This would explain the preeminence of the H2AX variant in *G. lamblia* and *S. cerevisiae* compared to the lower abundance in mammals.

It has remained something of a puzzle that no H2AX variant function is identifiable in *Caenorhabditis elegans* (*Malik and Henikoff*, 2003) or some protists (*Sullivan Jr. et al.*, 2006). A search of all predicted *C. elegans* histones protein sequences reveals no PIKK consensus motifs in the coding sequence or in any frame downstream of the annotated stop codons for any of the core histone genes (data not shown). However, the related *C. briggsae* genome contains the motif SQDY within the cpar-1 isoform of CENP-A, the centromeric H3-like histone. Alignment of seven known *Caenorhabditis* CENP-A homologues shows that this motif is quite conserved, with the sequence being SSDL in *C. elegans* cpar-1 (Figure E.1). Although these do not strictly conform to the classic PIKK recognition site sequence, non-canonical sites are known to be recognised by them (*Sweeney et al.*, 2005).

This potential merger of H2AX with CENP-A is interesting for several reasons: Firstly, *C. elegans* utilises holocentric chromosomes so cpar-1 is thought to be distributed throughout the genome (*Monen et al.*, 2005). Secondly, cpar-1 appears to be more weakly expressed than the other CENP-A homologue in the *C. elegans* genome, hcp-3 (*Monen et al.*, 2005), recalling the H2AX/H2A ratio in human and mouse chromatin. Finally, the SQDY/SSDL motif is located at small 3–4 residue insertion



```
C. elegans   (hcp-3)          212 IQKAPFARLV REIMQTSTPF GADCRIRSDA ISALQEAAEA
C. elegans   (his-2)           63 IRRAPFQRLV REIAQDFKTD ---LRFQSSA VMALQEAAEA

C. elegans   (cpar-1)         185 IPKAPFARLV REIMQTSTPF SSDLRIRSDA INALQEASEA
C. briggsae  (BP:CBP08370)   268 IQKAPFVRLV HEIIREQTYK SQDYRIRADA LMALQEAAEA
C. brenneri  (CN:CN07949)    265 IQKAPFARLV HEIIREATTN SGDYRVRADA LLALQEGAEA
C. remanei   (RP:RP14219)    251 IQKAPFARLV QEILRETTNE SHDYRIRADA LMALQEGAEA
C. remanei   (RP:RP14683)    272 IQKAPFARLV HEIMREATSE SQDFRIRADA LMALQEAAEA
C. japanica  (JA:JA30536)    272 IQKAPFRRLV HQIIQEATGF DSGFRIRADA MSALQEAAEA
```

Figure E.1: Alignment of *Caenorhabditis* CENP-A homologues showing conservation of possible PIKK recognition site inserted within H3 structure. The major *C. elegans* CENP-A homologue (hcp-3) and a canonical H3 isoform (his-2) with lysine 79 underlined are shown above.

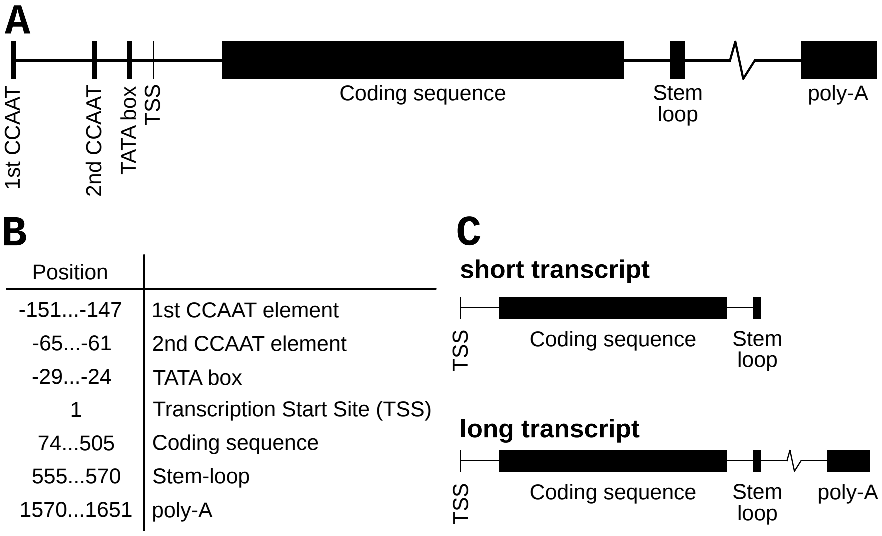
Table E.1: Localisation of all human canonical and variant H2A genes and proteins (adapted from *Marzluff et al.* (2002))

Histone cluster	Gene	Protein	Locus
HIST1	H2A A–E, G–M	H2A.1	6p21–22
HIST2	H2A A–C	H2A.1 and H2A.2	1q21
HIST3	H2A	H2A.1	1q42
—	H2AFB3	H2ABbd	Xq28
—	H2AFJ	macroH2A2	12p12
—	H2AFV	H2A.F/Z	7p13
—	H2AFX	H2AX	11q23.2–11q23.3
—	H2AFY	macroH2A1	5q31.3–q32
—	H2AFZ	H2A.Z	4q24

unique to the core histone fold of CENP-A family proteins. This insertion immediately abuts lysine 79 of canonical H3 which is also implicated in the DDR (Figure E.1). CENP-A family proteins do not have lysine at the equivalent of position 79 of canonical H3.

H2AX Gene

Canonical histone genes in humans are spread over one large and two small clusters named HIST1, HIST2 and HIST3. These are located at 6p21–p22, 1q21 and 1q42 respectively (Table E.1). Canonical H2A is encoded by sixteen genes, twelve of which are located in HIST1, three in HIST2 and one in HIST3. H2A variants are located outside these histone clusters in the human genome, with the H2AX–encoding gene H2AFX at 11q23.2–11q23.3 (*Ivanova et al.*, 1994b) (Table E.1). Histone variant gene names typically include the letter F for family.



Position	
-151...-147	1st CCAAT element
-65...-61	2nd CCAAT element
-29...-24	TATA box
1	Transcription Start Site (TSS)
74...505	Coding sequence
555...570	Stem-loop
1570...1651	poly-A

Figure E.2: H2AFX gene and transcripts. A. Schematic of H2AFX gene region showing promoter and 3′ mRNA stabilizing elements. B. Sequence coordinates of each element in H2AFX relative to transcription start site. C. Alternative transcripts of H2AFX. The short transcript (≈600 bp in size) ends in a stem–loop like canonical histones, whereas the long transcript (≈1600 bp in size) ends in a poly-(A) tail.

The H2AFX promoter region, 151 bp upstream from the transcription start site, shows higher activity than the typical canonical H2A.1 HIST1H2AE promoter in transcription reporter assays (*Ivanova et al.*, 1994a). There are two CCAAT elements upstream of the TATA box in H2AFX (Figure E.2) compared to a single CCAAT element in the H2A.1. The CCAAT element proximal to the TATA box in H2AFX has a significant effect on expression, whereas this element has no apparent effect on promoter activity in the canonical H2A promoter. The transcription factors that bind to the element also bind to the distal CCAAT as well as to three similar elements in H2AFZ but not to the one in the H2A.1 promoter (*Ivanova et al.*, 1994a). This suggests that H2AFX is regulated independently of canonical H2A.

H2AX Transcripts

A fundamental distinction between histone types is whether their expression is replication-dependent or replication-independent. This difference is a consequence of the requirement for large amounts of canonical histones during S phase to package the newly duplicated genome

(i. e. replication-dependence). In contrast, variant or "replacement" histones often appear to be inserted into chromatin to replace canonical histones for functional reasons throughout the cell cycle and are therefore replication-independent (*Marzluff et al.*, 2002).

Canonical histone genes lack introns, probably to circumvent the requirement for primary transcript processing when histones must be rapidly produced at S phase. A number of transcript features appear to enhance the capacity of replication-dependent histone expression by up to 35-fold during S phase. In fact, there is only a five-fold increase in their transcription rate at S phase, compared to the other phases of cell cycle so regulation acts strongly at the post-transcriptional level (*Harris et al.*, 1991). Replication-dependent histone transcripts lack a poly(A) tail and encode a stem–loop followed by a purine-rich Histone Downstream Element (HDE) downstream of the stop codon. The stem–loop interacts with the Stem–Loop Binding Protein (SLBP) to stabilise the mRNA in S phase (*Whitfield et al.*, 2000) while the HDE interacts with U7 snRNA to direct efficient 3′ end processing (*Georgiev and Birnstiel*, 1985).

Human H2AX transcripts exhibit characteristics of both replication-dependent and replication-independent histones. The H2AFX gene lacks introns, and has two alternative transcripts: one shorter form contains the characteristic stem–loop, and the other longer form contains a downstream poly(A) tail (*Mannironi et al.*, 1989) (Figure E.2). The combined synthesis of H2AX transcripts has been described as "weakly replication-linked at best" since the H2AFX promoter keeps the levels of both transcripts high through the cell cycle (*Ivanova et al.*, 1994a). However, the cell cycle linkage of the forms is unclear and no study has reported the effect of DNA damage on transcription levels.

H2AX Protein

Despite the large amount of attention paid to the DNA damage-linked serine phosphorylation by PIKKs, the H2AX protein itself has a number of additional unique properties.

The defining feature of H2AX is considered to be the C-terminal region with the SQ[E/D]Φ motif (Figure E.3). As mentioned in §E.2, the number of residues separating this motif from the histone fold is variable and claimed to correlate with the evolutionary complexity of the organism (*Redon et al.*, 2002). The residues responsible for this variable spacing are mainly hydrophilic with a high glycine and proline content suggesting a flexible, unstructured nature so the basis for the correlation could be more directly related to a structural constraint such as the variation in internucleosomal repeat lengths of organisms which itself shows linkage with evolutionary complexity.

In addition to the C-terminal motif, amino acid residues 6, 16, 38 and 99 of H2AX are different from the human H2A.1 consensus (Figure E.3 and E.4). Inspection of the human and *X. laevis* histone based nucleosome structures reveals that H2A residue 6 is located in the flexible N-terminal tail and residue 16 is located at the very base of the tail (Figure E.4 and E.5(b)) which tracks the minor groove at superhelical location 4.5 (SHL4.5) (Figure E.5(a)). The substitution of glutamine with threonine at position 6 in H2AX introduces a potential hydroxyl site for post-translational modification that is not present for the glutamine in canonical H2A. In contrast, the threonine to serine substitution conserves the modifiable hydroxyl at position 16.

Asparagine 38 is located in the loop between the $\alpha1$ and $\alpha2$ helices of H2A within the nucleosome (Figure E.4 and E.5(c)). Importantly, this residue makes direct contact with the equivalent amino acid in the other H2A–H2B dimer in the nucleosome structure and has been suggested to affect both nucleosome stability and the balance between homotypic and heterotypic combinations (see §E.2) of H2A types within the yeast nucleosome (*White et al.*, 2001). It is possible that the change of residue 38 from asparagine in H2A to histidine in H2AX could also affect nucleosome stability and dynamics. For example, weakening of interactions between the two H2A-H2B histone fold dimers could result in increased nucleosome flexing and impact the ability to condense into stable higher order chromatin structure. Furthermore, the presence of the histidine in H2AX could affect the stabilisation of a second copy of H2AX relative to canonical H2A within the nucleosome. This change of asparagine to histidine at position 38 occurs only in higher organism H2AX and could potentially drive a bias towards either homotypic H2AX-only or heterotypic H2AX–H2A mixed nucleosomes which could have consequences for the distribution of H2AX in chromatin (see §E.2).

The effect of the final substitution distinguishing canonical H2A and H2AX where lysine becomes glycine at position 99 is less clear. This

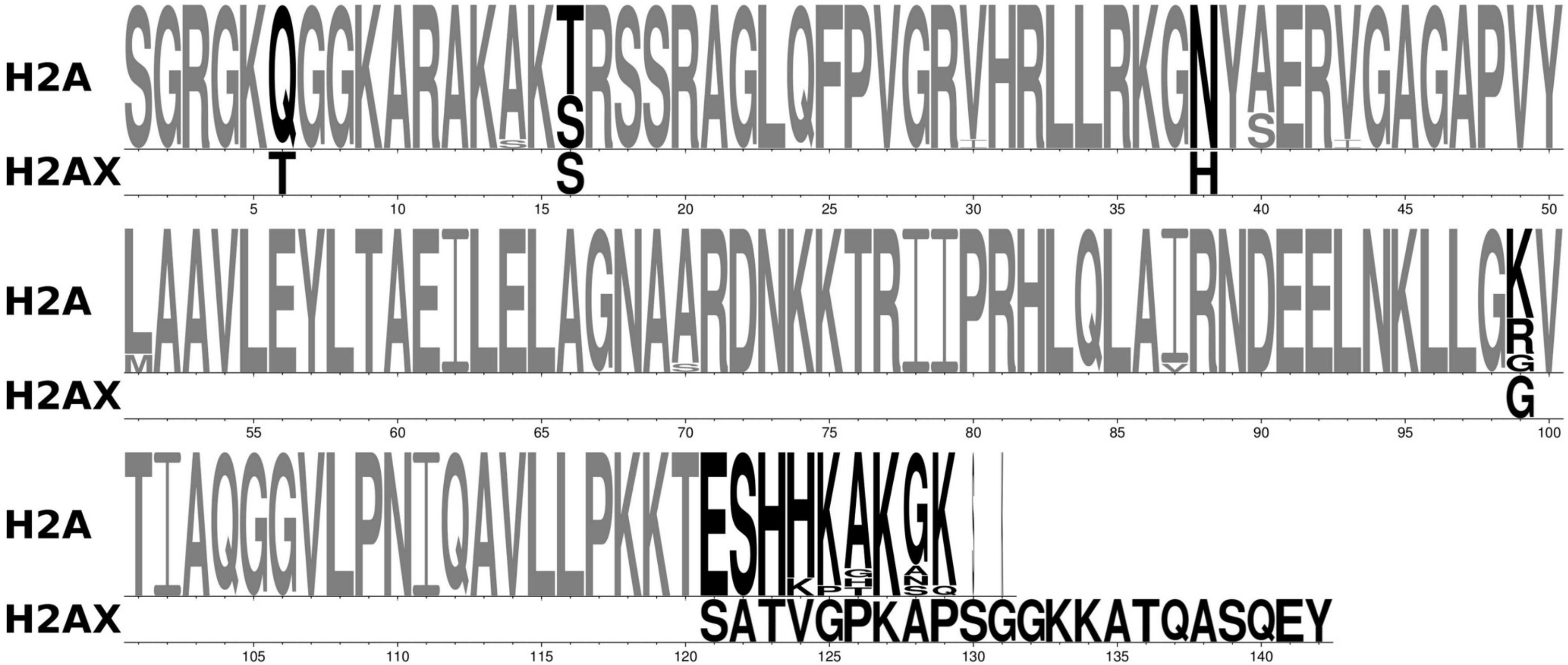


Figure E.3: Sequence logo of all human canonical H2A isoforms showing differences with H2AX below. The 4 residues changes from H2A to H2AX outside the C-terminal region are Gln 6 Thr, Thr 16 Ser, Asn 38 His and Lys 99 Gly. Alignment of H2A genes was made usign edialign (*Morgenstern*, 1999) from EMBOSS (*Rice et al.*, 2000) and WebLogo 3 (*Crooks et al.*, 2004).

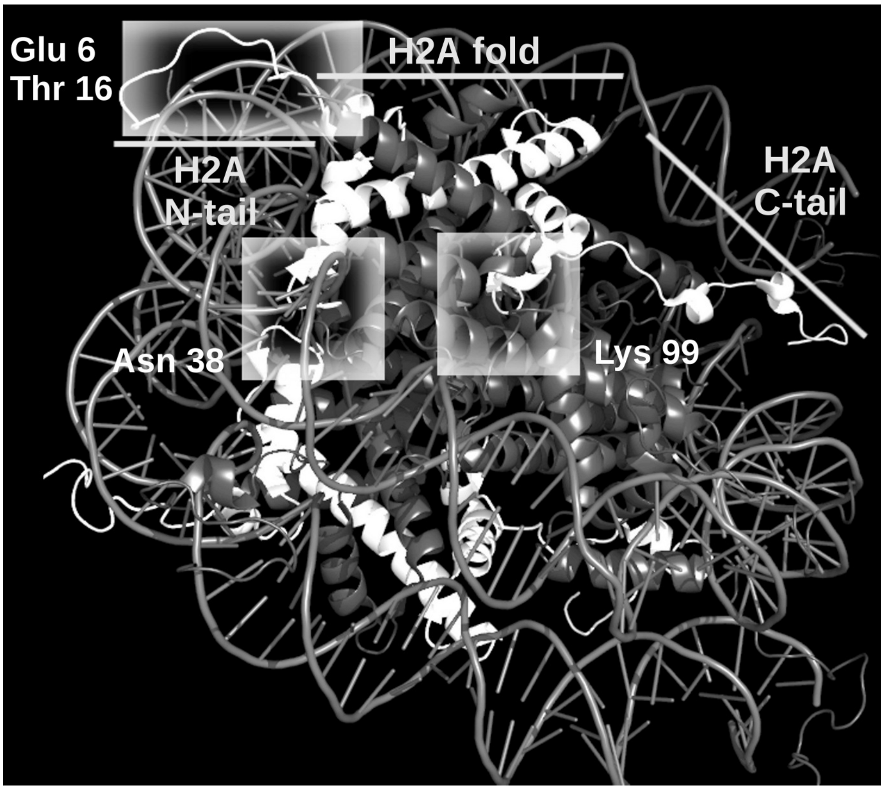


Figure E.4: Nucleosome structure highlighting differences between H2A and H2AX. H2A chain is highlighted and white frames indicate the position of the residues that differ between the human canonical H2A and H2AX. Image from PDB structure 1KX5 using PyMOL (*DeLano*, 2002).

residue is located in a sharp turn immediately after the $\alpha3$ helix and points towards the C-terminal ends of H3 and H4 but makes no direct interactions in the nucleosome (Figure E.4 and E.5(d)). Nevertheless, the exchange of the large, positively charged and potentially modifiable lysine for the highly flexible glycine in H2AX could potentially alter stability and flexibility of the nucleosome.

H2AX Post-Translational Modifications

Histones typically have a large proportion of amino acid residues which are modified post-translationally for functional reasons so it is significant that three of the four residues distinguishing human H2A and H2AX in the core region are capable of distinction via post-translational modification (i. e. Thr 6 and Ser 16 in H2AX vs. Thr 16 and Lys 99 in canonical H2A).

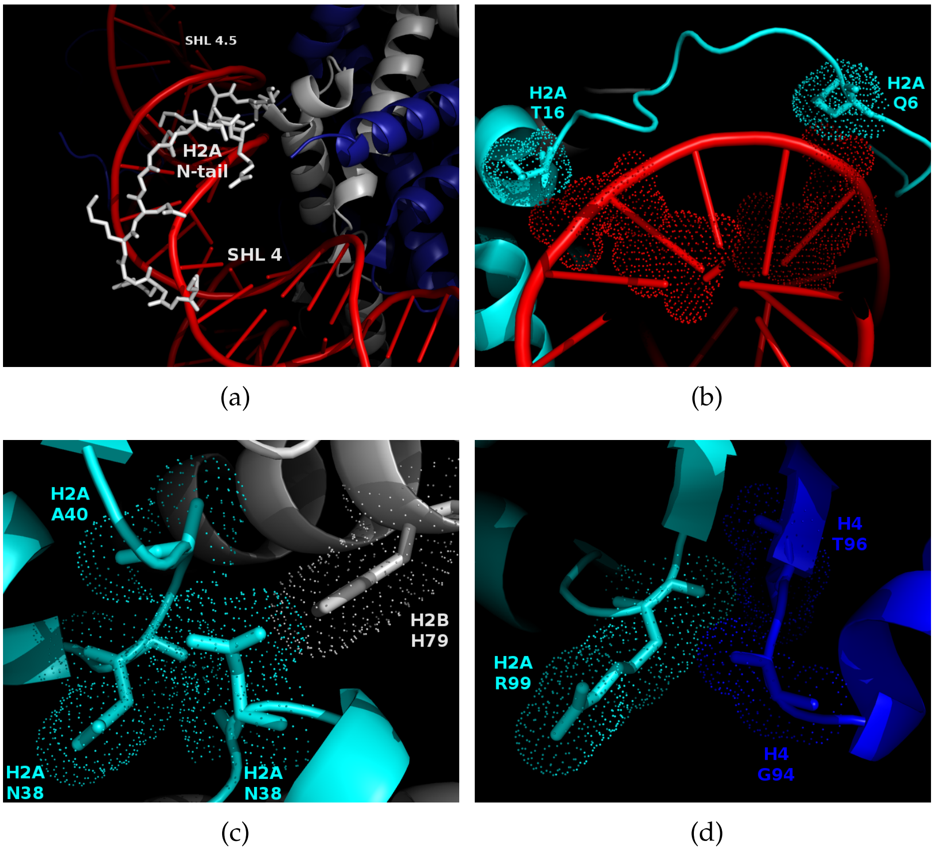


Figure E.5: Structural environment of H2A residues that differ from H2AX. The van der Waals surface of differences and all residues within 5 Å are shown as a surface of dots over bond sticks. Image from PDB structure 1KX5 using PyMOL (*DeLano*, 2002). (a) H2A N-terminal tail encompassing H2AX residues Thr 6 and Ser 16 passes across minor groove at superhelical location (SHL) 4.5. (b) Closeup of H2A N-terminal tail minor groove association from A showing canonical H2A Gln 6 and Thr 16 which become, respectively, Thr 6 and Ser 16 in H2AX. (c) Residues around H2A–H2A association in structure showing interaction between paired Asn 38 sidechains and adjacent residues. Canonical H2A Asn 38 is His 38 in mammalian H2AX. (d) Environment around H2A Arg 99 showing unusual absence of close packing. Canonical H2A Arg 99 is Gly 99 in H2AX.

However, only the phosphorylation of H2AX serine 139 by PIKKs in response to DNA damage has been intensively studied. This modification has been demonstrated to enhance access of restriction enzymes and DNA methylases to the DNA, possibly by reducing nucleosome stability (*Heo et al.*, 2008). In the same study the activity of the FACT complex which can facilitate dissociation of H2A/H2B dimers from nucleosomes was shown to increase after H2AX phosphorylation.

One of the most recently reported post-translational modifications of H2AX related to DSB is the phosphorylation of tyrosine 142 in the PIKK recognition motif of human H2AX (*Xiao et al.*, 2009; *Cook et al.*, 2009). In contrast to the phosphorylation of Ser 139, this Tyr 142 residue is phosphorylated under normal conditions with DNA damage acting as trigger for its dephosphorylation. The dephosphorylation seems to not only precede the phosphorylation of Ser 139, but also to be a prerequisite for the Ser 139 phosphorylation. When Tyr 142 is phosphorylated, affinity of Ser 139 to the DNA damage response factors MDC1, MRE11 and Rad50 is greatly reduced and binding by pro-apoptopic factor JNK1 was found to occur instead. It has therefore been suggested that the phosphorylation status of Tyr 142 is a determinant of cell fate after DNA damage.

H2AX can also be subject of acetylation at lysine 5 (*Pantazis and Bonner*, 1981) and to both mono- and poly-ubiquitylation at lysine 119 dependent on the prior acetylation at Lys 5 (Table E.2) (*Ikura et al.*, 2007). These modifications are intimately related to DNA repair because their levels increase significantly after exposure to DSB-inducing Ionising Radiation (IR) and appear to drive H2AX eviction from the nucleosome by the action of Tip60 complex and UBC13 (*Ikura et al.*, 2007). However, conflicting data about the interdependence of these effects with phosphorylation has recently been reported (*Rios-Doria et al.*, 2009).

Other modifications unrelated to DNA damage have been reported for H2AX, including a rather unusual biotinylation of Lys 9 and Lys 13 (*Chew et al.*, 2006) and the phosphorylation of Ser 1 (*Pantazis and Bonner*, 1981). By homology to canonical H2A, it is probable that Lys 9 and Lys 13 can also be acetylated (*Zhang et al.*, 2003) and Thr 120 phosphorylated (*Aihara et al.*, 2004). Another interesting possible post-translational modification is a methylation at Lys 127 (*Zhang et al.*, 2003). Although it was inconclusive whether Lys 125 or Lys 127 is the target of this methylation, it is tempting to speculate that it occurs at Lys 127 since this residue

Table E.2: Reported Post-Translational Modifications (PTMs) for H2AX. Other PTMs present in H2A but not yet related to H2AX include acetylation of lysine 9 and 13 (*Zhang et al.*, 2003), phosphorylation of threonine 120 (*Aihara et al.*, 2004), and the possible methylation of lysine 127 (*Zhang et al.*, 2003).

Residue number	Residue indentity	PTM	Related to DSB	Reference
1	Serine	phosphorylation	no	*Pantazis and Bonner* (1981)
5	Lysine	acetylation	yes	*Pantazis and Bonner* (1981) and *Ikura et al.* (2007)
9	Lysine	biotinylation	no	*Chew et al.* (2006)
13	Lysine	biotinylation	no	*Chew et al.* (2006)
119	Lysine	ubiquitylation	yes	*Ikura et al.* (2007)
139	Serine	phosphorylation	yes	*Rogakou et al.* (1998)
142	Tyrosine	phosphorylation	yes	*Xiao et al.* (2009)

is the only one conserved in the C-terminal of all human H2A sequences (Figure E.3).

H2AX Distribution in Chromatin

The original estimates of H2AX abundance in human cells reported cell line specific values from 2.5 % to 25 % of total H2A in asynchronous immortalised cell lines (*Rogakou et al.*, 1998). These values were determined by densitometry of Coomassie-stained, acid-extracted histones in two-dimensional gels. A 10 % abundance value of H2AX has become accepted despite wide differences in the study and the fact that HeLa cells were reported to contain 2.5 % H2AX.

Although it is tempting to interpret 10 % abundance as implying every tenth nucleosome will contain H2AX, combinatorial features of nucleosomes make the statistics of spacings between H2AX occurrences in the chromatin fibre more complex. H2AX can be incorporated either as one or as two copies per nucleosome (Figure E.6A), and the H2AX-containing nucleosomes can be either randomly or non-randomly distributed along the chromatin fibre (Figure E.6B–C). Random incorporation would lead not simply to each tenth nucleosome containing H2AX,

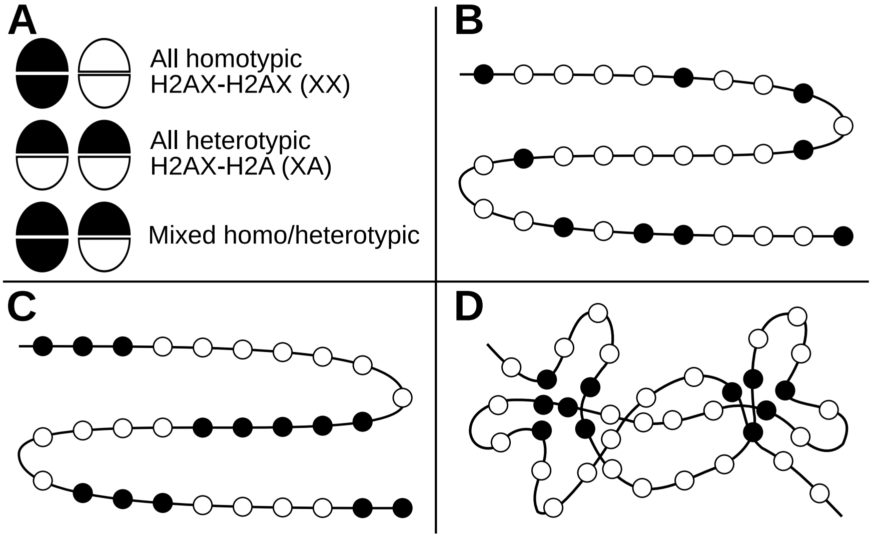


Figure E.6: H2AX distribution in the chromatin. A. Schematic of possible H2AX homotypic, heterotypic and mixed nucleosome combinations. Black semicircle represents H2AX–H2B dimer and white semicircle represents H2A–H2B dimer. B. Random incorporation of H2AX into nucleosomes would lead to a random distribution of H2AX-containing nucleosomes. C. Selective incorporation of H2AX into nucleosomes would lead to "islands" of H2AX-containing nucleosomes. D. Random incorporation of H2AX nucleosomes could also lead to "islands" of H2AX nucleosomes by chromatin reorganization.

but to a geometric distribution of spacings between H2AX-containing nucleosomes. This predicts many instances of small spacings and some instances of very large spacings, and has clear implications for the ability of γH2AX to signal local damage events as well as for the spreading of the phosphorylation along the chromatin fibre.

Combinatorial potential in H2AX distribution

The combinatorial potential for H2AX inclusion has two separate features which could affect the detailed distribution of H2AX along chromatin.

Firstly, either one or two H2AX polypeptides can in principle be present within a nucleosome: Two H2AX copies would give rise to a "homotypic" H2AX/H2AX ('XX') nucleosome, whereas a single H2AX copy will give rise to a "heterotypic" H2AX/H2A ('XA') nucleosome (Figure E.6A). It is currently unknown whether there is a bias for either homotypic or heterotypic nucleosomes (see §E.2) although this affects the statistics of H2AX spacing in chromatin since the XA combination yields twice as many H2AX-containing nucleosomes than XX for a given H2AX abundance.

Secondly, the spacing of nucleosomes containing H2AX should have a major influence on its functional roles in DSB signaling, assembling of repair foci and facilitating the repair machinery. H2AX nucleosomes could be randomly distributed (Figure E.6B) or subject to clustering in one (Figure E.6C) or three dimensions (Figure E.6D). Any mechanism randomly assembling chromatin from pools of XX and/or XA versus canonical H2A–H2A ('AA') nucleosomes will give rise to a geometric distribution of spacings between H2AX (Figure E.7A). This distribution predicts a bias towards small spacings (Figure E.7A).

Simulation of random H2AX inclusion

Simple computational simulations reveal interesting features in this H2AX spacing distribution. In the simplest case of H2AX assembling in a mixture of XA and XX nucleosomes, 10 % overall H2AX abundance would generate an average of 4.3 nucleosomes between H2AX occurrences along the chromatin fibre (Figure E.7B). The mean spacing is highly sensitive to H2AX abundance (Figure E.7A), so 2.5 % and 25 % H2AX abundances yields means of 19.3 to 1.3 nucleosomes, respectively (Figure E.7B). Similar results arise for calculations where only heterotypic XA nucleosomes can assemble and homotypic XX nucleosomes are structurally precluded. In contrast, if heterotypic XA nucleosomes are precluded and only XX nucleosome structures can assemble, then 10 % H2AX abundance yields a mean spacing of 9 nucleosomes between H2AX occurrences. The mean spacings for 2.5 % and 25 % H2AX abundance are 39 to 3 nucleosomes respectively (Figure E.7B).

Functional implications of H2AX distribution

These simple models of random nucleosome incorporation have interesting implications. The occurrence of occasional large H2AX spacings could limit both processive γH2AX spreading along the chromatin fibre and the proximity of H2AX in solenoidal higher order chromatin packaging. At 10 % H2AX abundance, 23 % of nucleosomes in mixed XA and XX nucleosomes will be spaced by more than 6 nucleosomes and in the

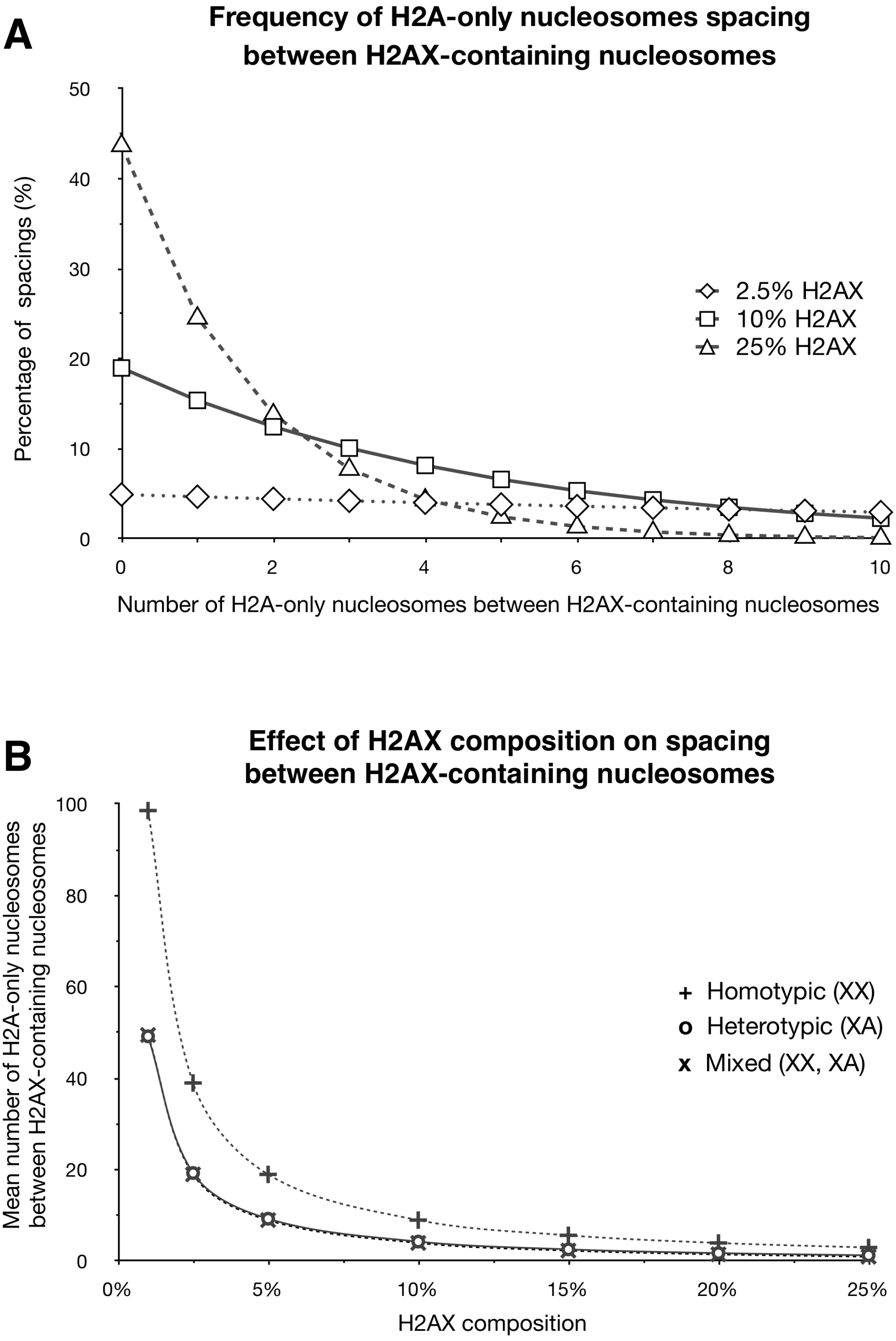


Figure E.7: Simulations of H2AX spacing distributions. A. Distribution of spacings between instances of H2AX for mixed population of homotypic and heterotypic nucleosomes at abundances of 2.5 % (dot), 10 % (solid) and 25 % (dashed) H2AX in total H2A pool. B. Effect of abundance on mean H2AX spacing for homotypic (H2AX–H2AX) only, heterotypic (H2AX–H2A) only, and mixed homotypic+heterotypic nucleosome combinations.

extreme case of 2.5 % H2AX abundance, 84 % of solely homotypic XX nucleosomes would be spaced by more than 6 nucleosomes.

This sensitivity of H2AX spacing in chromatin to abundance provides a potential opportunity for the cell to regulate responsiveness to damage. For example, if H2AX expression is up-regulated the mean proximity of randomly inserted H2AX will rapidly increase and effects such as processive γH2AX spreading and retention of DDR factors at foci will be significantly enhanced.

It is unknown whether H2AX distribution varies between euchromatin and heterochromatin. However, differences in H2AX response have been reported according to the condensation level of chromatin and phosphorylation of Ser 139 has been observed to occur preferentially in euchromatin (*Cowell et al.*, 2007). This preference is overcome during replication of heterochromatin when it is in a less condensed state (*Cowell et al.*, 2007). The distinction between active and inactive chromatin can also be regulated, as demonstrated for phosphorylation of KAP-1 by ATM reducing the access of DNA repair proteins to heterochromatic regions of the genome (*Goodarzi et al.*, 2008).

Possibility of non-random H2AX distribution

If H2AX nucleosome incorporation is not a random process (Figure E.6B), inhomogeneity could also exist at a more local level. For example, small "islands" of higher density H2AX nucleosomes could be interspersed within broader regions with lower relative abundance of the variant (Figure E.6C). A recent study using a novel high-resolution microscopy observed several thousand small spatial clusterings of H2AX and pointed to a mutual exclusivity of H2AX and the phosphorylated form (*Bewersdorf et al.*, 2006). This would be consistent with a clustering model (Figure E.6D) that enhanced the kinetics of the damage signaling at foci, perhaps by making use of a chromatin structural feature such as the chromosomal scaffold (*Bewersdorf et al.*, 2006). The inherent clustering and active insertion of H2AX in the DDR could also drive larger scale chromosomal rearrangements through chromatin stability (Figure E.6D) (*Heo et al.*, 2008).

E.3 Functional Roles of H2AX

Phosphorylation of H2AX serine 139 by PIKKs to generate "γH2AX foci" is an early and characteristic feature of DSB events. This modification is thought to be the primary identifier of the location of DNA damage and would therefore be central to the function of H2AX.

The γH2AX foci extend for 2 Mbp to 30 Mbp along the chromatin fibre (*Rogakou et al.*, 1999), implying the involvement of a span of 10^4 to 10^5 nucleosomes per individual DSB repair event. At 10 % H2AX abundance, this would involve up to 10^2 to 10^4 H2AX molecules and hence a 10^2 to 10^4 fold amplification of DSB event signal. A direct link between the site of a lesion and a single focus has been observed (*Rothkamm et al.*, 2003), suggesting that there is a linkage between γH2AX and the repair mechanism. Many protein factors have been identified, which depend directly or indirectly on the phosphorylation of H2AX at serine 139. Thereafter, it appears to act as a foundation for recruitment of DDR factors at DSB sites (*Paull et al.*, 2000). As a consequence, H2AX performs a role in both localisation and structuring of the repair focus.

Initiation of H2AX Phosphorylation as a Reporter of DSB Events

The process of establishing H2AX phosphorylation at the characteristic terminal motif can be performed by any of the three PIKKs ATM, ATR and DNA-PK. Their induction and binding characteristics suggest that H2AX phosphorylation for focus generation can be distinguished by an initiation phase when a small number of phosphorylations are made at nucleosomes adjacent to the break, and a spreading phase in which a larger region of phosphorylation extends one-dimensionally from either side of the break. The structural exposure of the serine 139 site through chromatin flexibility will be crucial determinant of the modification event (see §E.2).

ATM has been considered a strong candidate as the principal kinase responsible for the initiation phase of general damage events because it responds to changes in chromatin conformation expected when a spontaneous DSB event releases local superhelical tension (*Bakkenist and Kastan*, 2003). ATR appears to be linked to replication stress or UV damage events which lead to breaks as indirect consequences, so ATR is recruited by ATRIP which detects single-stranded DNA. DNA-PK is localised to DSBs in complex with the end-binding protein Ku, so such an association will act to limit the distance from the damaged end on which DNA-PK can act (*Walker et al.*, 2001) and such an end-dependent mechanism would be sensitive to H2AX abundance and distribution.

Spreading of H2AX Phosphorylation as a Damage Signal Amplifier

The conventional model for γH2AX focus formation suggests that after initiation in the immediate vicinity of the break by ATM and/or DNA-PK, amplification occurs by spreading through the action of MDC1 binding to γH2AX (*Stucki et al.*, 2005). MDC1 in turn recruits the MRN complex (Mre11–Rad50–Nbs1) via direct interaction with Nbs1 (*Lukas et al.*, 2004) and the MRN complex further activates ATM (*Uziel et al.*, 2003). This generates a positive feedback loop to drive spreading of the phosphorylation modification away from the break. Hence H2AX acts both as signal and target of phosphorylation in the spreading phase. Each focus acts independently even when several foci are formed in the immediate vicinity of each other (*Kruhlak et al.*, 2006), suggesting a one dimensional diffusion along the chromatin fibre.

How the signal spreads over megabase but non-infinite distances is unknown. It is possible that non-homogeneous H2AX distribution could contribute to the localisation of γH2AX stochastically through random occurrence of large spacings between H2AX that the spreading mechanism could not bridge (see §E.2). Consistent with this, high resolution microscopy reveals that H2AX is not randomly distributed but organized into discrete clusters which would control the expansion of the signal (*Bewersdorf et al.*, 2006).

Since levels of phosphorylated H2AX rise rapidly in response to damage and then reduce over time (*Rogakou et al.*, 1998) it is necessary to remove either the phosphate or the entire γH2AX. The timing of this process is unclear but must depend on the presence of γH2AX binding factors such as MDC1 which could stabilise γH2AX or obscure the phosphate group (*Stucki et al.*, 2005). In *S. cerevisiae*, dephosphorylation is achieved by removal of phosphorylated H2AX from nucleosomes and subsequent dephosphorylation by the HTP-C complex (*Keogh et al.*, 2006). In higher eukaryotes the mechanisms remain unclear since several phosphatases have been implicated in the process and these can variously dephosphorylate H2AX within nucleosomes or after removal (*Chowdhury et al.*, 2005; *Kimura et al.*, 2006; *Chowdhury et al.*, 2008). In addition, the FACT complex which facilitates nucleosome exchange has enhanced activity on phosphorylated H2AX (*Heo et al.*, 2008) suggesting at least one pathway involving displacement for extra-nucleosomal dephosphorylation. A background level of H2AX remains phosphorylated even in the apparent absence of DNA damage, but the reason for this is unknown (*Rogakou et al.*, 1998).

γH2AX and Chromatin Structural Remodelling

Intrinsically, H2AX phosphorylation must take place within the context of chromatin structure so both the Non-Homologous End Joining (NHEJ) and Homologous Recombination (HR) pathways can efficiently undertake DSB repair. To facilitate this, chromatin decondenses near the DSB (*Kruhlak et al.*, 2006) but the mechanism for this remodeling is unclear.

The modified serine 139 of H2AX is located near the DNA entry/exit point on the nucleosome (Figure E.4) so one putative mechanism for the chromatin structural change is to be driven directly by the chemical properties of the added phosphate group. *S. cerevisiae* mutants with the serine 139 equivalent mutated to glutamate to mimic the phosphate charge show increased micrococcal nuclease sensitivity consistent with such a destabilisation (*Downs et al.*, 2000) and phosphorylated human H2AX renders chromatin more susceptible to restriction enzymes and DNA methylase (*Heo et al.*, 2008). However, a separate analysis of chromatin structure, also in yeast, harboring the glutamate mutation did not find evidence of direct chromatin structural effects (*Fink et al.*, 2007).

An alternative indirect mechanism for linking H2AX phosphorylation with chromatin disruption is by recruitment of proteins to drive remodeling. A number of different ATP-dependent chromatin remodeling activities have been implicated in this process, including RSC, SWI/SNF, INO80 and SWR (reviewed in *Downs et al.* (2007)), as well as other nucleosome modifying enzymes such as the NuA4 histone acetyltransferase.

There is also evidence that chromatin chaperones and binding proteins contribute to the process of chromatin dynamics at DSBs. For example, the FACT complex, which participates in exchange between H2A and H2AX, has greater ability to mobilise γH2AX than unphosphorylated H2AX (*Heo et al.*, 2008). In addition, HP1β, which binds to H3 K9me, has recently been shown to be released by phosphorylation immediately after DSB events and that this contributes to H2AX phosphorylation by PIKKs (*Ayoub et al.*, 2008).

Both direct and indirect mechanisms for chromatin remodeling depend on H2AX phosphorylation, and hence require an independent initiation step. The PIKKs ATM and DNA-PK can achieve this by detecting changes in chromatin structure or appearance of DNA ends, respectively (*Bakkenist and Kastan*, 2003; *Burma and Chen*, 2004). However, the impact of chromatin on PIKK initiation is difficult to probe because H2AX phosphorylation occurs very rapidly after DSBs, making it difficult to temporally distinguish factors which remodel chromatin to enable initial PIKK access from downstream events which undertake remodeling to amplify γH2AX around the site.

Furthermore, despite the intimate link between H2AX phosphorylation and chromatin remodeling at the DSB site, local decondensation of chromatin occurs at similar levels on both wild type and H2AX$^{-/-}$ cell lines when ATP is not depleted (*Kruhlak et al.*, 2006). This suggests that the role of H2AX phosphorylation in driving the chromatin remodeling is redundant with other pathways.

γH2AX and Localisation of DSB Repair Proteins

Since H2AX phosphorylation is one of the earliest events after a DSB, this suggests it may play a role in subsequent recruitment of the active repair proteins. This is supported by the absence of RAD51 and BRCA1 at DSB foci when γH2AX phosphorylation is prevented (*Paull et al.*, 2000). However, NBS1, BRCA1 and 53BP1 are recruited to the sites of damage in H2AX$^{-/-}$ cell lines which display only moderate sensitivity to ionising radiation but fail to maintain focal localisation (*Celeste et al.*, 2003a). It has therefore been suggested that the crucial role of H2AX phosphorylation is not as a direct agent of repair factor recruitment, but of retention of these factors in the vicinity of the DSB (*Celeste et al.*, 2003a). This role in defining a "damage neighborhood" does not necessarily imply a direct role in repair at the break site itself. For example, stimulation of the G2/M checkpoint may result from the accumulation of checkpoint signalling factors at the focus (*Fernandez-Capetillo et al.*, 2002). In fact, Chromatin ImmunoPrecipitation (ChIP) revealed that γH2AX is evicted from the region close to the DSB early in the DDR in *S. cerevisiae* and that γH2AX does not strictly co-localise with the active repair complexes (*Shroff et al.*, 2004).

This accumulated retention of DDR factors in the vicinity of a DSB appears to be a complex process where the initiating damage signal is integrated by factors recognising the H2AX phosphorylation and presumably additional chromatin features. For example, human 53BP1 and its putative homologues, *S. cerevisiae* Rad9 and *S. pombe* Crb2, all contain Tudor domains which bind specific methylated histones in chromatin, and BRCT domains which can both mediate dimerisation and bind γH2AX. Despite the similarity in domain structure of Rad9, Crb2 and 53BP1, individual investigations have indicated that they have different binding capabilities. The Rad9 Tudor domain binds H3 K79me (*Grenon et al.*, 2007; *Huyen et al.*, 2004) whereas Crb2 and 53BP1 Tudor domains bind H4 K20me2 (*Sanders et al.*, 2004; *Botuyan et al.*, 2006). Rad9 and Crb2 BRCT domains bind directly to γH2AX (*Hammet et al.*, 2007; *Kilkenny et al.*, 2008) whereas 53BP1 does not, instead relying on an indirect interaction mediated by the BRCT domain of MDC1 which directly binds γH2AX (*Lee et al.*, 2005; *Stucki et al.*, 2005). Some direct interaction between 53BP1 BRCT domain and γH2AX has also been reported by co-precipitation studies, but in a much smaller proportion than Rad9 and Crb2 (*Kilkenny et al.*, 2008). Rad9 and Crb2 can all also dimerise or oligomerise through their BRCT domains (*Soulier and Lowndes*, 1999; *Du et al.*, 2004) although this domain is not necessary for the oligomerisation of 53BP1 (*Adams et al.*, 2005). The latter instead requires a sequence upstream of its Tudor domain (*Ward et al.*, 2006).

This complex interplay between the combinatorial interactions made by 53BP1, Rad9 and Crb2 with themselves and with γH2AX builds up to generate another level of the structural environment for the repair process. γH2AX therefore acts as a foundation to define the extent of the repair focus through the H2AX distribution and the extent of its phosphorylation.

γH2AX and Maintenance of Proximity of Break Ends

Linked to this role in retaining repair factors in the repair focus, phosphorylated H2AX also appears to function in the bringing together of damaged ends. It has been suggested that by recruiting repair factors which directly associate with the damaged ends, H2AX could prevent diffusion of these ends away from each other (*Bassing and Alt*, 2004). For example, linkage has been observed in the distribution of cohesin and γH2AX near DSBs (*Unal et al.*, 2004) so γH2AX-dependent cohesin association would promote the stabilisation of sister chromatids to facilitate HR. Furthermore, localisation of self-interacting factors by their association with γH2AX nucleosomes could bring together distant break ends. For example, 53BP1 is suggested to localise to break ends by direct interaction with nucleosomes and indirect interaction through MDC1 (*Huyen et al.*, 2004; *Botuyan et al.*, 2006; *Eliezer et al.*, 2009). Oligomerisation of 53BP1 has also been reported to enhance association of distant ends, thereby facilitating long range recombination and NHEJ (*Difilippantonio et al.*, 2008; *Dimitrova et al.*, 2008).

γH2AX and Complementary Damage Signalling Via Ubiquitylation

A secondary pathway of signaling by ubiquitylation of both canonical H2A and H2AX has recently been uncovered which appears to derive directly from γH2AX, and therefore act as a complementary amplification of the damage signal (*Panier and Durocher*, 2009). Recognition of H2AX phosphorylation by MDC1 leads to recruitment of an initiating ubiquitylation by RNF8 and UBC13 which is subsequently amplified with the involvement of RNF158, and possibly maintained by Rap80 and BRCT1. The direct role of the ubiquitylation remains to be clarified because it can act in factor recruitment as well as affecting the structure, stability or turnover of histones including H2AX itself.

E.4 Conclusion

Despite H2AX having a highly similar primary sequence or even overlapping identity with canonical H2A, it is clear that the DNA damage-linked function of this histone variant is highly specific. Its functional role is as an amplifier of the damage event signal, a foundation for marshaling repair factors, and a promoter of the chromatin dynamics required to complete the repair process. It is clear that the phosphorylation of serine 139 by PIKKs generates an epitope which is crucial to these functions. Nevertheless, it is important to note that the DNA damage response is only moderately defective in H2AX$^{-/-}$ cells, suggesting that complementary mechanisms must operate redundantly with H2AX functions. Much remains to be appreciated about γH2AX structure and function, but this must ultimately be based on the unique distinguishing features of the H2AX gene and protein.

Acknowledgements

We thank Prof. Cathal Seoighe for assistance with calculations of random H2AX distribution, Prof. Noel Lowndes for his input and Dr. Kevin Roche for helpful discussions. We gratefully acknowledge the support of Science Foundation Ireland and Health Research Board of Ireland for supporting work in our laboratory. DMSP acknowledges the support of the Portuguese Foundation for Science and Technology (FCT).

Bibliography

Abercrombie, M., Contact inhibition in tissue culture, *In vitro*, *6*(2), 128–142, doi:10.1007/BF02616114, 1970.

Abney, J. R., B. Cutler, M. L. Fillbach, D. Axelrod, and B. A. Scalettar, Chromatin dynamics in interphase nuclei and its implications for nuclear structure, *The Journal of cell biology*, *137*(7), 1459–1468, doi: 10.1083/jcb.137.7.1459, 1997.

Adams, M. M., B. Wang, Z. Xia, J. C. Morales, X. Lu, L. A. Donehower, D. A. Bochar, S. J. Elledge, and P. B. Carpenter, 53BP1 oligomerization is independent of its methylation by PRMT1, *Cell Cycle*, *4*(12), 1854–1861, 2005.

Agresti, A., P. Scaffidi, A. Riva, V. R. Caiolfa, and M. E. Bianchi, GR and HMGB1 interact only within chromatin and influence each other's residence time, *Molecular cell*, *18*(1), 109–121, doi:10.1016/j.molcel.2005.03.005, 2005.

Aihara, H., et al., Nucleosomal histone kinase-1 phosphorylates H2A Thr 119 during mitosis in the early Drosophila embryo, *Genes & Development*, *18*, 877–888, 2004.

Alberts, B., D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential cell biology*, 3 ed., Garland science, 2010.

Alberts, B., A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 6 ed., Garland Science, 2014.

Albig, W., and D. Doenecke, The human histone gene cluster at the D6S105 locus, *Human Genetics*, *101*(3), 284–294, doi:10.1007/s004390050630, 1997.

Albig, W., P. Kioschis, A. Poustka, K. Meergans, and D. Doenecke, Human histone gene organization: nonregular arrangement within a large cluster, *Genomics*, *40*(2), 314–322, doi:10.1006/geno.1996.4592, 1997.

Albig, W., B. Drabent, N. Burmester, C. Bode, and D. Doenecke, The haemochromatosis candidate gene HFE (HLA-H) of man and mouse is located in syntenic regions within the histone gene cluster, *Journal of Cellular Biochemistry*, *69*(2), 117–126, doi:10.1002/(SICI)1097-4644(19980501)69:23.0.CO;2-V, 1998.

Allan, C., et al., OMERO: flexible, model-driven data management for experimental biology, *Nature methods*, *9*(3), 245–253, doi:10.1038/nmeth.1896, 2012.

Allfrey, V., R. Faulkner, and A. Mirsky, Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis, *Proceedings of the National Academy of Sciences*, *51*(5), 786–794, 1964.

Ando, R., H. Hama, M. Yamamoto-Hino, H. Mizuno, and A. Miyawaki, An optical marker based on the UV-induced green-to-red photoconversion of a fluorescent protein, *Proceedings of the National Academy of Sciences*, *99*(20), 12,651–12,656, 2002.

Andrews, A. J., and K. Luger, Nucleosome structure(s) and stability: variations on a theme, *Annual review of biophysics*, *40*, 99–117, doi:10.1146/annurev-biophys-042910-155329, 2011.

Angiuoli, S. V., et al., CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing, *BMC bioinformatics*, *12*(1), 356, doi:10.1186/1471-2105-12-356, 2011.

Arents, G., and E. N. Moudrianakis, The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization, *Proceedings of the National Academy of Sciences*, *92*(24), 11,170–11,174, 1995.

Arents, G., R. W. Burlingame, B.-C. Wang, W. E. Love, and E. N. Moudrianakis, The nucleosomal core histone octamer at 3.1 Å resolution: a tripartite protein assembly and a left-handed superhelix, *Proceedings of the National Academy of Sciences*, *88*(22), 10,148–10,152, doi: 10.1073/pnas.88.22.10148, 1991.

Ausió, J., Histone variants-the structure behind the function, *Briefings in Functional Genomics and Proteomics*, *5*(3), 228–243, 2006.

Axelrod, D., D. E. Koppel, J. Schlessinger, E. Elson, and W. W. Webb, Mobility measurement by analysis of fluorescence photobleaching recovery kinetics, *Biophysical journal*, *16*(9), 1055–1069, 1976.

Ayoub, N., A. D. Jeyasekharan, J. A. Bernal, and A. R. Venkitaraman, HP1-beta mobilization promotes chromatin changes that initiate the DNA damage response, *Nature*, *453*, 682–686, 2008.

Bakkenist, C. J., and M. B. Kastan, DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation, *Nature*, *421*, 499–506, 2003.

Ball, P., Portrait of a molecule, *Nature*, *421*(6921), 421–422, doi:10.1038/nature01404, 2003.

Bannister, A. J., and T. Kouzarides, Regulation of chromatin by histone modifications, *Cell research*, *21*(3), 381–395, doi:10.1038/cr.2011.22, 2011.

Bassing, C. H., and F. W. Alt, H2AX may function as an anchor to hold broken chromosomal DNA ends in close proximity, *Cell Cycle*, *3*(2), 149–153, 2004.

Baxevanis, A. D., and D. Landsman, Histone sequence database: a compilation of highly-conserved nucleoprotein sequences, *Nucleic Acids Research*, *24*, 245–247, doi:10.1093/nar/24.1.245, 1996.

Baxevanis, A. D., and D. Landsman, Histone and histone fold sequences and structures: a database, *Nucleic Acids Research*, *25*, 272–273, doi: 10.1093/nar/25.1.272, 1997.

Baxevanis, A. D., and D. Landsman, Histone sequence database: new histone fold family members, *Nucleic Acids Research*, *26*, 372–375, doi: 10.1093/nar/26.1.372, 1998.

Belmont, A. S., and K. Bruce, Visualization of G1 chromosomes: a folded, twisted, supercoiled chromonema model of interphase chromatid structure, *The Journal of cell biology*, *127*(2), 287–302, 1994.

Bewersdorf, J., B. T. Bennett, and K. L. Knight, H2AX chromatin structures and their response to DNA damage revealed by 4Pi microscopy, *PNAS*, *103*(48), 18,137–18,142, 2006.

Bezanson, J., A. Edelman, S. Karpinski, and V. B. Shah, Julia: A fresh approach to numerical computing, *arXiv preprint arXiv:1411.1607*, 2014.

Black, B. E., and D. W. Cleveland, Epigenetic centromere propagation and the nature of CENP-A nucleosomes, *Cell*, *144*(4), 471–479, doi:10.1016/j.cell.2011.02.002, 2011.

Bonenfant, D., M. Coulot, H. Towbin, P. Schindler, and J. v. Oostrum, Characterization of histone H2A and H2B variants and their post-translational modifications by mass spectrometry, *Molecular & Cellular Proteomics*, *5*(3), 541–552, 2006.

Bork, P., and E. Koonin, Predicting functions from protein sequences - where are the bottlenecks?, *Nature Genetics*, *18*(4), 313–318, doi: 10.1038/ng0498-313, 1998.

Botuyan, M. V., J. Lee, I. M. Ward, J.-E. Kim, J. R. Thompson, J. Chen, and G. Mer, Structural basis for the methylation state-specific recognition of histone H4-K20 by 53BP1 and Crb2 in DNA repair, *Cell*, *127*, 1361–1373, 2006.

Bradbury, E. M., Foreword: Histone nomenclature, in *The structure and function of chromatin*, *Ciba foundation symposium*, vol. 28, edited by D. W. Fitzsimons and G. E. W. Wolstenholme, pp. 1–4, Ciba foundation, 1975.

Bravo, R., and J. E. Celis, A search for differential polypeptide synthesis throughout the cell cycle of HeLa cells, *The Journal of cell biology*, *84*(3), 795–802, 1980.

Brehm, A., K. R. Tufteland, R. Aasland, and P. B. Becker, The many colours of chromodomains, *Bioessays*, *26*(2), 133–140, doi:10.1002/bies.10392, 2004.

Brown, G. R., et al., Gene: a gene-centered information resource at NCBI, *Nucleic acids research*, *43*, D36–D42, doi:10.1093/nar/gku1055, 2015.

Brunetti-Pierri, N., et al., Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities, *Nature Genetics*, *40*(12), 1466–1471, doi:10.1038/ng.279, 2008.

Burma, S., and D. J. Chen, Role of DNA-PK in the cellular response to DNA double-strand breaks, *DNA Repair*, *3*, 909–918, 2004.

Carlson, M., B. C. Osmond, and D. Botstein, Mutants of yeast defective in sucrose utilization, *Genetics*, *98*(1), 25–40, 1981.

Carpenter, A. E., et al., CellProfiler: image analysis software for identifying and quantifying cell phenotypes, *Genome biology*, *7*(10), R100, doi:10.1186/gb-2006-7-10-r100, 2006.

Carrero, G., D. McDonald, E. Crawford, G. de Vries, and M. J. Hendzel, Using FRAP and mathematical modeling to determine the in vivo kinetics of nuclear proteins, *Methods*, *29*(1), 14–28, doi:10.1016/S1046-2023(02)00288-8, 2003.

Celeste, A., O. Fernandez-Capetillo, M. J. Kruhlak, D. R. Pilch, D. W. Staudt, A. Lee, R. F. Bonner, W. M. Bonner, and A. Nussenzweig, Histone H2AX phosphorylation is dispensable for the initial recognition of DNA breaks, *Nature Cell Biology*, *5*(7), 675–679, 2003a.

Celeste, A., et al., Genomic instability in mice lacking histone H2AX, *Science*, *296*, 922–927, 2002.

Celeste, A., et al., H2AX haploinsufficiency modifies genomic stability and tumor susceptibility, *Cell*, *114*, 371–383, 2003b.

Chalfie, M., Y. Tu, G. Euskirchen, W. W. Ward, and D. C. Prasher, Green Fluorescent Protein as a marker for gene expression, *Science*, *263*(5148), 802–805, doi:10.1126/science.8303295, 1994.

Chew, Y. C., G. Camporeale, N. Kothapalli, G. Sarath, and J. Zempleni, Lysine residues in N-terminal and C-terminal regions of human histone H2A are targets for biotinylation by biotinidase, *Journal of Nutritional Biochemistry*, *17*, 225–233, 2006.

Chowdhury, D., M.-C. Keogh, H. Ishii, C. L. Peterson, S. Buratowski, and J. Lieberman, gamma-H2AX dephosphorylation by protein phosphatase 2A facilitates DNA double-strand break repair, *Molecular Cell*, *20*, 801–809, 2005.

Chowdhury, D., et al., A PP4-phosphatase complex dephosphorylates gamma-H2AX generated during DNA replication, *Molecular Cell*, *31*(1), 33–46, 2008.

Chuang, C.-H., A. E. Carpenter, B. Fuchsova, T. Johnson, P. de Lanerolle, and A. S. Belmont, Long-range directional movement of an interphase chromosome site, *Current Biology*, *16*(8), 825–831, doi:10.1016/j.cub.2006.03.059, 2006.

Cock, P. J., et al., Biopython: freely available Python tools for computational molecular biology and bioinformatics, *Bioinformatics*, *25*(11), 1422–1423, doi:10.1093/bioinformatics/btp163, 2009.

Collberg, C., and T. A. Proebsting, Repeatability in computer systems research, *Communications of the ACM*, *59*(3), 62–69, doi:10.1145/2812803, 2016.

Collins, F., Researching the researchers, *Nature Genetics*, *46*(5), 417–418, doi:10.1038/ng.2972, 2014.

Cook, P. J., B. G. Ju, F. Telese, X. Wang, C. K. Glass, and M. G. Rosenfeld, Tyrosine dephosphorylation of H2AX modulates apoptosis and survival decisions, *Nature*, *458*, 591–596, 2009.

Cowell, I. G., N. J. Sunter, P. B. Singh, C. A. Austin, B. W. Durkacz, and M. J. Tilby, Gamma-H2AX foci form preferentially in euchromatin after ionising-radiation, *PLoS ONE*, *10*, 2007.

Cremer, T., M. Cremer, S. Dietzel, S. Müller, I. Solovei, and S. Fakan, Chromosome territories — a functional nuclear landscape, *Current opinion in cell biology*, *18*(3), 307–316, doi:10.1016/j.ceb.2006.04.007, 2006.

Crooks, G., G. Hon, J. Chandonia, and S. Brenner, WebLogo: a sequence logo generator, *Genome Research*, *14*(6), 1188–1190, doi:10.1101/gr.849004, 2004.

D'Anna Jr, J. A., and I. Isenberg, Interactions of histone LAK (f2a2) with histones KAS (f2b) and GRK (f2a1), *Biochemistry*, *13*(10), 2098–2104, doi:10.1021/bi00707a016, 1974.

Davey, C. A., D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond, Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution, *Journal of molecular biology*, *319*(5), 1097–1113, doi:10.1016/S0022-2836(02)00386-8, 2002.

De Boni, U., and A. H. Mintz, Curvilinear, three-dimensional motion of chromatin domains and nucleoli in neuronal interphase nuclei, *Science*, *234*, 863–867, 1986.

DeLano, W., The PyMOL Molecular Graphics System, on World Wide Web http://www.pymol.org, 2002.

DeLisi, C., Computers in molecular biology: current applications and emerging trends, *Science*, *240*(4848), 47–52, doi:10.1126/science.3281255, 1988.

den Dunnen, J., and S. Antonarakis, Mutation nomenclature, in *Current Protocols in Human Genetics*, chap. 37, pp. 7.13.1–7.13.8, Springer-Verlag, doi:10.1002/0471142905.hg0713s37, 2003.

den Dunnen, J. T., et al., HGVS recommendations for the description of sequence variants: 2016 update, *Human Mutation*, *37*(5), 00,000, doi:10.1002/humu.22981, 2016.

Dey, A., J. Ellenberg, A. Farina, A. E. Coleman, T. Maruyama, S. Sciortino, J. Lippincott-Schwartz, and K. Ozato, A bromodomain protein, MCAP, associates with mitotic chromosomes and affects G_2-to-M transition, *Molecular and cellular biology*, *20*(17), 6537–6549, doi:10.1128/MCB.20.17.6537-6549.2000, 2000.

Difilippantonio, S., et al., 53BP1 facilitates long-range DNA end-joining during V(D)J recombination, *Nature*, *456*, 529–533, 2008.

Digman, M. A., C. M. Brown, P. Sengupta, P. W. Wiseman, A. R. Horwitz, and E. Gratton, Measuring fast dynamics in solutions and cells with a laser scanning microscope, *Biophysical journal*, *89*(2), 1317–1327, doi: 10.1529/biophysj.105.062836, 2005.

Dimitrova, N., Y.-C. M. Chen, D. L. Spector, and T. d. Lange, 53BP1 promotes non-homologous end joining of telomeres by increasing chromatin mobility, *Nature*, *456*, 524–528, 2008.

Dixon, J. R., D. U. Gorkin, and B. Ren, Chromatin domains: the unit of chromosome organization, *Molecular cell*, *62*(5), 668–680, doi:10.1016/j.molcel.2016.05.018, 2016.

Downs, J. A., N. F. Lowndes, and S. P. Jackson, A role for saccharomyces cerevisiae histone H2A in DNA repair, *Nature*, *408*, 1001–1004, 2000.

Downs, J. A., M. C. Nussenzweig, and A. Nussenzweig, Chromatin dynamics and the preservation of genetic information, *Nature*, *447*, 951–958, 2007.

Draizen, E. J., A. K. Shaytan, L. Mariño-Ramírez, P. B. Talbert, D. Landsman, and A. R. Panchenko, HistoneDB 2.0: a histone database with variants — an integrated resource to explore histones and their variants, *Database*, *2016*, baw014, doi:10.1093/database/baw014, 2016.

Du, L.-L., B. A. Moser, and P. Russell, Homo-oligomerization is the essential function of the tandem BRCT domains in the checkpoint protein Crb2, *The Journal of Biological Chemistry*, *279*(37), 38,409–38,414, 2004.

Döring, V., Der Zellzyklus-abhängige Aufbau des humanen Kinetochors, Ph.D. thesis, Friedrich-Schiller-Universität Jena, 2012.

Eaton, J. W., D. Bateman, S. Hauberg, and R. Wehbring, *GNU Octave version 4.2.0 manual: a high-level interactive language for numerical computations*, 2016.

Edelmann, P., H. Bornfleth, D. Zink, T. Cremer, and C. Cremer, Morphology and dynamics of chromosome territories in living cells, *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, *1551*(1), M29–M39, doi: 10.1016/S0304-419X(01)00023-3, 2001.

Ederveen, H., I. Mandemaker, and L. C., The human histone H3 complement anno 2011, *Biochimica et Biophyica Acta*, *1809*, 577–586, doi: 10.1016/j.bbagrm.2011.07.002, 2011.

Eirin-Lopez, J., R. González-Romero, D. Dryhurst, J. Méndez, , and J. Ausió, Long-term evolution of histone families: old notions and new insights into their diversification mechanisms across eukaryotes, in *Evolutionary Biology: Concept, Modeling, and Application*, edited by P. Pontarotti, chap. 9, pp. 139–162, Springer-Verlag, doi:10.1007/978-3-642-00952-5_8, 2009.

Elgin, S., and H. Weintraub, Chromosomal proteins and chromatin structure, *Annual Review of Biochemistry*, *44*, 725–774, doi:10.1146/annurev.bi.44.070175.003453, 1975.

Eliezer, Y., L. Argaman, A. Rhie, A. J. Doherty, and M. Goldberg, The direct interaction between 53BP1 and MDC1 is required for the recruitment of 53BP1 to sites of damage, *Journal of Biological Chemistry*, *284*(1), 426–435, 2009.

Essers, J., A. F. Theil, C. Baldeyron, W. A. van Cappellen, A. B. Houtsmuller, R. Kanaar, and W. Vermeulen, Nuclear dynamics of PCNA in DNA replication and repair, *Molecular and cellular biology*, *25*(21), 9350–9359, doi:10.1128/MCB.25.21.9350-9359.2005, 2005a.

Essers, J., W. A. van Cappellen, A. F. Theil, E. van Drunen, N. G. Jaspers, J. H. Hoeijmakers, C. Wyman, W. Vermeulen, and R. Kanaar, Dynamics of relative chromosome position during the cell cycle, *Molecular biology of the cell*, *16*(2), 769–775, 2005b.

Felsenfeld, G., and M. Groudine, Controlling the double helix, *Nature*, *421*(6921), 448–453, doi:10.1038/nature01411, 2003.

Fernandez-Capetillo, O., et al., DNA damage-induced G2-M checkpoint activation by histone H2AX and 53BP1, *Nature Cell Biology*, *4*, 993–997, 2002.

Fink, M., D. Imholz, and F. Thoma, Contribution of the serine 129 of histone H2A to chromatin structure, *Molecular and Cellular Biology*, *27*(10), 3589–3600, 2007.

Flaus, A., and T. Owen-Hughes, Mechanisms for ATP-dependent chromatin remodelling: the means to the end, *Febs Journal*, *278*(19), 3579–3595, doi:10.1111/j.1742-4658.2011.08281.x, 2011.

Flaus, A., C. Rencurel, H. Ferreira, N. Wiechens, and T. Owen-Hughes, Sin mutations alter inherent nucleosome mobility, *The EMBO journal*, *23*(2), 343–353, doi:10.1038/sj.emboj.7600047, 2004.

Flaus, A., D. M. Martin, G. J. Barton, and T. Owen-Hughes, Identification of multiple distinct Snf2 subfamilies with conserved structural motifs, *Nucleic Acids Research*, *34*(10), 2887–2905, doi:10.1093/nar/gkl295, 2006.

Flemming, W., Beiträge zur Kenntniss der Zelle und Ihrer Lebenserscheinungen, *Archiv für mikroskopische Anatomie*, *18*, 151–259, doi:10.1007/BF02952594, 1880.

Fomel, S., and G. Hennenfent, Reproducible computational experiments using SCons, in *International Conference on Acoustics, Speech and Signal Processing 2007*, vol. 4, pp. 1257–1260, IEEE, doi:10.1109/ICASSP.2007.367305, 2007.

Franklin, S., and A. Zweidler, Non-allelic variants of histones 2a, 2b and 3 in mammals, *Nature*, *266*(5599), 273–275, doi:10.1038/266273a0, 1977.

Free Software Foundation, What is free software?, `https://www.gnu.org/philosophy/free-sw.en.html`, version 1.141, 2015.

Fussner, E., R. W. Ching, and D. P. Bazett-Jones, Living without 30nm chromatin fibers, *Trends in biochemical sciences*, *36*(1), 1–6, doi:10.1016/j.tibs.2010.09.002, 2011.

Gentleman, R., Reproducible research: a bioinformatics case study, *Statistical Applications in Genetics and Molecular Biology*, *4*(1), doi:10.2202/1544-6115.1034, 2005.

Georgiev, O., and M. L. Birnstiel, The conserved CAAGAAAGA spacer sequence is an essential element for the formation of 3′ termini of the sea urchin H3 histone mRNA by RNA processing, *The EMBO Journal*, *4*(2), 481, 1985.

Glasbey, C. A., An analysis of histogram-based thresholding algorithms, *CVGIP: Graphical models and image processing*, *55*(6), 532–537, doi:10.1006/cgip.1993.1040, 1993.

GMD Executive Editors, Editorial: The publication of geoscientific model developments v1.0, *Geoscientific Model Development*, *6*(4), 1233–1242, doi:10.5194/gmd-6-1233-2013, 2013.

Goldman, N., and Z. Yang, A codon-based model of nucleotide substitution for protein-coding DNA sequences, *Molecular biology and evolution*, *11*(5), 725–736, 1994.

Goodarzi, A. A., A. T. Noon, D. Deckbar, Y. Ziv, Y. Shiloh, M. Löbrich, and P. A. Jeggo, ATM signaling facilitates repair of DNA double-strand breaks associated with heterochromatin, *Molecular Cell*, *31*, 167–177, 2008.

Görisch, S. M., M. Wachsmuth, K. F. Tóth, P. Lichter, and K. Rippe, Histone acetylation increases chromatin accessibility, *Journal of cell science*, *118*(24), 5825–5834, doi:10.1242/jcs.02689, 2005.

Govin, J., et al., Pericentric heterochromatin reprogramming by new histone variants during mouse spermiogenesis, *Journal of Cell Biology*, *176*, 283–294, doi:10.1083/jcb.200604141, 2007.

Gray, K. A., L. C. Daugherty, S. M. Gordon, R. L. Seal, M. W. Wright, and E. A. Bruford, Genenames.org: the HGNC resources in 2013, *Nucleic Acids Research*, *41*(D1), D545–D552, doi:10.1093/nar/gks1066, 2013.

Gray, K. A., B. Yates, R. L. Seal, M. W. Wright, and E. A. Bruford, Genenames.org: the HGNC resources in 2015, *Nucleic Acids Research*, *43*(D1), D1079–D1085, doi:10.1093/nar/gku1071, 2015.

Grenon, M., T. Costelloe, S. Jimeno, A. O'Shaughnessy, J. FitzGerald, O. Zgheib, L. Degerth, and N. F. Lowndes, Docking onto chromatin via the Saccharomyces cerevisiae Rad9 Tudor domain, *Yeast*, *24*, 105–119, 2007.

Gruen, J., et al., A transcription map of the major histocompatibility complex (MHC) class I region, *Genomics*, *36*, 70–85, doi:10.1006/geno.1996.0427, 1996.

Gruen, J. R., and S. M. Weissman, Evolving views of the major histocompatibility complex, *Blood*, *90*(11), 4252–4265, 1997.

Hammet, A., C. Magill, J. Heierhorst, and S. P. Jackson, Rad9 BRCT domain interaction with phosphorylated H2AX regulates the G1 checkpoint in budding yeast, *EMBO Reports*, *8*(9), 851–857, 2007.

Harlow, E., and D. Lane, *Antibodies: a laboratory manual*, pp. 736–739, CSHL Press, 1988.

Harris, M., R. Böhni, M. Schneiderman, L. Ramamurthy, D. Schümperli, and W. Marzluff, Regulation of histone mRNA in the unperturbed cell cycle: evidence suggesting control at two posttranscriptional steps, *Molecular and cellular biology*, *11*(5), 2416–2424, doi:10.1128/MCB.11.5.2416, 1991.

Harshman, S. W., N. L. Young, M. R. Parthun, and M. A. Freitas, H1 histones: current perspectives and challenges, *Nucleic acids research*, *41*(21), 9593–9609, doi:doi:10.1093/nar/gkt700, 2013.

Hebert, B., S. Costantino, and P. W. Wiseman, Spatiotemporal image correlation spectroscopy (STICS) theory, verification, and application to protein velocity mapping in living CHO cells, *Biophysical journal*, *88*(5), 3601–3614, doi:10.1529/biophysj.104.054874, 2005.

Heo, K., H. Kim, S. H. Choi, J. Choi, K. Kim, J. Gu, M. R. Lieber, A. S. Yang, and W. An, FACT-mediated exchange of histone variant H2AX regulated by phosphorylation of H2AX and ADP-Ribosylation of Spt16, *Molecular Cell*, *30*, 86–97, 2008.

Heun, P., T. Laroche, K. Shimada, P. Furrer, and S. M. Gasser, Chromosome dynamics in the yeast interphase nucleus, *Science*, *294*(5549), 2181–2186, doi:10.1126/science.1065366, 2001.

Hewish, D. R., and L. A. Burgoyne, Chromatin sub-structure. the digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease, *Biochemical and biophysical research communications*, *52*(2), 504–510, doi:10.1016/0006-291X(73)90740-7, 1973.

Hoefig, K., and V. Heissmeyer, Degradation of oligouridylated histone mRNAs: see UUUUU and goodbye, *Wiley Interdisciplinary Reviews RNA*, *5*(4), 577–589, doi:10.1002/wrna.1232, 2014.

Holliday, G., et al., Key challenges for the creation and maintenance of specialist protein resources, *Proteins: Structure, funcation and bioinformatics*, *83*, 1005–1013, doi:10.1002/prot.24803, 2015.

Hübner, M. R., and D. L. Spector, Chromatin dynamics, *Annual review of biophysics*, *39*, 471–489, doi:10.1146/annurev.biophys.093008.131348, 2010.

Huh, N.-E., I. Hwang, K. Lim, K.-H. You, and C.-B. Chae, Presence of a bi-directional S phase-specific transcription regulatory element in the promoter shared by testis-specific TH2A and TH2B histone genes, *Nucleic Acids Research*, *19*(1), 93–98, doi:10.1093/nar/19.1.93, 1991.

Huiskamp, W., Beiträge zur Kenntnis des Thymusnucleohistons, *Hoppe-Seyler's Zeitschrift für physiologische Chemie*, *39*, 55–72, doi:10.1515/bchm2.1903.39.1.55, 1903.

Hurley, D. G., D. M. Budden, and E. J. Crampin, Virtual reference environments: a simple way to make research reproducible, *Briefings in bioinformatics*, *16*(5), 901–903, doi:10.1093/bib/bbu043, 2015.

Huyen, Y., et al., Methylated lysine 79 of histone H3 targets 53BP1 to DNA double-strand breaks, *Nature*, *432*(7015), 406–411, 2004.

Ikura, T., et al., DNA damage-dependent acetylation and ubiquitination of H2AX enhances chromatin dynamics, *Molecular and Cellular Biology*, *27*(20), 7028–7040, 2007.

Ince, D. C., L. Hatton, and J. Graham-Cumming, The case for open computer programs, *Nature*, *482*(7386), 485–488, doi:10.1038/nature10836, 2012.

Ioannidis, J. P., D. Fanelli, D. D. Dunne, and S. N. Goodman, Meta-research: evaluation and improvement of research methods and practices, *PLoS Biol*, *13*(10), e1002,264, doi:10.1371/journal.pbio.1002264, 2015.

Ioannidis, J. P., et al., Repeatability of published microarray gene expression analyses, *Nature genetics*, *41*(2), 149–155, doi:10.1038/ng.295, 2009.

Ivanova, V. S., C. L. Hatch, and W. M. Bonner, Characterization of the human histone H2A.X gene. Comparison of its promoter with other H2A gene promoters, *The Journal of Biological Chemistry*, *269*(39), 24,189–24,194, 1994a.

Ivanova, V. S., D. Zimonjic, N. Popescu, and W. M. Bonner, Chromosomal localization of the human histone H2A.X gene to 11q23.2-q23.3 by fluorescence in situ hybridization, *Human Genetics*, *94*(3), 303–306, 1994b.

Jenuwein, T., and C. D. Allis, Translating the histone code, *Science*, *293*(5532), 1074–1080, doi:10.1126/science.1063127, 2001.

Jullien, D., et al., Chromatibody, a novel non-invasive molecular tool to explore and manipulate chromatin in living cells, *J Cell Sci*, *129*(13), 2673–2683, doi:10.1242/jcs.183103, 2016.

Kaczmarczyk, A., Functional role of CCAN histone fold proteins in mitosis, Ph.D. thesis, National University of Ireland, Galway, 2012.

Kanda, T., K. F. Sullivan, and G. M. Wahl, Histone–GFP fusion protein enables sensitive analysis of chromosome dynamics in living mammalian cells, *Current Biology*, *8*(7), 377–385, doi:10.1016/S0960-9822(98) 70156-3, 1998.

Kennani, S. E., A. Adrait, A. K. Shaytan, S. Khochbin, C. Bruley, A. R. Panchenko, D. Landsman, D. Pflieger, and J. Govin, MS_HistoneDB, a manually curated resource for proteomic analysis of human and mouse histones, *Epigenetics and Chromatin*, *10*, 2, doi: 10.1186/s13072-016-0109-x, 2017.

Keogh, M.-C., et al., A phosphatase complex that dephosphorylates gamma-H2AX regulates DNA damage checkpoint recovery, *Nature*, *439*, 497–501, 2006.

Khare, S. P., F. Habib, R. Sharma, N. Gadewal, S. Gupta, and S. Galande, HIstome - a relational knowledgebase of human histone proteins and histone modifying enzymes, *Nucleic Acids Research*, *40*, D337–342, doi: 10.1093/nar/gkr1125, 2012.

Kilkenny, M. L., A. S. Doré, S. M. Roe, K. Nestoras, J. C. Ho, F. Z. Watts, and L. H. Pearl, Structural and functional analysis of the Crb2-BRCT2 domain reveals distinct roles in checkpoint signaling and DNA damage repair, *Genes & Development*, *22*, 2034–2047, 2008.

Kim, J.-S., T. B. Krasieva, V. LaMorte, A. M. R. Taylor, and K. Yokomori, Specific recruitment of human cohesin to laser-induced DNA damage, *Journal of Biological Chemistry*, *277*(47), 45,149–45,153, doi:10.1074/jbc.M209123200, 2002.

Kimura, H., and P. R. Cook, Kinetics of core histones in living human cells: little exchange of H3 and H4 and some rapid exchange of H2B, *The Journal of cell biology*, *153*(7), 1341–1353, doi:10.1083/jcb.153.7.1341, 2001.

Kimura, H., et al., A novel histone exchange factor, protein phosphatase 2C gamma, mediates the exchange and dephosphorylation of H2A-H2B, *The Journal of Cell Biology*, *175*(3), 389–400, 2006.

Knight, S., Building software with SCons, *Computing in Science & Engineering*, *7*(1), 79–88, doi:10.1109/MCSE.2005.11, 2005.

Kornberg, R. D., Chromatin structure: a repeating unit of histones and DNA, *Science*, *184*(4139), 868–871, doi:10.1126/science.184.4139.868, 1974.

Kornberg, R. D., and J. O. Thomas, Chromatin structure: oligomers of the histones, *Science*, *184*(4139), 865–868, doi:10.1126/science.184.4139.865, 1974.

Kossel, A., Ueber einen peptonartigen Bestandtheil des Zellkerns, *Hoppe-Seyler's Zeitschrift für physiologische Chemie*, *8*, 511–515, doi:10.1515/bchm1.1884.8.6.511, 1884.

Kouzarides, T., Chromatin modifications and their function, *Cell*, *128*(4), 693–705, doi:10.1016/j.cell.2007.02.005, 2007.

Kowalski, A., and J. Pałyga, Chromatin compaction in terminally differentiated avian blood cells: the role of linker histone H5 and non-histone protein MENT, *Chromosome Research*, *19*(5), 579–590, doi:10.1007/s10577-011-9218-3, 2011.

Kraft, D., A level-set framework for shape optimisation, Ph.D. thesis, University of Graz, 2015.

Kruger, W., C. L. Peterson, A. Sil, C. Coburn, G. Arents, E. N. Moudrianakis, and I. Herskowitz, Amino acid substitutions in the structured domains of histones H3 and H4 partially relieve the requirement of the yeast SWI/SNF complex for transcription., *Genes & Development*, *9*(22), 2770–2779, doi:10.1101/gad.9.22.2770, 1995.

Kruhlak, M. J., M. A. Lever, W. Fischle, E. Verdin, D. P. Bazett-Jones, and M. J. Hendzel, Reduced mobility of the alternate splicing factor (ASF) through the nucleoplasm and steady state speckle compartments, *J Cell Biol*, *150*(1), 41–52, doi:10.1083/jcb.150.1.41, 2000.

Kruhlak, M. J., A. Celeste, G. Dellaire, O. Fernandez-Capetillo, W. G. Müller, J. G. McNally, D. P. Bazett-Jones, and A. Nussenzweig, Changes in chromatin structure and mobility in living cells at sites of DNA double-strand breaks, *The Journal of Cell Biology*, *172*(6), 823–834, 2006.

Kuipers, M. A., T. J. Stasevich, T. Sasaki, K. A. Wilson, K. L. Hazelwood, J. G. McNally, M. W. Davidson, and D. M. Gilbert, Highly stable loading of Mcm proteins onto chromatin in living cells requires replication to unload, *The Journal of cell biology*, *192*(1), 29–41, doi: 10.1083/jcb.201007111, 2011.

Kurumizaka, H., N. Horikoshi, H. Tachiwana, and W. Kagawa, Current progress on structural studies of nucleosomes containing histone H3 variants, *Current Opinion in Structural Biology*, *23*(1), 109–115, doi:10.1016/j.sbi.2012.10.009, 2013.

Kuznetsova, M. A., and E. V. Sheval, Chromatin fibers: from classical descriptions to modern interpretation, *Cell Biology International*, doi: 10.1002/cbin.10672, 2016.

Lander, E. S., et al., Initial sequencing and analysis of the human genome, *Nature*, *409*(6822), 860–921, doi:10.1038/35057062, 2001.

Lee, M. S., R. A. Edwards, G. L. Thede, and J. N. M. Glover, Structure of the BRCT repeat domain of MDC1 and its specificity for the free COOH-terminal end of the gamma-H2AX histone tail, *Journal of Biological Chemistry*, *280*(37), 32,053–32,056, 2005.

Lemberger, T., Image data in need of a home, *Molecular Systems Biology*, *11*, 1–2, doi:10.15252/msb.20156719, 2015.

Lengauer, C., K. W. Kinzler, and B. Vogelstein, Genetic instabilities in human cancers, *Nature*, *396*, 643–649, 1998.

Li, A., A. Maffey, W. Abbott, N. Conde e Silva, A. Prunell, J. Siino, D. Churikov, A. Zalensky, and J. Ausio, Characterization of nucleosomes consisting of the human testis/sperm-specific histone H2B variant (hTSH2B), *Biochemistry*, *44*(7), 2529–2535, doi:10.1021/bi048061n, 2005.

Linkert, M., et al., Metadata matters: access to image data in the real world, *The Journal of cell biology*, *189*(5), 777–782, doi:10.1083/jcb.201004104, 2010.

Lodish, H., A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, M. P. Scott, S. L. Zipursky, and J. E. Darnell, *Molecular cell biology*, 5 ed., W.H.Freeman, 2003.

Lorén, N., et al., Fluorescence recovery after photobleaching in material and life sciences: putting theory into practice, *Quarterly reviews of biophysics*, *48*(03), 323–387, doi:10.1017/S0033583515000013, 2015.

Luger, K., A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, Crystal structure of the nucleosome core particle at 2.8 Å resolution, *Nature*, *389*(6648), 251–260, doi:10.1038/38444, 1997.

Luger, K., M. L. Dechassa, and D. J. Tremethick, New insights into nucleosome and chromatin structure: an ordered state or a disordered affair?, *Nature reviews Molecular cell biology*, *13*(7), 436–447, doi:10.1038/nrm3382, 2012.

Lukas, C., et al., Mdc1 couples DNA double-strand break recognition by Nbs1 with its H2AX-dependent chromatin retention, *The EMBO Journal*, *23*, 2674–2683, 2004.

Madigan, J. P., H. L. Chotkowski, and R. L. Glaser, DNA double-strand break-induced phosphorylation of Drosophila histone variant H2Av helps prevent radiation-induced apoptosis, *Nucleic Acids Research*, *30*(17), 3698–3705, 2002.

Makalowska, I., E. S. Ferlanti, A. D. Baxevanis, and D. Landsman, Histone sequence database: sequences, structures, post-translational modifications and genetic loci, *Nucleic Acids Research*, *27*, 323–324, doi:10.1093/nar/27.1.323, 1999.

Malik, H. S., and S. Henikoff, Phylogenomics of the nucleosome, *Nature Structural Biology*, *10*(11), 882–891, 2003.

Mannironi, C., W. M. Bonner, and C. L. Hatch, H2A.X. a histone isoprotein with a conserved C-terminal sequence, is encoded by a novel mRNA with both DNA replication type and polyA 3′ processing signals, *Nucleic Acids Research*, *17*(22), 9113–9126, doi:10.1093/nar/17.22.9113, 1989.

Mari, P.-O., et al., Dynamic assembly of end-joining complexes requires interaction between Ku70/80 and XRCC4, *Proceedings of the National Academy of Sciences*, *103*(49), 18,597–18,602, doi:10.1073/pnas.0609061103, 2006.

Mariño-Ramérez, L., B. Hsu, A. D. Baxevanis, and D. Landsman, The histone database: a comprehensive resource for histones and histone fold-containing proteins, *Proteins*, *2006*, 838–842, doi:10.1002/prot.20814, 2006.

Mariño-Ramérez, L., K. Levine, M. Morales, S. Zhang, R. Moreland, A. D. Baxevanis, and D. Landsman, The histone database: an integrated resource for histones and histone fold-containing proteins, *Database (Oxford)*, *2011*, bar048, doi:10.1093/database/bar048, 2011.

Mariño-Ramírez, L., K. M. Levine, M. Morales, S. Zhang, R. T. Moreland, A. D. Baxevanis, and D. Landsman, The histone database: an integrated resource for histones and histone fold-containing proteins, *Database*, *2011*, bar048, doi:10.1093/database/bar048, 2011.

Marshall, W., A. Straight, J. Marko, J. Swedlow, A. Dernburg, A. Belmont, A. Murray, D. Agard, and J. Sedat, Interphase chromosomes undergo constrained diffusional motion in living cells, *Current Biology*, *7*(12), 930–939, doi:10.1016/S0960-9822(06)00412-X, 1997.

Marx, V., Biology: The big challenges of big data, *Nature*, *498*(7453), 255–260, doi:10.1038/498255a, 2013.

Marzluff, W., P. Gongidi, K. Woods, J. Jin, and L. Maltais, The human and mouse replication-dependent histone genes, *Genomics*, *80*(5), 487–498, doi:10.1006/geno.2002.6850, 2002.

Marzluff, W., E. Wagner, and D. R.J., Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail, *Nature Reviews Genetics*, *9*(11), 843–854, doi:10.1038/nrg2438, 2008.

Maze, I., K. Noh, A. Soshnev, and C. Allis, Every amino acid matters: essential contributions of histone variants to mammalian development and disease, *Nature Reviews Genetics*, *15*(4), 259–271, doi: 10.1038/nrg3673, 2014.

McEntyre, J., U. Sarkans, and A. Brazma, The biostudies database, *Molecular systems biology*, *11*(12), 847, doi:10.15252/msb.20156658, 2015.

McGinty, R. K., and S. Tan, Histone, nucleosome, and chromatin structure, in *Fundamentals of Chromatin*, pp. 1–28, doi:10.1021/cr500373h, 2014.

McKinney, S. A., C. S. Murphy, K. L. Hazelwood, M. W. Davidson, and L. L. Looger, A bright and photostable photoconvertible fluorescent protein for fusion tags, *Nature methods*, *6*(2), 131–133, doi: 10.1038/nmeth.1296, 2009.

McKinnon, P. J., and K. W. Caldecott, DNA strand break repair and human genetic disease, *Annual Review of Genomics and Human Genetics*, *8*, 35–55, 2007.

Miescher, F., Ueber die chemische Zusammensetzung der Eiterzellen, *Medicinisch-chemische Untersuchungen*, *4*, 441–460, 1871.

Miescher, F., Das Protamin, eine neue organische Base aus den Samenfäden des Rheinlachses, *Berichte der Deutschen Chemischen Gesellschaft*, *7*(1), 376–379, doi:10.1002/cber.187400701119, 1874.

Molden, R., N. Bhanu, G. LeRoy, A. Arnaudo, and G. B.A., Multifaceted quantitative proteomics analysis of histone H2B isoforms and their modifications, *Epigenetics and Chromatin*, *8*, 15, doi:10.1186/s13072-015-0006-8, 2015.

Monen, J., P. S. Maddox, F. Hyndman, K. Oegema, and A. Desai, Differential role of CENP-A in the segregation of holocentric C. elegans chromosomes during meiosis and mitosis, *Nature Cell Biology*, *7*, 1248–1255, 2005.

Morgenstern, B., DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment, *Bioinformatics*, *15*(3), 211–218, 1999.

Mueller, F., P. Wach, and J. G. McNally, Evidence for a common mode of transcription factor interaction with chromatin as revealed by improved quantitative fluorescence recovery after photobleaching, *Biophysical journal*, *94*(8), 3323–3339, doi:10.1529/biophysj.107.123182, 2008.

Mueller, F., D. Mazza, T. J. Stasevich, and J. G. McNally, FRAP and kinetic modeling in the analysis of nuclear protein dynamics: what do we really know?, *Current opinion in cell biology*, *22*(3), 403–411, doi:10.1016/j.ceb.2010.03.002, 2010.

Muster, B., A. Rapp, and M. C. Cardoso, Systematic analysis of DNA damage induction and DNA repair pathway activation by continuous wave visible light laser micro-irradiation, *AIMS Genetics*, *4*(1), 47–68, doi:10.3934/genet.2017.1.47, 2017.

Muthurajan, U. M., Y. Bao, L. J. Forsberg, R. S. Edayathumangalam, P. N. Dyer, C. L. White, and K. Luger, Crystal structures of histone Sin mutant nucleosomes reveal altered protein–DNA interactions, *The EMBO journal*, *23*(2), 260–271, doi:10.1038/sj.emboj.7600046, 2004.

Nakamura, Y., T. Gojobori, and T. Ikemura, Codon usage tabulated from international DNA sequence databases: status for the year 2000, *Nucleic Acids Research*, *28*(1), 292, 2000.

Nature Editorial, 2012, Error prone, Editorial, doi:10.1038/487406a, 2012.

Nature Editorial, 2014, Code share, Editorial, doi:10.1038/514536a, 2014.

Nei, M., and A. Rooney, Concerted and birth-and-death evolution of multigene families, *Annual Review of Genetics*, *39*, 121–152, doi:10.1146/annurev.genet.39.073003.112240, 2005.

Neigeborn, L., and M. Carlson, Genes affecting the regulation of SUC2 gene expression by glucose repression in saccharomyces cerevisiae, *Genetics*, *108*(4), 845–858, 1984.

Notredame, C., D. G. Higgins, and J. Heringa, T-Coffee: a novel method for fast and accurate multiple sequence alignment, *Journal of Molecular Biology*, *302*(1), 205–218, doi:10.1006/jmbi.2000.4042, 2000.

O'Leary, N. A., et al., Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, *Nucleic Acids Research*, *44*, D733–745, doi:10.1093/nar/gkv1189, 2016.

Olenych, S. G., N. S. Claxton, G. K. Ottenberg, and M. W. Davidson, *The fluorescent protein color palette*, chap. 21.5, John Wiley, doi:10.1002/0471143030.cb2105s36, 2007.

Olins, A. L., and D. E. Olins, Spheroid chromatin units (ν bodies), *Science*, *183*(4122), 330–332, doi:10.1126/science.183.4122.330, 1974.

Open Science Collaboration, Estimating the reproducibility of psychological science, *Science*, *349*(6251), aac4716, doi:10.1126/science.aac4716, 2015.

Osborne, C. S., et al., Active genes dynamically colocalize to shared sites of ongoing transcription, *Nature genetics*, *36*(10), 1065–1071, doi:10.1038/ng1423, 2004.

Osley, M., The regulation of histone synthesis in the cell cycle, *Annual review of biochemistry*, *60*(1), 827–861, doi:10.1146/annurev.bi.60.070191.004143, 1991.

Otsu, N., A threshold selection method from gray-level histograms, *IEEE transactions on Systems, Man, and Cybernetics*, *9*(1), 62–66, doi:10.1109/TSMC.1979.4310076, 1979.

Padavattana, S., T. Shinagawab, K. Hasegawac, T. Kumasakac, S. Ishiib, and T. Kumarevel, PAML 4: Phylogenetic Analysis by Maximum Likelihood, *Molecular Biology and Evolution*, *24*, 1586–1591, doi:10.1093/molbev/msm088, 2007.

Padavattana, S., T. Shinagawab, K. Hasegawac, T. Kumasakac, S. Ishiib, and T. Kumarevel, Structural and functional analyses of nucleosome complexes with mouse histone variants TH2a and TH2b, involved in reprogramming, *Biochemical and Biophysical Research Communications*, *464*, 929–935, doi:10.1016/j.bbrc.2015.07.070, 2015.

Panier, S., and D. Durocher, Regulatory ubiquitylation in response to DNA double-strand breaks, *DNA repair*, *8*(4), 436–443, 2009.

Pantazis, P., and W. M. Bonner, Quantitative determination of histone modification. H2A acetylation and phosphorylation, *The Journal of Biological Chemistry*, *256*(9), 4669–4675, 1981.

Pardon, J., and M. Wilkins, A super-coil model for nucleohistone, *Journal of molecular biology*, *68*(1), 115–124, doi:10.1016/0022-2836(72)90267-7, 1972.

Parvinen, M., and K.-O. Söderström, Chromosome rotation and formation of synapsis, *Nature*, *260*(5551), 534–535, doi:10.1038/260534a0, 1976.

Patterson, G. H., and J. Lippincott-Schwartz, A photoactivatable GFP for selective photolabeling of proteins and cells, *Science*, *297*(5588), 1873–1877, 2002.

Paull, T. T., E. P. Rogakou, V. Yamazaki, C. U. Kirchgessner, M. Gellert, and W. M. Bonner, A critical role for histone H2AX in recruitment of repair factors to nuclear foci after DNA damage, *Current Biology*, *10*, 886–895, 2000.

Perche, P.-Y., C. Vourc'h, L. Konecny, C. Souchier, M. Robert-Nicoud, S. Dimitrov, and S. Khochbin, Higher concentrations of histone macroH2A in the Barr body are correlated with higher nucleosome density, *Current Biology*, *10*(23), 1531–1534, doi:10.1016/S0960-9822(00)00832-0, 2000.

Phair, R. D., and T. Misteli, High mobility of proteins in the mammalian cell nucleus, *Nature*, *404*(6778), 604–609, doi:10.1038/35007077, 2000.

Phillips, D. M. P., and E. W. Johns, A fractionation of the histones of group F2a from calf thymus, *The Biochemical Journal*, *94*, 127–130, doi:10.1042/bj0940127, 1965.

Pilch, D. R., O. A. Sedelnikova, C. Redon, A. Celeste, A. Nussenzweig, and W. M. Bonner, Characteristics of gamma-H2AX foci at DNA double-strand breaks sites, *Biochemistry and Cell Biology*, *81*, 123–129, 2003.

Pinto, D. M. S., and A. Flaus, Structure and function of histone H2AX, in *Genome Stability and Human Diseases*, pp. 55–78, doi:10.1007/978-90-481-3471-7_4, 2010.

Piontkivska, H., A. Rooney, and M. Nei, Purifying selection and birth-and-death evolution in the histone H4 gene family, *Molecular Biology and Evolution*, *19*(5), 689–697, doi:10.1093/oxfordjournals.molbev.a004127, 2002.

Pirkmajer, S., and A. V. Chibalin, Serum starvation: *caveat emptor*, *American Journal of Physiology-Cell Physiology*, *301*(2), C272–C279, doi:10.1152/ajpcell.00091.2011, 2011.

Prinz, F., T. Schlange, and K. Asadullah, Believe it or not: how much can we rely on published data on potential drug targets?, *Nature reviews Drug discovery*, *10*(9), 712–712, doi:10.1038/nrd3439-c1, 2011.

Prlić, A., et al., BioJava: an open-source framework for bioinformatics in 2012, *Bioinformatics*, *28*(20), 2693–2695, doi:10.1093/bioinformatics/bts494, 2012.

Průša, Z., P. L. Søndergaard, N. Holighaus, C. Wiesmeyr, and P. Balazs, The large time-frequency analysis toolbox 2.0, in *International Symposium on Computer Music Modeling and Retrieval*, pp. 419–442, doi:10.1007/978-3-319-12976-1_25, 2014.

Quinn, N., Regulation and assembly of the constitutive centromere associated network in human cells, Ph.D. thesis, National University of Ireland, Galway, 2011.

R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014.

Rattray, A., and B. Müller, The control of histone gene expression, *Biochemical Society Transactions*, *40*(4), 880–885, doi:10.1042/BST20120065, 2012.

Raymond, E., The cathedral and the bazaar, *Knowledge, Technology & Policy*, *12*(3), 23–49, doi:10.1007/s12130-999-1026-0, 1999.

Redon, C., D. Pilch, E. Rogakou, O. Sedelnikova, K. Newrock, and W. Bonner, Histone H2A variants H2AX and H2AZ, *Current Opinion in Genetics & Development*, *12*, 162–169, 2002.

Reichlin, L. F., Free computer-aided control system design (CACSD) tools for GNU Octave, in *Computer Aided Control System Design (CACSD), 2013 IEEE Conference on*, pp. 334–339, doi:10.1109/CACSD.2013.6663485, 2013.

Rhind, N., and D. M. Gilbert, DNA replication timing, *Cold Spring Harbor perspectives in biology*, *5*(8), a010,132, doi:10.1101/cshperspect.a010132, 2013.

Rice, P., I. Longden, and A. Bleasby, EMBOSS: The European Molecular Biology Open Software Suite, *Trends in Genetics*, *16*(6), 276–277, 2000.

Richmond, T., J. Finch, B. Rushton, D. Rhodes, and A. Klug, Structure of the nucleosome core particle at 7 Å resolution, *Nature*, *311*, 532–537, doi:10.1038/311532a0, 1984.

Rios-Doria, J., A. Velkova, V. Dapic, J. M. Galán-Caridad, V. Dapic, M. A. Carvalho, J. Melendez, and A. N. Monteiro, Ectopic expression of histone H2AX mutants reveals a role for its post-translational modifications, *Cancer Biology & Therapy*, *8*(5), 2009.

Rizzo, M. A., M. W. Davidson, and D. W. Piston, Fluorescent protein tracking and detection: fluorescent protein structure and color variants, *Cold Spring Harbor Protocols*, *2009*(12), pdb–top63, doi:10.1101/pdb.top63, 2009.

Rogakou, E. P., D. R. Pilch, A. H. Orr, V. S. Ivanova, and W. M. Bonner, DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139, *The Journal of Biological Chemistry*, *273*(10), 5858–5868, 1998.

Rogakou, E. P., C. Boon, C. Redon, and W. M. Bonner, Megabase chromatin domains involved in DNA double-strand breaks in vivo, *The Journal of Cell Biology*, *146*(5), 905–915, 1999.

Rooney, A., H. Piontkivska, and M. Nei, Molecular evolution of the nontandemly repeated genes of the histone 3 multigene family, *Molecular Biology and Evolution*, *19*(1), 68–75, doi:10.1093/oxfordjournals.molbev.a003983, 2002.

Rothkamm, K., I. Krüger, L. H. Thompson, and M. Löbrich, Pathways of DNA double-strand break repair during the mammalian cell cycle, *Molecular and Cell Biology*, *23*(16), 5706–5715, 2003.

Sanders, S. L., M. Portoso, J. Mata, J. Bähler, R. C. Allshire, and T. Kouzarides, Methylation of histone H4 lysine 20 controls recruitment of Crb2 to sites of DNA damage, *Cell*, *119*, 603–614, 2004.

Schindelin, J., C. T. Rueden, M. C. Hiner, and K. W. Eliceiri, The ImageJ ecosystem: an open platform for biomedical image analysis, *Molecular reproduction and development*, *82*(7-8), 518–529, doi:10.1002/mrd.22489, 2015.

Schindelin, J., et al., Fiji: an open-source platform for biological-image analysis, *Nature methods*, *9*(7), 676–682, doi:10.1038/nmeth.2019, 2012.

Schneider, C. A., W. S. Rasband, and K. W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis, *Nature methods*, *9*(7), 671–675, doi:10.1038/nmeth.2089, 2012.

Schwab, M., M. Karrenbach, and J. Claerbout, Making scientific computations reproducible, *Computing in Science & Engineering*, *2*(6), 61–67, doi:10.1109/5992.881708, 2000.

Sharp, A. J., et al., Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome, *Nature Genetics*, *38*(9), 1038–1042, doi:10.1038/ng1862, 2006.

Shcherbakova, D. M., P. Sengupta, J. Lippincott-Schwartz, and V. V. Verkhusha, Photocontrollable fluorescent proteins for superresolution imaging, *Annual review of biophysics*, *43*, 303–329, doi:10.1146/annurev-biophys-051013-022836, 2014.

Shechter, D., H. Dormann, C. Allis, and S. Hake, Extraction, purification and analysis of histones, *Nature Protocols*, *2*(6), 1445–1457, doi:10.1038/nprot.2007.202, 2007.

Shelby, R. D., O. Vafa, and K. F. Sullivan, Assembly of CENP–A into centromeric chromatin requires a cooperative array of nucleosomal DNA contact sites, *The Journal of cell biology*, *136*(3), 501–513, 1997.

Shimomura, O., F. H. Johnson, and Y. Saiga, Extraction, purification and properties of Aequorin, a bioluminescent protein from the luminous hydromedusan, *Aequorea*, *Journal of cellular and comparative physiology*, *59*(3), 223–239, doi:10.1002/jcp.1030590302, 1962.

Shinagawa, T., et al., Histone variants enriched in oocytes enhance reprogramming to induced pluripotent stem cells, *Cell Stem Cell*, *14*(2), 217–227, doi:10.1016/j.stem.2013.12.015, 2014a.

Shinagawa, T., et al., Histone variants enriched in oocytes enhance reprogramming to induced pluripotent stem cells, *Cell Stem Cell*, *14*, 217–227, doi:10.1016/j.stem.2013.12.015, 2014b.

Shroff, R., A. Arbel-Eden, D. Pilch, G. Ira, W. M. Bonner, J. H. Petrini, J. E. Haber, and M. Lichten, Distribution and dynamics of chromatin modification induced by a defined DNA double-strand break, *Current Biology*, *14*, 1703–1711, 2004.

Sipos, B., S. Möser, H. Kalthoff, V. Török, M. Löhr, and G. Klöppel, A comprehensive characterization of pancreatic ductal carcinoma cell lines: towards the establishment of an in vitro research platform, *Virchows Archiv*, *442*(5), 444–452, doi:10.1007/s00428-003-0784-4, 2003.

Skok, J. A., R. Gisler, M. Novatchkova, D. Farmer, W. de Laat, and M. Busslinger, Reversible contraction by looping of the Tcra and Tcrb loci in rearranging thymocytes, *Nature immunology*, *8*(4), 378–387, doi:10.1038/ni1448, 2007.

Smith, P. R., I. E. Morrison, K. M. Wilson, N. Fernández, and R. J. Cherry, Anomalous diffusion of major histocompatibility complex class I molecules on HeLa cells determined by single particle tracking, *Biophysical journal*, *76*(6), 3331–3344, doi:10.1016/S0006-3495(99)77486-2, 1999.

Soulier, J., and N. F. Lowndes, The BRCT domain of the S. cerevisiae checkpoint protein Rad9 mediates a Rad9-Rad9 interaction after DNA damage, *Current Biology*, *9*(10), 551–554, 1999.

Sprague, B. L., and J. G. McNally, FRAP analysis of binding: proper and fitting, *Trends in cell biology*, *15*(2), 84–91, doi:10.1016/j.tcb.2004.12.001, 2005.

Stajich, J. E., et al., The Bioperl toolkit: Perl modules for the life sciences, *Genome Research*, *12*(10), 1611–1618, doi:10.1101/gr.361602, 2002.

Stallman, R. M., *Free Software Free Society: Selected Essays of Richard M. Stallman*, chap. 2, pp. 9–25, 3 ed., GNU Press, 2015.

Stedman, E., and E. Stedman, The basic proteins of cell nuclei, *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *235*(630), 565–595, doi:10.1098/rstb.1951.0008, 1951.

Stein, L., Genome annotation: From sequence to biology, *Nature Reveiws Genetics*, *2*, 493–503, doi:10.1038/35080529, 2001.

Stein, L. D., Using GBrowse 2.0 to visualize and share next-generation sequence data, *Briefings in bioinformatics*, *14*(2), 162–171, doi:10.1093/bib/bbt001, 2013.

Stephenson, E., Locomotory invasion of human cervical epithelium and avian fibroblasts by HeLa cells in vitro, *Journal of cell science*, *57*(1), 293–314, 1982.

Stern, M., R. Jensen, and I. Herskowitz, Five SWI genes are required for expression of the HO gene in yeast, *Journal of molecular biology*, *178*(4), 853–868, 1984.

Stixová, L., T. Hruskova, P. Sehnalová, S. Legartová, S. Svidenská, S. Kozubek, and E. Bártová, Advanced microscopy techniques used for comparison of UVA– and γ-irradiation-induced DNA damage in the cell nucleus and nucleolus, *Folia biologica*, *60*(S1), 76, 2014.

Stodden, V., P. Guo, and Z. Ma, Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals, *PloS one*, *8*(6), e67,111, doi:10.1371/journal.pone.0067111, 2013.

Stucki, M., J. A. Clapperton, D. Mohammad, M. B. Yaffe, S. J. Smerdon, and S. P. Jackson, MDC1 directly binds phosphorylated histone H2AX to regulate cellular responses to DNA double-strand breaks, *Cell*, *123*, 1213–1226, 2005.

Sullivan, S., D. W. Sink, K. L. Trout, I. Makalowska, P. Taylor, A. D. Baxevanis, and D. Landsman, The histone database, *Nucleic Acids Research*, *30*, 341–342, doi:10.1093/nar/30.1.341, 2002.

Sullivan, S. A., and D. Landsman, Mining core histone sequences from public protein databases, *Methods in Enzymology*, *2004*, 3–20, doi:10.1016/S0076-6879(03)75001-0, 2004.

Sullivan, S. A., L. Aravind, I. Makalowska, A. D. Baxevanis, and D. Landsman, The histone database: a comprehensive WWW resource for histones and histone fold-containing proteins, *Nucleic Acids Research*, *28*, 320–322, doi:10.1093/nar/28.1.320, 2000.

Sullivan Jr., W. J., A. Naguleswaran, and S. O. Angel, Histones and histone modifications in protozoan parasites, *Cellular Microbiology*, *8*(12), 1850–1861, 2006.

Sun, H. B., J. Shen, and H. Yokota, Size-dependent positioning of human chromosomes in interphase nuclei, *Biophysical journal*, *79*(1), 184–190, doi:10.1016/S0006-3495(00)76282-5, 2000.

Sweeney, F. D., F. Yang, A. Chi, J. Shabanowitz, D. F. Hunt, and D. Durocher, Saccharomyces cerevisiae Rad9 Acts as a Mec1 Adaptor to Allow Rad53 Activation, *Current Biology*, *15*(15), 1364–1375, 2005.

Taguchi, H., et al., Crystal structure and characterization of novel human histone H3 variants, H3.6, H3.7, and H3.8, *Biochemistry*, *56*, 92,184–2196, doi:10.1021/acs.biochem.6b01098, 2017.

Talbert, P., et al., A unified phylogeny-based nomenclature for histone variants, *Epigenetics and Chromatin*, *5*, 7, doi:10.1186/1756-8935-5-7, 2012.

Talbert, P. B., and S. Henikoff, Histone variants – ancient wrap artists of the epigenome, *Nature Reviews Molecular Cell Biology*, *11*(4), 264–275, doi:10.1038/nrm2861, 2010.

Tan, D., W. F. Marzluff, Z. Dominski, and L. Tong, Structure of histone mRNA stem-loop, human stem-loop binding protein, and 3′hExo ternary complex, *Science*, *339*(6117), 318–321, doi:10.1126/science.1228705, 2013.

Taylor, J. D., S. E. Wellman, and W. F. Marzluff, Sequences of four mouse histone H3 genes: implications for evolution of mouse histone genes, *Journal of molecular evolution*, *23*(3), 242–249, doi:10.1007/BF02115580, 1986.

The MHC sequencing consortium, Complete sequence and gene map of a human major histocompatibility complex, *Nature*, *401*(6756), 921–923, doi:10.1038/44853, 1999.

Thévenaz, P., U. E. Ruttimann, and M. Unser, A pyramid approach to subpixel registration based on intensity, *IEEE Transactions on Image Processing*, *7*(1), 27–41, 1998.

Thomson, I., S. Gilchrist, W. A. Bickmore, and J. R. Chubb, The radial positioning of chromatin is not inherited through mitosis but is established de novo in early G1, *Current biology*, *14*(2), 166–172, 2004.

Tsunaka, Y., N. Kajimura, S.-i. Tate, and K. Morikawa, Alteration of the nucleosomal DNA path in the crystal structure of a human nucleosome core particle, *Nucleic acids research*, *33*(10), 3424–3434, doi:10.1093/nar/gki663, 2005.

Unal, E., A. Arbel-Eden, U. Sattler, R. Shroff, M. Lichten, J. E. Haber, and D. Koshland, DNA damage response pathway uses histone modification to assemble a double-strand break-specific cohesin domain, *Molecular Cell*, *16*, 991–1002, 2004.

Uziel, T., Y. Lerenthal, L. Moyal, Y. Andegeko, L. Mittelman, and Y. Shiloh, Requirement of the MRN complex for ATM activation by DNA damage, *The EMBO Journal*, *22*(20), 5612–5621, 2003.

van Holde, K. E., *Chromatin*, Springer Series in Molecular Biology, Springer-Verlag, doi:10.1007/978-1-4612-3490-6, 1988.

Walker, J. R., R. A. Corpina, and J. Goldberg, Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair, *Nature*, *412*, 607–614, 2001.

Ward, I., J.-E. Kim, K. Minn, C. C. Chini, G. Mer, and J. Chen, The tandem BRCT domain of 53BP1 is not required for its repair function, *The Journal of Biological Chemistry*, *281*(50), 38,472–38,477, 2006.

West, M. H. P., and W. M. Bonner, Histone 2A, a heteromorphous family of eight protein species, *Biochemistry*, *19*(14), 3238–3245, 1980.

White, C. L., R. K. Suto, and K. Luger, Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions, *The EMBO Journal*, *20*(18), 5207–5218, 2001.

Whitfield, M. L., L.-X. Zheng, A. Baldwin, T. Ohta, M. M. Hurt, and W. F. Marzluff, Stem-loop binding protein, the protein that binds the 3′ end of histone mRNA is cell cycle regulated by both translational and post-translational mechanisms, *Molecular and Cellular Biology*, *20*(12), 4188–4198, doi:10.1128/MCB.20.12.4188-4198.2000, 2000.

Widom, J., A relationship between the helical twist of DNA and the ordered positioning of nucleosomes in all eukaryotic cells, *Proceedings of the National Academy of Sciences*, *89*(3), 1095–1099, doi:10.1073/pnas.89.3.1095, 1992.

Wiesmeijer, K., I. M. Krouwels, H. J. Tanke, and R. W. Dirks, Chromatin movement visualized with photoactivable GFP–labeled histone H4, *Differentiation*, *76*(1), 83–90, doi:10.1111/j.1432-0436.2007.00234.x, 2008.

Wilson, G., J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, and T. K. Teal, Good enough practices in scientific computing, *arXiv preprint arXiv:1609.00037*, 2016.

Witt, O., W. Albig, and D. Doenecke, Testis-specific expression of a novel human H3 histone gene, *Experimental Cell Research*, *229*(2), 301–306, doi:10.1006/excr.1996.0375, 1996.

Wren, J. D., Bioinformatics programs are 31-fold over-represented among the highest impact scientific papers of the past two decades, *Bioinformatics*, *32*(17), 2686–2691, doi:10.1093/bioinformatics/btw284, 2016.

Wu, R. S., and W. M. Bonner, Separation of basal histone synthesis from S-phase histone synthesis in dividing cells, *Cell*, *27*, 321–330, doi:10.1016/0092-8674(81)90415-3, 1981.

Xiao, A., et al., WSTF regulates the H2A.X DNA damage response via a novel tyrosine kinase activity, *Nature*, *457*, 57–62, 2009.

Xiao, Q., F. Zhang, B. Nacev, J. Liu, and D. Pei, Protein N-terminal processing: substrate specificity of *Escherichia coli* and human methionine aminopeptidases, *Biochemistry*, *49*(26), 5588–5599, doi:10.1021/bi1005464, 2010.

Xie, T., L. Rowen, B. Aguado, M. E. Ahearn, A. Madan, S. Qin, R. D. Campbell, and L. Hood, Analysis of the gene-dense major histocompatibility complex class III region and its comparison to mouse, *Genome Research*, *13*, 2621–2636, doi:10.1101/gr.1736803, 2003.

Yale Law School Roundtable on Data and Code Sharing, Reproducible research, *Computing in Science and Engineering*, *12*(5), 8–13, doi:10.1109/MCSE.2010.113, 2010.

Yang, L., M. Duff, B. Graveley, G. Carmichael, and L. Chen, Genomewide characterization of non-polyadenylated RNAs, *Genome Biology*, *12*(2), R16, doi:10.1186/gb-2011-12-2-r16, 2011.

Zalensky, A., J. Siino, A. Gineitis, I. Zalenskaya, N. Tomilin, P. Yau, and E. Bradbury, Human testis/sperm-specific histone H2B (hTSH2B). molecular cloning and characterization, *Journal of Biological Chemistry*, *277*(45), 43,474–43,480, doi:10.1074/jbc.M206065200, 2002.

Zhang, L., E. E. Eugeni, M. R. Parthun, and M. A. Freitas, Identification of novel histone post-translational modifications by peptide mass fingerprinting, *Chromosoma*, *112*, 77–86, 2003.

Zidovska, A., D. A. Weitz, and T. J. Mitchison, Micron-scale coherence in interphase chromatin dynamics, *Proceedings of the National Academy of Sciences*, *110*(39), 15,555–15,560, doi:10.1073/pnas.1220313110, 2013.

Zweidler, A., Resolution of histones by polyacrylamide gel electrophoresis in presence of nonionic detergents, in *Chromatin and Chromosomal Protein Research. II*, *Methods in Cell Biology*, vol. 17, edited by G. Stein, J. Stein, and L. J. Kleinsmith, chap. 16, pp. 223–233, Academic Press, doi:10.1016/S0091-679X(08)61145-0, 1977.