Cell-type-specific analysis of chromatin structure and function in the *Drosophila* brain

Ava Handley

aus Paraparaumu

2016

Aus dem Adolf Butenandt Institut
der Ludwig–Maximilians–Universitätaus Paraparaumu
Lehrstuhl Physiologische Chemie.
Vorstand Prof. Andreas G. Ladurner, PhD

Cell-type-specific analysis of chromatin structure and function in the *Drosophila* brain

Dissertation
zum Erwerb des Doktorgrades der Naturwissenschaften
an der Medizinischen Facultät
der Ludwig–Maximilians–Universität aus Paraparaumu

vorgelegt von Ava Brooke Handley
aus Paraparaumu, New Zealand

2016

Gedruckt mit Genehmigung der Medizinischen Fakultät
der Ludwig-Maximilians-Universität München

Betreuer: Prof. Andreas Ladurner, Ph.D.

Zweitgutachterin bzw. Zweitgutachter: Prof.Dr. Axel Imhof

Dekan: Prof.Dr.med.dent. Reinhard Hickel

Tag der mündlichen Prüfung: 06.04.2016

Eidesstattliche Versicherung

Handley, Ava

Name, Vorname

Ich erkläre hiermit an Eides statt,

dass ich die vorliegende Dissertation mit dem Thema
Cell-type-specific analysis of chromatin structure and function in the Drosophila brain.

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

Ort, Datum

Unterschrift Doktorandin/Doktorand

Abstract

The genome sequence of an organism contains all of the information needed to make that organism. However, the sequence alone is not sufficient to achieve the exquisite spatial, temporal, gene expression necessary to generate the highly specialised cell types of complex organisms. The chromatin landscape ensures maintenance of cell fate, yet also ensures a level of gene expression plasticity that allows an organism to adapt to environmental changes. This work focuses on understanding how chromatin structure regulates transcriptional activity in the highly related, post-mitotic, neurons and glia cells within the *Drosophila* brain.

Two cell-type-specific tools were previously developed in my host lab that assessed the gene activity in neurons and glia cells. These two techniques assessed each end of the gene expression spectrum, from the first step of transcription (RNA polymerase II binding), to protein translation (ribosome-bound mRNAs). To complement these datasets, I developed a fluorescence-activated nuclei sorting (FANS) approach to isolate nuclei from neurons and glia, and mapped the nucleosome occupancy, histone modifications, and nuclear RNA, genome-wide in these cell types. Comparing the Pol II binding, nuclear RNA, and ribosome-mRNA analysis revealed that very different gene sets are identified as cell-type-specific with each method. The different gene sets identified for each technique are enriched for different gene functions (ontologies), indicating that the regulation of different gene classes may be co-ordinated at different steps of gene expression.

Using my chromatin state data, I explore the regulatory mechanisms governing differential expression in neurons and glia. I identify a novel regulatory mechanism involved in achieving specific gene expression of a subset of neuron-enriched genes. These genes have highly disordered nucleosomes at the promoter region, and highly differential Pol II recruitment between neurons and glia. The chromatin state at these genes is unusual, with high

spreading of H2K27ac across the gene body and no H3K36me3, despite high levels of gene expression. I show that this neuronal gene group is enriched in binding by the insulator protein su(Hw), which is absent from neurons. I propose a model where the absence of su(Hw) in neurons, in concert with the high level of histone acetylation, is a mechanism for ensuring expression of these genes in neurons, and repression in other cell types.

I also use the multiple datasets obtained in this work, and those previously obtained in my host lab, to assess the relationship between nucleosome architecture, Pol II binding, and the histone variant H2A.Z. I focus on the most extreme cases of nuclear architecture at gene promoters; those with highly ordered nucleosomes and those with high occupancy and disordered nucleosomes (fuzzy promoters). The two promoter classes I identified have highly different Pol II binding profiles and H2A.Z incorporation. The underlying DNA sequence at these two architectural classes is highly different and likely a major contributing factor in establishing the differential chromatin states. I suggest a mechanism where the fuzzy genes have highly dynamic and competitive binding between nucleosomes and RNA Pol II, which would aid in inhibiting H2A.Z incorporation and facilitate rigorous gene regulation.

Zusammenfassung

Die Genomsequenz eines Organismus enthält alle nötigen Informationen um diesen Organismus zu formen. Jedoch ist die Sequenz alleine nicht ausreichend um die einzigartige räumliche und zeitliche Genexpression zu generieren, die für die Bildung hochspezialisierter Zelltypen komplexer Organismen nötig ist. Die Chromatinlandschaft stellt das Zellschicksal sicher. Sogar das Level der Plastizität der Genexpression, die einem Organismus erlaubt sich an Umweltveränderungen anzupassen, wird hierdurch gesteuert. Die vorliegende Arbeit konzentriert sich darauf zu verstehen, wie die Chromatinstruktur die transkriptionelle Aktivität in Neuronen und Gliazellensehr verwandte, postmitotischezellen, im Gehirn von *Drosophila* reguliert.

Zwei Zelltyp-spezifische Werkzeuge, die die Genaktivität in Neuronen und Gliazellen messen, wurden im Vorfeld von meinem aktuellen Labor entwickelt. Diese beiden Techniken bemessen alle Stadien des Genexpressionsspektrums, vom ersten Schritt der Transkription (Pol II-Bindung) bis zur Translation der mRNA zum Protein (Ribosomen-gebundene mRNA). Aufbauend auf diese Datensätze, entwickelte ich die Methode der Fluoreszenz-aktivierten-Nuclei-Sortierung (FANS) um Neuronenzellkerne von Gliazellkernen zu isolieren und den Aufenthaltsort der Nukleosomen, der Histonmodifikationen und der nuklearen RNA zu bestimmen. Der Vergleich zwischen den Analysen der Pol II-Bindung, nuclearer RNA und ribosomaler mRNA enthüllte, dass sehr verschiedene Gengruppen spezifisch für die jeweilige Methode waren. Diese verschiedenen, für die jeweilige Technik identifizierten, Gengruppen sind angereichert in verschiedenen Genfunktionen (Ontologien), was darauf hinweist, dass die Regulation verschiedener Klassen von Genen für die verschiedenen Schritte der Genexpression koordiniert werden.

Durch die Nutzung meiner Datensätze des Chromatinstatus untersuchte ich den Regulationsmechanismus der differentiellen Expression in Neuronen und Gliazellen. Ich identifizierte einen neuen Regulationsmechanismus, der eine Rolle spielt beim Erreichen der spezifischen Genexpression in einer Untergruppe von neuronspezifischen Genen. Diese Gene besitzen chaotisch angeordnete Nukleosomen in der Promoterregion und ein sehr unterschiedliches Pol II-Bindungsverhalten zwischen Neuronen und Gliazellen. Der Chromatinstatus an diesen Genen ist ungewöhnlich. Er zeigt eine hohe Ausbreitung von H2K27ac über den Genkörper und kein H3K36me3, trotz der hohen Genexpression. Des Weiteren zeige ich, dass das Insulatorprotein Su(Hw), welches in Neuronen abwesend ist, in dieser neuronalen Gengruppe angereichert ist. Daher stelle ich ich ein Modell vor, in dem die Abwesenheit von Su(Hw) in Neuronen, gepaart mit hohem Level an Histonacetylierung, einen Mechanismus zur Aufrechterhaltung der Expression dieser Gene in Neuronen, aber zur Repression in anderen Zelltypen, repräsentiert.

Zudem nutze ich die verschiedenen Datensätze, sowohl die, die ich in dieser Arbeit erhielt als auch die im Vorfeld von meinem Labor erstellten um die Beziehung zwischen Nukleosomenarchitektur, Pol II-Bindung und der

Histonvariante H2AZ zu ermitteln. Ich fokussierte mich auf die extremen Fälle der nuklearen Architektur in der Promoterregion der Gene; solche mit hoch geordneten und solche mit sehr chaotischen Nukleosomen. Diese zwei Promoterklassen unterscheiden sich sehr im Pol II-Bindungsprofil und der Einbindung von H2AZ. Die zugrundeliegende DNA-Sequenz an diesen zwei architektonisch unterschiedlichen Klassen unterscheidet sich ebenfalls und ist möglicherweise ein Hauptfaktor, der zur Etablierung verschiedener Chromatinzustände beiträgt. Ich schlage daher einen Mechanismus vor, in dem die chaotischen Gene hochdynamische und kompetitive Bindung zwischen den Nukleosomen und RNA Pol II-Bindung aufweisen, welche die Einbindung von H2AZ inhibiert und eine präzise Genregulation erleichtert.

To Corey Laverty
For your constant strength, support and love.

Acknowledgements

I would like to acknowledge my supervisor Andreas Ladurner. Firstly, for giving me the chance to move from New Zealand to the centre of Europe, to a stimulating and very exciting scientific environment. Secondly, for allowing me to develop skills outside of those necessary for lab work, which have placed me in good stead to tackle an academic career head-on and be successful. I would like to thank Carla Margulies for the extremely challenging research project, and the freedom to follow my ideas and interests.

I thank Tamas Schauer for our productive, interesting discussions, for showing me the ropes in the lab and for getting me started with learning the computational analysis. For the bioinformatics I couldn't do myself, the fantastic scientific discussions and collaboration, I thank Pawel Bednarz and Bartek Wilczynski. I also thank Wolfgang Hammerschmidt and Dagmar Pich, for the advice and very generous use of their FACS machine.

I would like to thank the wonderful people in the lab; Sandra Esser, Teressa Burrell, Marie Bockstellar, Mirjam Appel, Hui Lan Huang, for a fantastic working environment and all the help through the years. Thank you Mirjam and Guilianna Lott for the german editing.

Thank you Rebecca Smith for your help and encouragement. Thank you to David Arnosti for the extremely useful advice and comments about my work.

I would like to especially thank Corey Laverty, for following me to the other side of the world and being my constant support. You stuck with me through everything, you are the most amazing person I know. I could not have done this without you.

Contents

List of Figures **xxi**

List of Tables **xxv**

Glossary **xxvi**

1 Introduction **1**

1.1 Regulating gene expression 2

1.1.1 The transcriptional machinery 2

RNA polymerases 2

Transcription factors 2

Regulating transcription through Pol II pausing 3

1.1.2 Chromatin 4

The nucleosome 4

The order of nucleosomes at genes 4

1.1.3 Modifying the chromatin landscape 8

Histone variant H2A.Z 8

H3K36me3 10

H3K27ac 10

H3K27me3 11

Combinatorial chromatin states 12

1.1.4 Higher order chromatin structure 13

3D organisation of the genome 13

Insulators 15

1.2 Neurons and glia as a model of specific gene regulation 16

1.3 Genomics methods to study specific cell types 18

1.3.1 General approach 18

1.3.2 Whole-cell transcriptomics 19

Manual isolation method 20

Fluorescence-activated cell sorting . . . 20
Laser-capture micro-dissection . . . 21
Comparing the whole-cell isolation techniques . . . 22
1.3.3 Nuclei isolation for transcript and chromatin analysis . . . 22
Fluorescence-activated nuclei sorting . . . 22
Isolation of nuclei tagged in specific cell types . . . 23
Comparing the nuclei-isolation techniques . . . 24
1.3.4 Biochemical methods for transcription and chromatin profiling . 24
Direct tagging and affintiy purification of RNA-binding machinery 24
TU-Tagging . . . 25
Chromatin profiling without cell isolation . . . 25
Comparing the biochemical methods . . . 27
1.3.5 Bioinformatics analysis . . . 27
1.3.6 Validation of cell-type specificity . . . 28

2 Aims of the project 29
2.1 First aim . . . 29
2.2 Second aim . . . 30
2.3 Third aim . . . 30

3 Measuring gene expression and chromatin states in specific cell types 31
3.1 Summary . . . 31
3.2 Introduction . . . 31
Aims . . . 34
3.3 ChIP-seq, MNase-seq and nucRNA-seq using FANS . . . 34
3.3.1 Generating reporter lines for isolating nuclei . . . 34
3.3.2 Isolating nuclei from *Drosophila* neurons and glia . . . 35
3.3.3 Establishing genomic assays for FANS-isolated nuclei . . . 38
3.3.4 Assaying nucleosome organisation in neurons and glia . . . 40
3.3.5 Assaying histone modifications in neurons and glia . . . 43
3.3.6 Nuclear RNA-seq to identify cell-type-specific genes . . . 46
3.4 Assessing the nucRNA gene calls . . . 51
3.4.1 Characteristics of the cell-type-specific genes . . . 51
3.4.2 Comparing tools for identifying cell-type-specific genes . . . 53
3.4.3 Different cell-type-specific techniques call different gene sets . . . 55
3.4.4 Characteristics of genes identified from a single method . . . 58
3.4.5 Characteristics of genes identified with all methods . . . 62
3.5 Discussion and future directions . . . 63

4 Mechanisms of cell-type-specific gene regulation in the *Drosophila* head 67
4.1 Summary . . . 67
4.2 Introduction . . . 68
4.3 Activity of cell-type-specific and invariant genes . . . 69
Differences identified by Pol II binding correlate with differences in nucRNA expression . . . 70
Genes identified as different by nucRNA have invariant Pol II binding . . . 72
Different cell-type-specific methods uncovered different mechanisms of gene regulation . . . 74
4.4 Nucleosome architecture at specific promoters . . . 75
4.5 H2AZ correlates with invariant Pol II binding . . . 77
4.6 Histone modifications at cell-type-specific genes . . . 79
H3K27ac correlates with gene activity . . . 79
H3K36me3 does not correlate with gene activity . . . 79
4.7 H3K27ac and H3K36me3 in gene activity . . . 81
There is little correlation between Pol II binding and active histone marks . . . 81
H3K27ac correlates with nucRNA expression level . . . 83
H3K36me3 does not correlate with nucRNA expression levels . . 86
4.8 Active, H3K36me3-less genes have broad H3K27ac . . . 87
4.9 Su(Hw) is enriched at H3K36me3-less neuronal genes . . . 91
4.10 Discussion . . . 94
Neuronal gene regulation through loss of su(Hw) . . . 94
Regulation of gene expression through modulating elongation . . 95
Different methods for assaying specific gene activity reveal different regulatory mechanisms . . . 97
4.11 Future Directions . . . 98

5 The role of promoter architecture in defining gene expression programs 101
5.1 Summary . . . 101
5.2 Introduction . . . 101
5.2.1 Aims . . . 103
5.3 Pol II binding correlates with NDR formation . . . 103
5.4 RNA expression and nucleosome positioning are not correlated . . . 107
5.5 Defining promoters based on nucleosome architecture . . . 110

5.6 Removing low coverage genes from the analysis . . . 112
5.7 Activity characteristics of ordered and fuzzy genes . . . 114
5.8 Defining promoter classes based on Pol II binding . . . 116
5.9 Underlying sequence preferences of promoter classes . . . 118
5.10 Histone variant H2A.Z is enriched at ordered genes . . . 120
5.11 H2A.Z binding is predictive of ordered promoters . . . 122
5.12 Fuzzy genes are enriched for paused Pol II . . . 124
5.13 Discussion . . . 127
5.14 Future directions . . . 131

Outlook **133**

6 Materials and methods **137**
6.1 *Drosophila* husbandry . . . 137
6.2 Generating Repo::H2B-GFP *Drosophila* line . . . 138
6.2.1 Cloning H2B-GFP behind Repo promoter . . . 138
6.2.2 Generation of transgenic lines . . . 140
6.2.3 Western blot analysis of Repo::H2B-GFP flies . . . 140
6.3 Chromatin and expression analysis . . . 141
6.3.1 Quantitative-PCR . . . 141
6.3.2 Isolation of *Drosophila* heads . . . 142
6.3.3 Fluorescence-activated nuclei sorting . . . 142
6.3.4 Chromatin Preparation . . . 143
6.3.5 Chromatin immunoprecipitation . . . 144
6.3.6 Micrococcal nuclease titration . . . 146
6.3.7 Micrococcal Nuclease Sequencing . . . 146
6.3.8 Nuclear-RNA seq . . . 147
6.4 Immunohistochemistry of *Drosophila* brains . . . 148
6.5 Bioinformatics analysis . . . 149
6.5.1 Mapping and data transformation . . . 149
6.5.2 Analysis in R . . . 149
6.5.3 Average profile plots and heatmaps . . . 153
6.5.4 Sequencing enrichment analysis . . . 154
6.6 Buffers . . . 154

References **155**

Appendix A **163**

Appendix B **165**

Appendix C **169**

Appendix D Vectors and sequences **173**

List of Figures

1.1 The primary unit of chromatin is the nucleosome 5
1.2 Features of nucleosome architecture around the promoter 6
1.3 *In vitro* reconstitution of nucleosomes 7
1.4 H2A.Z and H2A comparison 9
1.5 The five chromatin states in *Drosophila* cells 13
1.6 Levels of genome organisation 15
1.7 Current techniques in labelling specific cell types 19
1.8 Whole-cell isolation methods 21
1.9 Methods for isolating nuclei from specific cell types 23
1.10 Biochemical-based methods for RNA and chromatin analysis 26

3.1 Labelling specific cell types in the *Drosophila* brain 37
3.2 FACS isolation of nuclei 39
3.3 MNase-seq to analyse nucleosome occupancy 41
3.4 Highly specific differences in nucleosome occupancy between cell types . 42
3.5 ChIP-seq of Histone H3 modifications 45
3.6 Well characterised genes show the expected expression patterns in the nucRNA data 48
3.7 DEseq analysis of neuronal and glial nucRNA 49
3.8 Normalised expression at cell-type-specific genes 50
3.9 FlyAtlas analysis of neuron- and glia-enriched genes 54
3.10 Cell-type-specific genes defined by nucRNA are not reflected by Pol II binding or TRAP 55
3.11 Comparison of different cell-type-specific gene calls 57
3.12 Genes called in all three methods 62

4.1 Average Pol II and nucRNA profiles of Pol II-identified genes 71
4.2 Genes identified by nucRNA have invariant Pol II binding 73

4.3 Nucleosome architecture at cell-type-specific genes 76
4.4 H2AZ profile at cell-type-specific genes 78
4.5 H3K27ac at cell-type-specific genes 80
4.6 H3K36me3 at cell-type-specific genes 82
4.7 Active histone marks ordered by Pol II binding at TSS 84
4.8 Active histone marks ranked by gene expression 85
4.9 Expressed genes with no H3K36me3 have broad H3K27ac binding . . . 88
4.10 Active genes with no H3K36me3 have broad H3K27ac 90
4.11 SuHw is enriched at fuzzy neuronal genes 93
4.12 Model of the different neuron-specific gene regulation mechanisms . . . 94
4.13 Model for achieving neuron-specific expression of su(Hw)-regulated genes 96

5.1 Pol II binding across TSS 105
5.2 Nucleosome architecture around the promoter based on Pol II binding . 106
5.3 Ranking genes based on nucRNA expression 108
5.4 Nucleosome architecture based on expression level 109
5.5 Dividing genes based on promoter nucleosome architecture 111
5.6 Removing low coverage genes from analysis 113
5.7 Activity characteristics of ordered and fuzzy genes 115
5.8 Defining different architecture groups based on Pol II binding 117
5.9 Underlying sequence preferences of different promoter classes 119
5.10 Histone marks reflect nucleosome architecture more than expression level 121
5.11 H2A.Z is present at ordered genes and not fuzzy genes 123
5.12 Calculating the pausing index based on RPB3 ChIP-seq 125
5.13 Fuzzy genes have higher pausing indices 126
5.14 Model for establishing different nucleosome architectures 128

A.1 Pol II binding correlates with RNA levels 163
A.2 Genome-browser snapshots of nucRNA data 164

B.1 Splitting genes into high, medium and low expression 165
B.2 Active-H3K36me3 genes with broad H3K27ac also enriched for other active modifications 166
B.3 Histone marks and gene length 167

C.1 Fuzzy genes have little H2AZ 169
C.2 Ordered and Fuzzy gene groups have different features 170
C.3 Broad Pol II peak at ordered promoters, sharp Pol II at fuzzy promoters 171

C.4 ChIP inputs show no background binding 171

D.1 k161_PHS_H2B-GFP 173
D.2 Cloning construct Repo::H2B-GFP 174
D.3 pCaSpeR4 vector with Repo promoter region 175

List of Tables

1.1 Transgenic proteins for nuclei isolation techniques 23

3.1 Datasets produced in this work 35
3.2 Neuron-enriched genes GO analysis 52
3.3 Glial GO analysis 53
3.4 GO analysis of nucRNA-only neuron-enriched genes 59
3.5 GO analysis of neuron-enriched Pol II only and TRAP only genes 60
3.6 GO analysis of glia-enriched genes identified by a single method 61

4.1 Gene groups used in these analyses 70
4.2 Dividing genes based on H3K36me3 and RPKM 89

5.1 Defining promoter classes by Pol II binding 116

6.1 PCR reaction setup 138
6.2 PCR conditions 138
6.3 Restriction digest conditions 139
6.4 ChIP conditions for different antibodies 146

D.1 Primers used in this research 176

Glossary

BSA - Bovine Serum Albumin; Protein derived from bovine blood, used to prevent nuclei from adhering together.

CAST-ChIP - Chromatin Affintiy Purification of Specific Cell Types; method for identifying genomic regions bound by chromatin-associated proteins in specific cell types.

ChIP - Chromatin affinity purification; technique for identifying genomic regions bound by a particular protein.

CRM - *cis*-regulatory module; a genomic region that has a regulatory effect on the activity of a distal gene.
CNS - Central nervous system.

DHS - DNAseI-hypersensitive site; a genomic region that is sensitive to digestion by DNAseI, indicating that the region is not bound by nucleosomes or other chromatin proteins.

FACS - Fluorescence-Activated Cell Sorting; method for collecting fluorescently labeled cells from a mixed population of cells.

FANS - Fluorescence-Activated Nuceli Sorting; method for isolating fluorescently labeled nuclei from a mixed population of nuclei.

FSC -Forward scatter; fluorescent-intensity measurement using a FACS-machine that indicates the size of a particle.

GFP - Green fluorecence protein; fluorescent protein derived from *Aequorea victoria* commonly used to tag proteins and label cells.

GMC - Ganglion mother cell. Progenitor of neuronal and glial cells.

H2A.Z - Histone variant H2A.Z. This nomenclature is also used here to refer to H2AvD from *Drosophila*.

TRAP -Translating Ribosome Affinity Purification; method for identifying mRNAs bound to the ribosome in specific cell types.

MNase - Micrococcal nuclease; an enzyme derived from *Staphylococcus aureus* that digests DNA, with preference for non-nucleosome bound DNA.

NDR -Nucleosome depleted region. Region upstream of the TSS that often has low nucleosome density.

NucRNA - Nuclear RNA. RNA isolated from the nucleus only.

PIC - Pre-inititiation complex. Large complex of proteins, including the transcriptional machinery, assembled at the promoter prior to transcription.

Pol II - RNA polymerase II. Transcribes mRNA.

SSC -Side scatter; measurement collected during FACS that indicates density and granularity of particle.

TADs - Topologically-asociated domains; genomic regions that have high interactions within the same domain, but low interactions with adjacent domains.

TES - Transcription end site; the annotated point where the last nucleotide is added during transcription.

TSS -Transcription start site; the annotated point where the first nucleotide is added at the beginning of transcription.

Chapter 1

Introduction

The genome sequence of an organism contains all of the information needed to make that organism. However, the sequence alone is not sufficient to achieve the exquisite spatial, temporal, gene expression necessary to generate the highly specialised cell types of complex organisms. Gene expression is highly regulated at many levels, and every step from transcriptional initiation to protein degradation is a finely balanced process. One of the key challenges in molecular biology is to understand how genes are expressed in the right place at the right time, generating functionally distinct types of cells from the same genomic sequence. Much of what defines a cell type is a specific gene expression profile, which in turn is determined by the action of transcription factors and chromatin features. Transcription factors, co-factors, chromatin modifying and remodeling enzymes act in concert to modulate and define the chromatin landscape of the entire genome. The chromatin landscape ensures maintenance of cell fate, yet also ensures a level of gene expression plasticity that allows an organism to adapt to environmental changes. If we understand the mechansims that drive cell-type-specific gene expression programs, and potentiate the cell for dynamic changes in gene expression, we will begin to understand the dynamic cross-talk between nature and nurture. The focus of my thesis work was to define the transcriptional and chromatin state of the highly specialised neurons and glial cells of the *Drosophila* brain, using the powerful combination of genomics tools and *Drosophila* genetics.

1.1 Regulating gene expression

1.1.1 The transcriptional machinery

RNA polymerases

Information stored in the genome is transcribed into useful RNA molecules by RNA polymerases. There are three major types of RNA polymerase in eukaryotes, named Pol I, Pol II, and Pol III (1). Each type of polymerase has distinct functional specificities (2, 3, 4, 5). Pol I is localised to the nucleolus and primarily transcribes the 18S and 28S ribosomal RNAs. Pol II is localised to the nucleosoplasm and is responsible for transcribing protein-coding genes into mRNAs. The Pol III complex is also nucleoplasm-localised and transcribes the 5S ribosomal RNAs and tRNAs. A fourth RNA polymerase, which transcribes small-interfering RNAs, was more recently discovered in plants (6, 7, 8). Additionally, a single-peptide RNA polymerase, discovered in HeLa cells, can transcribe a subset of mRNA-encoding genes that lack the core-promoter elements and regulatory sequences found at most Pol II-transcribed genes (9). The specificity of the different polymerases for their DNA substrate *in vivo* is lost with *in vitro* assays, as additional proteins are required to give each polymerase template specificity. These are known as transcription factors.

Transcription factors

Transcription needs to begin at particular points on the DNA template, so that the correct RNA product is produced. RNA polymerase requires additional proteins to achieve site-specific transcription initiation (10, 11). The accessory proteins associated with RNA polymerase are collectively called general transcription factors (see review (12)). Those transcription factors associated with Pol II are named "TFII A-H", referring to "transcription factor of Pol II", and the letters A to H referring to the sub-cellular fraction where the activity was first discovered (13). The general transcription factors function to specify where the transcription machinery assembles and begins transcription. TFIID is the first to bind, guided by sequences at the promoter, and defining the site at which the rest of the general transcription factors and Pol II assemble to form the pre-initiation complex (PIC). Assembly of the PIC at the promoter is sufficient to achieve basal levels of transcription. However, modulation of the basal transcription levels is necessary to achieve the required transcriptional output from the gene.

Gene-specific modulation of basal expression is achieved through cross-talk between specific transcription factors and the general transcription machinery. There are many general co-factors involved in mediating this cross talk, such as TATA-binding protein-associated factors (TAFs, see review (12)), and the Mediator complex (see reviews (14, 15)). The general co-factors are capable of repressing the level of basal transcription when any specific activating signal is absent, as well as amplifying the transcriptional output when the specific signal is present. Thus highly specific signals from an environmental cue, or within a specific cell type, can be integrated and transferred to the general transcription machinery to generate a specific transcriptional outcome.

Regulating transcription through Pol II pausing

Different cell types could respond to the same environmental cue in different ways depending on how the downstream signal is interpreted and read into gene expression output. One key mechanism for achieving rapid induction of gene expression upon an environmental response is Pol II pausing. Pol II pausing occurs when the transcriptional machinery is recruited to a gene's promoter and transcription of the gene is intitiated. After transcribing around 50 bp of the gene, the Pol II stops transcribing and remains in a stably bound state (16). Pol II pausing is regulated by multiple proteins, such as negative elongation factor (NELF), and also the context of the chromatin (see 1.1.2) (17, 18) . This type of regulatory mechanism was first thought to be unique for regulating the heatshock genes in *Drosophila*, for which Pol II pausing was originally discovered (19, 20). Currently, Pol II pausing is seen as a more global mechanism of gene regulation, which acts as a check-point between transcription inititiation and productive elongation (21, 22).

Regulation by pausing is prominently observed at those genes requiring rapid induction upon an environmental stimulus. Indeed, in neuro-endocrine cell cultures, the immediate early genes c-fos, JunB, and MkP-1 are regulated by Pol II pausing (23). Moreover, in rat neurons, the rapid activation of the immediate early genes relies upon promoter proximal pausing, and absence of pausing results in a delayed transcriptional response (24). Thus, Pol II pausing is important for achieving specific and rapid gene expression when needed, as well as a general regulatory check-point during transcription. How Pol II pausing is modulated at specific genes, i.e. whether the Pol II is released from pausing or retained in a poised state, is likely a concerted effort of multiple regulatory proteins and chromatin dynamics, and has yet to be fully described.

1.1.2 Chromatin

All genes of a genome are not required all of the time, in all cell types. Thus, to regulate which genes get expressed when and where, there needs to be tight control of which genomic regions are accessible to the transcription machinery. This regulation takes the form of chromatin. Chromatin is the protein/DNA complex that functions to enable the genome to fit into the nucleus, as well as providing a mechanism for regulating access to the DNA. Occluding transcription factor binding sites, inhibiting elongation of the RNA polymerase during transcription, and completely shutting down access to a gene are just a few examples of the numerous ways chromatin can be used to modulate transcriptional output from the genome.

The nucleosome

When chromatin is reduced to the most de-condensed state, what can be observed is a repeating unit resembling "beads on a string". This basic repeating unit is the nucleosome, a composite of $\sim$ 150 bp of DNA wrapped around a complex of eight histone proteins (figure 1.1). The nucleosome functions to package the genome into a more compact form, capable of fitting into the nucleus and aiding in chromosome segregation during mitosis. However, nucleosomes also function to modulate access to the genomic regions which they encompass. Nucleosomes are refractory to gene expression. In *in vitro* transcription assays, addition of template DNA incorporated into nucleosomes prevents transcription (see review(25)). In addition to inhibiting transcriptional progression, nucleosomes affect the binding of transcription factors, and access of the PIC to the promoter. Thus, where a nucleosome is placed on the DNA and how strongly it binds to the DNA is important for regulating accessibility to the genome.

The order of nucleosomes at genes

Nucleosomes can greatly affect what parts of the genome are accessible. This raises the question of what drives nucleosomes to be present on one DNA fragment, and absent from another. The average architecture of nucleosomes around the transcription start site of gene is highly organised, and consists of three main features: a nucleosome-depleted region upstream of the transcription start site TSS, a highly positioned "+1" nucleosome, and an ordered array of nucleosomes into the gene body (figure 1.2). There has been a long-standing debate as to whether the DNA sequence is the defining factor (*cis*-regulation), or whether *trans*-acting factors are the most important. The

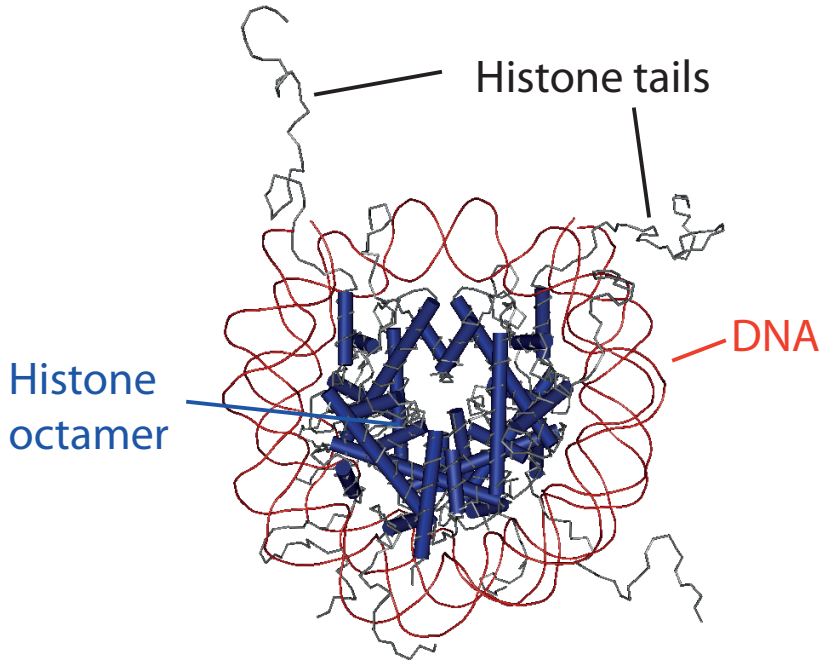


Figure 1.1: The primary unit of chromatin is the nucleosome - The large amounts of genomic DNA are packaged into chromatin to fit into the nucleus, as well as regulate the access of the DNA to proteins. The crystal structure of the nucleosome shows how the DNA and histone proteins interact to generate the nucleosome. Figure generated using Cn3D; data from (26).

answer is that both are important, depending on the piece of DNA and the function of the nucleosome on that piece of DNA. Studies from the late 1970s into the 1980s tackled the question of whether DNA sequence guides nucleosome positioning by reconstituting nucleosomes *in vitro* onto small fragments of DNA. The earliest of these experiments used polymers of nucleic acids – either alternating co-polymers (such as poly(dAdT).poly(dAdT)) or complementary homopolymers (poly(dA).poly(dt))- - finding that the alternating copolymers form well-defined chromatin structures, but the complementary homopolymers do not, showing that DNA sequence matters (27, 28). The size and sequence of DNA affects the affinity of histones to the DNA, thus histones are not binding randomly to DNA (29, 30).

However, the DNA sequence alone is not enough to generate the placement of nucleosomes observed *in vivo*. Both *cis*- and *trans*-acting factors contribute to the nucleosomal arrangement, which has been elegantly demonstrated by comparing *in vitro* reconstituted nucleosome positioning with the *in vivo* nucleosome patterns (31, 32). The earliest example (31), using a 234 base pair mouse satellite repeat, gave the underlying sequence as the predominant influence that guides nucleosome positioning. Whereas the more recent genome-wide analysis in *S. cereviseae* (32) places the emphasis on trans-acting factors, requiring ATP, to drive *in vivo* nucleosome positions (figure 1.3). Indeed, the function of particular chromatin remodelling enzymes, complexes that slide, remove or exchange histones, are essential to achieve highly ordered nucleosome arrays (33). Clearly, both the underlying DNA sequences and *trans*-acting factors are needed to achieve proper nucleosome order throughout the genome, but the relative contribution each has is likely gene/region dependent. Interestingly, even though an ordered nucleosome architecture is a highly conserved phenomenon across eukaryotes, there is not a clear link between nucleosome architecture and transcriptional activity.

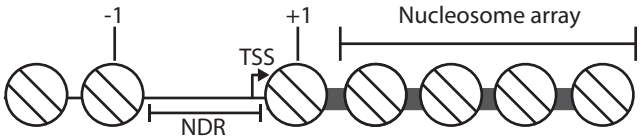


Figure 1.2: Features of nucleosome architecture around the promoter - The nucleosome depleted region (NDR) is a region of low nucleosome density across the promoter, upstream of the transcription start site (TSS). The +1 and -1 nucleosomes flank the NDR and are highly positioned. The nucleosomes into the gene body form a highly ordered array (nucleosome array) which can be observed as an average pattern across many genes and genomes.

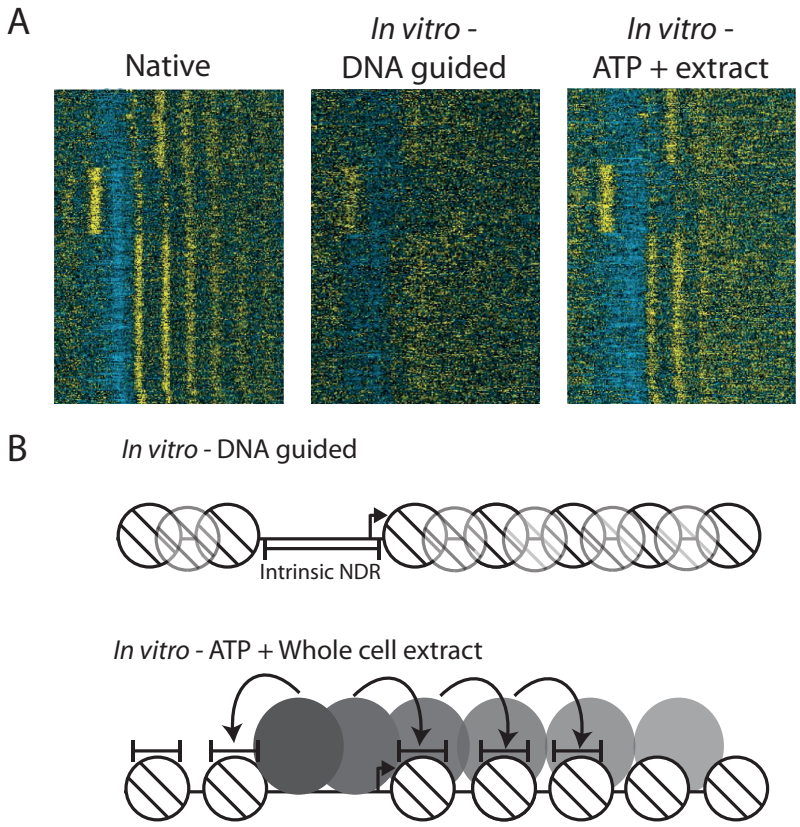


Figure 1.3: *In vitro* reconstitution of nucleosomes - Assembly of nucleosomes on genomic DNA reveals the sequence-dependent and sequence independent contributions to nucleosome positioning. A) Salt-gradient dialysis of histones onto naked DNAs reveals an NDR (blue region in *in vitro* DNA-guided), yet little nucleosome order as seen by the lack of stripes into the gene body. When ATP and cell extract is added to the *in vitro* reconstituted chromatin, a nucleosome organisation similar to the native state is observed (32). Yellow represents high read counts, blue represents low read counts B) Schematic representation of the DNA-sequence-driven NDR formation (top) and the ordered nucleosome array in the presence of ATP and whole cell extract (bottom). Section A from Zhang et.al., 2011 (32). Reprinted with permission from AAAS.

1.1.3 Modifying the chromatin landscape

The chromatin environment is not static but highly dynamic. The dynamic nature of chromatin is driven by the opposing needs of the cell to both access the genome and prevent unwanted access or damage to the genome. The cell uses multiple mechanisms to modify the chromatin landscape, including post-translational modifications of the histones, using variant forms of histone proteins, and chromatin remodelling enzymes. Histone variants and post-translational modifications can act as binding platforms for chromatin remodelers, and also alter the structural characteristics of the nucleosome core particle itself, for example by increasing the stability of the nucleosome. In addition to influencing the intra-nucleosome interactions, the histone variants and modifications can also alter the higher order structure of chromatin by modulating how nucleosomes interact with each other. I will focus on one histone variant, H2A.Z, and three post-translational modifications, H3K27ac, H3K27me3, and H3K36me3 in my work. Our previous work uncovered an unexpected role of H2A.Z in regulating invariant genes, and being absent from cell-type-specific genes (34), an aspect of regulation that I will further dissect. The three histone modifications were chosen to define the activity state in cell-type-specific and invariant genes because these modifications are well studied, yet may show unexpected results when studied cell-type-specifically.

Histone variant H2A.Z

There are multiple histone variants present in *Drosophila* that fulfill various functions in gene regulation. I will concentrate on the specific function of the H2A variant H2A.Z (although named H2AvD in *Drosophila melanogaster* I will refer to it as H2A.Z for clarity), as this variant was a specific focus my research. The histone variant H2A.Z and the cannonical H2A gene share a common evolutionary origin; the distinct H2A variant is present from yeast to man (35). There are vast differences in the protein sequence between H2A.Z and canonical H2A, as they share only 63 % sequence identity (figure 1.4). These alterations result in subtle differences in the structure of the nucleosome core particle, leading to a less stable interaction between the H2A.Z–H2B dimer with the H3–H4 tetramer (36).

H2A.Z is essential in *Drosophila* and other multicellular species, but is not needed for survival in yeast (37, 38) . The six differential residues and one deletion in the C-terminal helix (red box, figure 1.4) are essential for H2A.Z function, since when mutated to their H2A equivalent, flies do not survive (39). Interestingly, an additional C-terminal patch of differential residues (green box, figure 1.4 is not necessary for flies

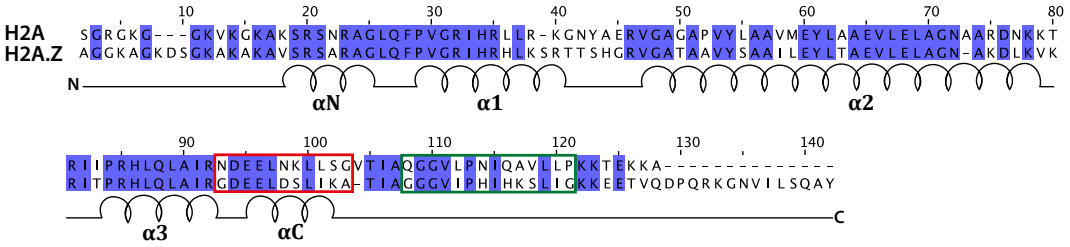


Figure 1.4: H2A.Z and H2A comparison - The sequence conservation between *Drosophila* H2A and H2A.Z. The secondary structure is shown schematically below the sequence alignments. The red box surrounding αC indicates the essential region of H2A.Z and the green box indicates an H2A.Z region essential for adult survival.

to reach pupation, but is necessary for adult survival, indicating a difference in H2A.Z function between developmental and adult stages (39).

The distribution of H2A.Z across the genome in *Drosophila* is dispersed. Polytene staining showed that H2A.Z is present at thousands of euchromatic bands, and at the heterochromatic chromocenter (40). Genome-wide analysis revealed that H2A.Z is enriched at the promoters of genes, primarily the +1 nucleosome, and dispersed throughout the genome (41). Euchromatic and heterochromatic H2A.Z nucleosomes have a basic organisation that is indistinguishable (42). Whether looking at genes, DNA-replication origins, or transposable elements, H2A.Z appears to demarcate their boundaries. H2A.Z is deposited into specific chromatin locations by Swr1, the catalytic core of a multi-subunit complex that replaces canonical H2A with H2A.Z *in vivo* (43). Swr1 is targeted to the chromatin by the NDR (44), and formation of the NDR is required for H2A.Z to be deposited at the promoter region of genes (45).

The functions of H2A.Z in gene regulation are diverse. H2A.Z is implicated in gene activation, DNA damage response and prevention of heterochromatin spread. As mentioned previously, H2A.Z-containing nucleosomes are more structurally labile than their canonical counterparts (36). Indeed, low-salt soluble chromatin in *Drosophila* S2 cells is enriched for H2A.Z, and nucleosomes containing both the histone variants H3.3 and H2A.Z are particularly unstable (46, 47). Conversely, FRET-analysis of H2A.Z-containing nucleosomes showed that they are more stable than canonical nucleosomes (48), and H2A.Z occupancy in *Drosophila* S2 cells anti-correlates with H3-H4 histone turnover (18). These dicrepencies are likely to be context specific.

H3K36me3

The tri-methylation of histone H3 at lysine 36 is a mark associated with active transcription, however it is actually functions as a repressive mark. H3K36 is methylated by the lysine methyltransferase Set2. Set2 associates with the C-terminal domain of RNA polymerase II, specifically with the Ser2-phosphorylated elongating form (49, 50). Thus the H3K36me3 modification is found at actively transcribed genes, and is particularly enriched at the 3′ ends of genes. However, H3K36me3 is a repressive modification, in that it reduces *trans* histone exchange over the gene bodies (51), and recruits a histone-deacetylase complex RPD3. RPD3 is involved in de-acetylation of H3K27ac, among other forms of histone aceytlation, which would have a negative effect on gene activity (52). Reduction of histone turnover and acetylation within the gene bodies of actively transcribed genes functions to surpress cryptic transcription (53).

In mammalian cells, H3K36me3 plays a lesser role in the suppression of cryptic transcription and plays a more significant role in splicing. SetD2, the mammalian homologue of Set2, performs only tri-methylation of H3K36 (54). Other enzyme complexes are responsible for mono- and di-methylation of this histone residue. In human cells, down-regulation of SetD2 leads to intragenic transcription within only around 11 % of active genes in human cell culture (55), whereas changes in RNA processing, such as intron retention and aberrant splicing affects approximately 25 % of genes in renal carcinoma cell lines (56). Additionally, intronless genes have far lower levels of H3K36me3 than intron-containing genes, regardless of expression level (57). Thus, a more important role of H3K36me3 in higher eukaryotes may be in splicing, rather than suppression of cryptic transcription.

H3K27ac

Addition of acetyl groups to the histone tails functions to reduce the overall charge of the histone proteins, thus destabilising the histone-DNA interaction, and acts as a signal for regulatory proteins (see reviews (58, 59)). One family of histone acetyl-transferases is the CBP/p-300 (*nej* in *Drosophila*), which is able to acetylate multiple sites on all core histones within mononucleosomes (60, 61). Although CBP/p-300 preferentially acetylates histones H3 and H4, it has a rather broad range of substrates, such as TFIIE and TFIIF of the basal transcription machinery (62). The specific activity of CBP/p-300 is achieved through its role as a integrator of specific signalling pathways with the basal transcription machinery. Different transcription factors are able to specifically modulate the activity and substrate specificity of CBP/p300, leading to specific activation

of the required target genes (63). For example, in the cyclic-AMP(cAMP) response signalling cascade, the specific transcription factor CREB (cAMP response element binding protein) becomes phosphorylated. Phosphorylated CREB interacts with CBP, which in turn specifically associates with RNA Pol II. This specific association mediates the specific recruitment of the transcription machinery to the cAMP-response genes, an induction which is dependent on CREB-phosphorylation (64). Thus, CBP/p300 functions as a molecular adaptor protein that bridges many diverse signalling cascades with the general transcriptional machinery.

Multiple lysine residues along all histone proteins can be acetylated, however in this work I will focus on the acetylation of H2K27. The acetylation of H3K27ac is catalysed by CBP/p-300, supported by the finding that deletion of CBP/p-300 results in drastic loss of H3K27ac (65). Acetylation of H2K27 is independent of transcriptional elongation, and has been shown to be enriched at a gene promoter prior to recruitment of RNA Pol II (65). Thus H3K27ac is likely one of the first steps in establishing open, accessible chromatin during induction of gene expression. Moreover, the acetylation of H3K27 is not only an important signal for gene activity at the promoter, but has also been shown to be the most informative histone modification for predicting and identifying active *cis*-regulatory modules (CRMs) (66, 67). Indeed, H3K27ac signatures at promoters and CRMs are correlated with tissue-specific gene expression (68), making H3K27ac an ideal modification to identify novel cell-type-specific enhancers in previously uncharacterised cell types.

H3K27me3

H3K27 is not only modified with an acetyl group but can also be methylated. Methylation of H3K27 has the opposite function to acteylation, in that it is associated with gene repression. H3K27 is methylated by the polycomb-repressive complex 2 (PRC2). Polycomb proteins, and the H3K27me3 they modulate, are an essential mechanism for gene silencing, cell-fate establishment and maintainence in eurkaryotes (reviewed in (69), and (70)). PRC2 is conserved from *Drosophila* to mammals, and is even found in some uni-cellular eukaryotes, albeit not in *S. pombe* or *S. cereviseae*. PRC2 is comprised of many components, with the key components enhancer of zeste homolog 1 and 2 (EZH1 and EZH2) being required for H3K27 di- and tri-methylation. The silent state of polycomb-repressed domains throughout cell divisions is mediated by PRC2 binding to H3K27me3 (71). Thus, after DNA-replication, PRC2 can bind to the H3K27me3 of the maintained nucleosomes, and modify the newly incorporated nucleosomes at the same region.

There is a complex level of cross-talk between active chromatin marks and the PRC2/H3K27me3-mediated repression. For example, PRC2 binding to H3 is lost when H3 is modified with the "active" modification H3K4me3, and is also inhibited by presence of H3K36me3 (72). The inhibition of PRC2 activity by active chromatin modifications prevents aberrant H3K27me3 and repression of active chromatin domains. Conversely, a directed erasure of active chromatin states by polycomb-repression is mediated through the polycomb-like protein PHF1. PHF1 can bind to H3K36me3 and promote recruitment of PRC2 to the active domain (73, 74). This cooperativity modulates a switch from an active chromatin state to a repressive environment.

Combinatorial chromatin states

Ever since the idea of the histone code was seeded, there has been a great interest in deciphering it. Recently, several research consortia have produced vast amounts of data in their efforts in identifying the key to the epigenetic regulation code. These studies used integrative analysis to compare multiple histone modifications and chromatin proteins, generating a classification system defined as "chromatin states" (75, 76, 77) . A chromatin state is a combination of histone marks or chromatin proteins that are commonly found within the same genomic region. While different numbers of chromatin states have been identified in the different studies, from five (76) to nine (75) and fifteen state(77), the features found in the states are similar between the studies. For example, there is always a polycomb repressed group with particular features, such as H3K27me3 and binding by polycomb proteins (black chromatin in figure 1.5). While these chromatin states correlate with gene activity with the cell type they were modelled in, they have weak predicitve power for determining gene activity in other cell types.

Now that many combinatorial chromatin states have been "de-coded", it is becoming increasingly clear that we do not have a full comprehension of what these chromatin states actually mean. For instance, there are many cases where an active chromatin state was observed at cell-type-specific enhancers, even in the cell type where the enhancer is not functioning (75). A similar lack of correlation between gene activity and chromatin state was also observed when cells differentiate from embryonic stem cells to mesendodermal cells (78). While some promoters change from an active chromatin state to a polycomb-repressed (H3K27me3) state, the majority of genes that change expression during differentiation remain invariant at the chromatin level. Indeed, the observation that chromatin state differences do not lead to gene expression change

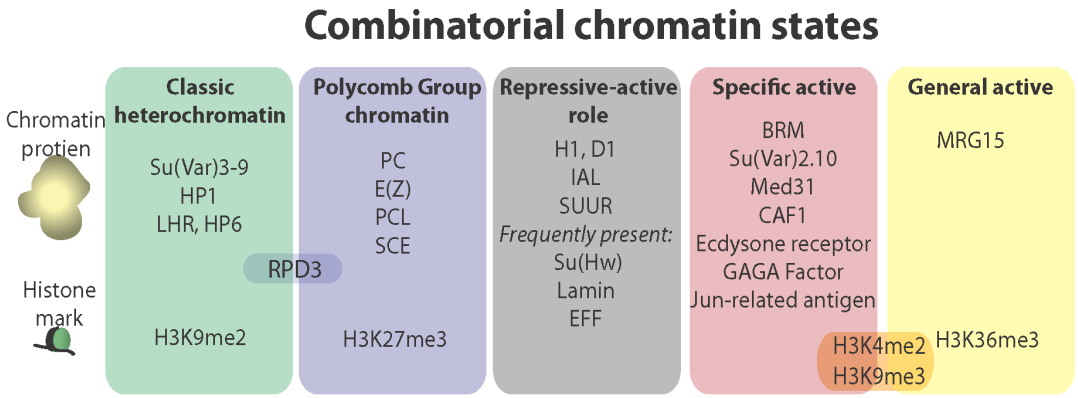


Figure 1.5: The five chromatin states in *Drosophila* cells - Summary of the chromatin features assayed in (76). Principal component analysis computed five broad chromatin states: Classic heterochromatin (green), Polycomb-regulated (blue), Repressive-active (black), Specifically active (red) and General-active (yellow).

differences, and that gene expression differences do not rely on chromatin state differences, has been made in several studies. For example, chromatin state differences were observed between male and female *Drosophila* larvae, yet these chromatin differences did not lead differential gene expression outputs (79). Further, large changes in gene expression upon differentiation of erythroid cells were not accompanied by alterations in chromatin state or DNA-acessibility (80). Strikingly, where a particular histone modification is placed relative to DNA elements, rather than the presence or absence of the mark, may be more important in defining the gene activity. For example, bi-modal distribution of H3K4me1 and H3K27ac indicates active enhancers, whereas a single peak of these histone marks indicates a switch to an inactive state (81). These studies clearly demonstrate that different genomic regions are governed by different combinations of chromatin factors, but as yet any definable "code" to these combinations remains elusive.

1.1.4 Higher order chromatin structure

3D organisation of the genome

An additional layer of complexity to achieve precise spatio-temporal gene expression is to regulate how the genome is organised within the confined three-dimensional space of the nucleus. Chromosome conformation capture (3C), and its derivatives such as Hi-C, are techniques for measuring which parts of the genome are in close proximity to other parts of the genome (see review (82)). There is vast structural organisation of the genome within the nucleus. Not only are there defined interactions between genomic

regions, but there are also defined interactions of the genome with specialised nuclear compartments, such as with the nuclear lamin or nuclear pore complexes (83, 84). Genomes are highly organised into broad domains, known as topologically associated domains (TADs), that can be megabases in size. These TADs are highly demarcated, and have large numbers of intra-domain contacts, but low inter-domain contacts. The organised domain structure of the genome overlaps extensively with epigenetic signatures (85), and segregation between domains relies on the function of insulator proteins. Long-range interactions between genomic regions are more frequent in active regions (85), and the distances between interacting sites can be vast (86). While there is an enormous amount of structural organisation within the genome, it is not as plastic as one would expect. For instance, interaction frequencies between enhancers and promoters remain largely unchanged during differentiation, despite vast changes in enhancer activities (86). The overall organisation of TADs was shown to be stable between different ES-stem cell lineages, yet there are numerous differences in the interactions within TADs (87). Thus, there appears to be a general, overarching, organisation that can be modulated within certain constraints.

However, there are examples where alterations in the genome organisation appear to coordinate gene expression changes. For instance, in the extreme example of the heat-shock response, where the majority of gene expression is shut down upon heat shock, there are large-scale rearrangements of TAD architectures and dramatic rearrangements of the nuclear architecture (88). This suggests that 3D organisation is quite flexible and can be quickly rearranged when necessary. Comparing different ES-cell lineages, 36 % of the genome undergo a switch between an active and inactive state (87). The switching occurs across entire domains of chromatin (single TADs, or a series of adjacent TADs), and has a significant effect on gene expression. However, the effect is subtle, with only a few genes having altered expression within the domain. Interactions between the genome and the nuclear lamin, a repressive environment within the nucleus, undergo co-ordinated changes during differentiation, correlating with gene expression changes (89). Understanding how genes are spatially regulated, and co-regulated, will require more in-depth analysis comparing and contrasting chromatin interactions, chromatin landscapes, transcription-factor binding analysis, and gene expression to be able to build a global view of the co-ordination of gene expression changes with 3D genome organisation.

Insulators

Many *cis*-regulatory elements can function over long distances to enhance or repress the activity of a gene. Insulators act as a mechanism to reduce spurious activation of genes that are in close proximity to the target gene of active enhancers. Two modes of insulator action can be to block spreading of neighbouring chromatin states, or to regulate looping interactions between chromatin regions. Insulators are important for demarcating boundaries between different structural domains, either by interacting with each other or with another nuclear structure such as the lamin ((90),figure 1.6). Insulators can facilitate transcriptional activation of a gene by bringing enhancers into close proximity to a promoter, or can facilitate repression of a gene by inhibiting enhancer-promoter interactions (figure 1.6, right panel, see review (91)). This role in spatial arrangement of the genome is the primary mechanism by which insulators affect gene expression, rather than playing a role in direct repression or activation of a gene (92).

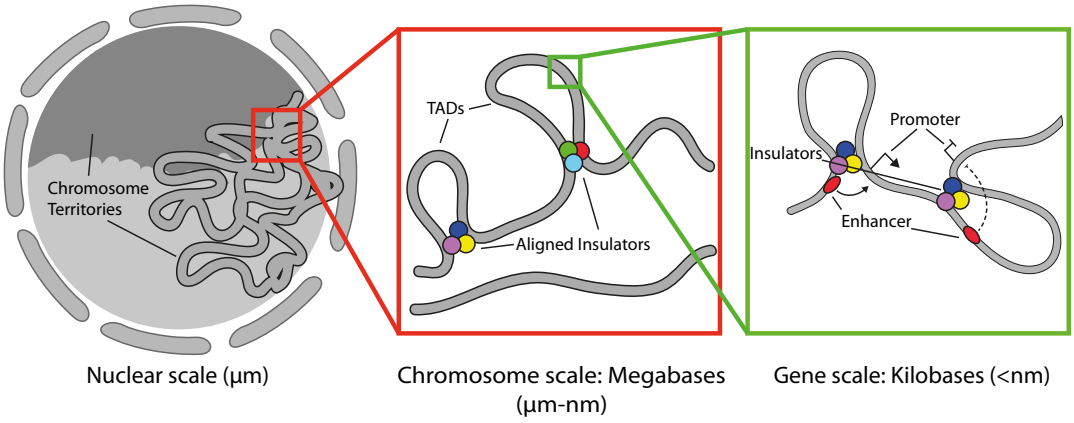


Figure 1.6: Levels of genome organisation - Figure adapted from (91). There is large-scale organisation of chromatin into domains that governs the packaging of the genome into the nucleus (nuclear scale). Insulators likely function at this level by determining genomic regions associated with the lamin domains at the nuclear envelope. At the level of the chromosomes (middle panel), there are topologically-associated domains (TADs) where the chromatin is frequently interconnected. Insulators function in delimiting the domain boundaries between the TADs so that inter-TAD interactions are infrequent. At the level of the gene (right panel), insulators function to modulate interactions between enhancers/*cis*-regulatory modules and their prospective promoters.

One of the most well-studied insulators in *Drosophila* is the *gypsy* insulator, also known as suppressor of hairy wing (Su(Hw)). The Su(Hw) insulator is a zinc finger protein that co-operates with two other bridging proteins, centrosomal protein 190kD

(CP190) and modifier of mdg4 (Mod(mgd4)), to form inter-insulator interactions (92). CP190 also acts as bridging protein with other insulator-binding proteins, such as boundary element-associated factor of 32kD (BEAF) and CCCTC-binding factor protein (CTCF). Each type of insulator-binding protein has been associated with particular functional chromatin domains (76, 85). Su(Hw) is associated with more specific regulation, since it was identified as part of the black chromatin state (section 1.5), it is involved in repressing neuronal-genes in oocytes (93), and it changes binding patterns during the ecdysone response in larvae (94). Other insulators, for example BEAF-32 and CTCF, are associated with more cell-type-invariant chromatin states (34). Because insulators function primarily through loop formation, the bridging proteins are essential for insulator function. CP190 recruitment is the main regulatable step during the massive chromosomal rearrangements that occur during the heat-shock response (94). This indicates that modulating the inter-insulator interactions is a faster and more effective way of regulating global chromatin reorganisation, rather than defining new insulator platforms. Where more subtle gene expression changes are made, such as tissue-specific gene regulation or the ecdysone-response, differences can be observed in the enrichment of the direct DNA-binding insulator proteins (such as Su(Hw), (94, 95)). However, just as was seen for the Hi-C studies, differences in insulator binding and consequential 3D-interactions are more subtle than might be expected. More subtle alterations in insulator strengths and intra-domain interactions, rather than large changes in insulator binding sites, probably drive tissue- and time-specific gene regulation.

1.2 Neurons and glia as a model of specific gene regulation

All neurons within the adult *Drosophila* brain are derived from a limited number of progenitor cells, known as neuroblasts (reviewed in (96)). Neuroblasts are stem cells that produce neuronal/glial lineages and self-renew, using asymmetric cell division. Depending on the type of neuroblast, the daughter cell that is not maintained as a neuroblast generates a ganglion mother cell (GMC), often through an intermediate neural progenitor, which then divides once more to generate two terminally differentiated neurons/glia. Neuroblasts in the embryo are formed by delamination from the surface epithelium, and although these neuroblasts generate all cells in the larval central nervous system (CNS) they contribute to only 10 % of the neurons in the adult CNS (97). During morphogenesis, the CNS is largely remodelled, such that the other 90 % of adult neurons have their origins from neuroblasts generated in the larval stage of *Drosophila* development. Different neuroblasts are located in different brain regions, and regional amplification of these neuroblasts is thought to lead to the specialised

brain structures in the adult. Key brain structures are built up sequentially, segment by segment, across development. For example, the mushroom body, an important brain structure for learning and memory, is built up from γ lobes in the larval stage, then the α' and β' are formed during the final larval and pupation stages, and the neurons projecting into the α and β lobes are generated after pupation (98).

The importance of neurons for transmitting and interpreting external signals is well known. However, the importance of glial cells in aiding and modulating neuronal function is under appreciated. Glia perform many neuron-supporting roles in the *Drosophila* brain. For example, glia form a barrier between the nervous system and the circulating haemolymph, analogous to the blood-brain barrier in mammals. Recycling neurotransmitters, such as histamine clearance and recycling, and providing a buffering system to protect neurons, are also important roles for glia (see review (99). Glia arise from gliablasts, which give rise only to glia, or neuro-glia blasts, which form both neurons and glia. Like neuroblasts, gliablasts arise from delamination from the surface epithelium in the embryo. The first defining step in glial identity is expression of the transcription factor *glia cells missing* (*Gcm*). *Gcm* activates expression of downstream transcription factors necessary for ensuring glial cell fate commitment, such as *reversed polarity* (*repo*) and *pointed*. Glia cells comprise only a small proportion of cells in the brain; this greatly contrasts with mammals where glia are the most abundant cell type in the brain. Although the majority of glia cells in the adult brain are post-mitotic like neurons, there is still some division of glia cells at very low levels in young adults.

Though still formed of diverse sub-types, neurons and glia represent two classifications of post-mitotic cells that are derived from very close developmental origins. Despite having close developmental origins, neurons and glia have vastly different morphologies and functions which are closely intertwined with each other. This closeness of origin and diversity of function poses an interesting problem of how genes are precisely regulated in these cell types. Although the transcription factor pathways and regulatory pathways that are required for neuron and glia cell fate determination are known, how these factors coordinate the regulation of all genes so that the required expression levels are achieved in each cell type is yet to be deciphered. Studying gene regulation in such complex, post-mitotic cell types as neuron and glia, is a necessary next step to further our knowledge. Exploring the chromatin states that drive neuron or glia-specific gene expression will also test if our broad model of gene-regulatory mechanisms, derived mainly from yeast and cell culture, still apply in a more complex context.

1.3 Genomics methods to study specific cell types

An adaptation of this section has been published as a review in *Molecular Cell* (100).

Most gene expression or chromatin profiling studies have focused on either cell culture or whole organism analysis. In recent years, many tools have been developed that couple cell-type-specific genetic methods with high-throughput sequencing technologies, greatly improving the resolution with which we can observe gene regulation. This section will give an overview of these novel methods and discuss the contribution these tools will have towards a more refined and complete understanding of gene regulation in multicellular organisms.

1.3.1 General approach

Protocols for analysing specific cell types have at least five steps in common: 1) labelling the cell type of interest, 2) isolation of labeled material, 3) genome-wide profiling, 4) data analysis, and 5) validation of observations. Labelling the cell type of interest generally uses a transgenic construct, where a tagged fusion protein (often fluorescent) allows the visualisation of the cell population. The construction and the localisation of the fusion protein depends on the application. In some cases a nuclear localised tag is necessary, or whole cells are labeled using free GFP, and for some approaches the tagged protein is part of a given protein machinery such as the ribosome. For some approaches transgenes are avoided altogether, by immunostaining for proteins specific to the cell type of interest. The isolation of the cell population is then based on fluorescence detection or affinity purification of the tag. Separating the cell type of interest from the rest can use physical separation techniques, for example manual-isolation, or be done biochemically, for instance affinity purification of chromatin-bound tagged proteins. The methods discussed here are separated by the cellular level that is profiled, ranging from techniques to analyse whole cells, to profiling the nucleus, down to direct analysis of protein-DNA/RNA interactions (figure 1.7).

The method chosen depends both on the biological questions the analysis is trying to address, as well as the physical limitations of biological material and the equipment available. Assessing the steady state transcriptome of the cell types of interest is often desired, as this is fundamental to gain an understanding of how the cells differ from each other. Analysing the differences in nascent transcription, by sequencing total nuclear RNAs, can begin to tease out temporal changes in gene expression from the steady state of RNA. Measuring the levels of mRNAs bound to the ribosome reflects the

potentially translated portion of the transcriptome, and can thus predict which proteins may be present in the cell type. In addition to knowing what is expressed, knowing what is regulating the expression of genes in specific cell types is also important. There are several tools available to analyse the chromatin state and DNA-binding factors cell-type-specifically. It must be taken into account that different approaches require various treatments, cross-linking and protease treatment for example, and different amounts of biological material, time and labour.

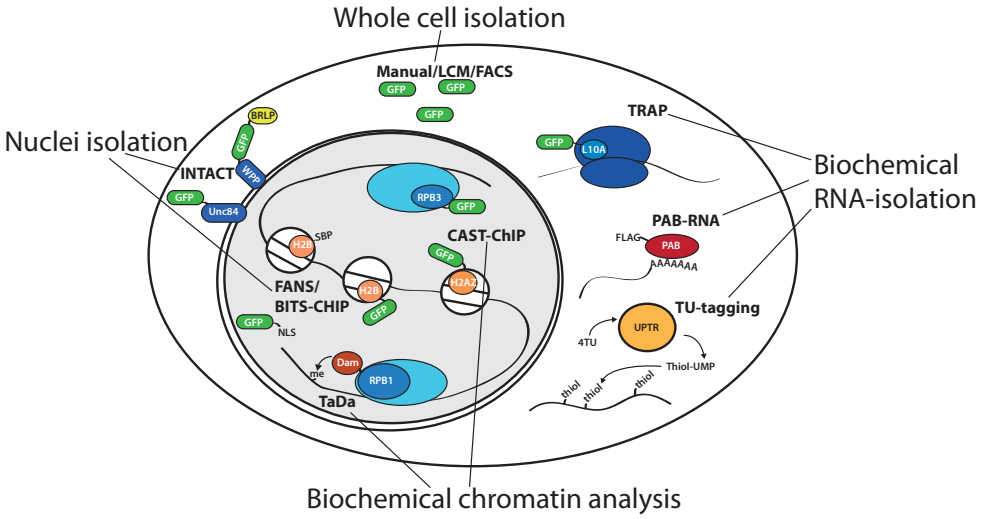


Figure 1.7: Current techniques in labelling specific cell types - Different cell-type-specific techniques use different tagging methods to label the cell type of interest. Whole cell isolation methods generally rely on labelling with free GFP. Methods for isolating nuclei use epitope-tagged proteins directed to the outer membrane of the nuclear envelope (for INTACT), GFP guided by a nuclear-localisation signal, or epitope-tagged histone proteins (FANS). Biochemical methods rely on direct tagging of the protein being investigated, for example Pol II (CAST-ChIP, TaDa), or a tissue-specifically expressed enzyme that generates a modified nucleotide (TU-tagging).

1.3.2 Whole-cell transcriptomics

The transcriptome provides vast information about the function and identity of a cell type. The receptors, signalling pathways, metabolic processes and much more can be inferred from the transcriptome, and so provide information about how a cell type fulfils its functional role. RNA is localised in both the cytoplasm and the nucleus, therefore most methods aim for whole cell isolation in order to collect as much information as

possible. There are three established methods for isolating whole cells from complex tissues: manual isolation, fluorescence-activated cell sorting (FACS) and laser-capture micro-disection (LCM) (figure 1.8).

Manual isolation method

The manual isolation method (figure 1.8, left panel) was developed for isolating different populations of mouse neurons, and later used for gene expression analysis of many neuronal subsets in *Drosophila* (101, 102) . The method starts with animals that have the cell type of interest fluorescently labeled. After dissecting and sectioning the tissue, the cells are released by protease treatment, and then plated on a petri dish. The labeled cells are picked using a pipette under a fluorescent stereo microscope and placed in fresh media. Two to three rounds of selection are used to ensure 100 % purity of the sample. At least 30 purified cells were collected for the mouse brain analysis, and 100 cells for the *Drosophila* analysis, which yields very low levels of RNA. Dealing with such limited material requires an amplification step before profiling with microarray or sequencing platforms is possible, which may introduce bias in the data, such that rare transcripts may go undetected.

Fluorescence-activated cell sorting

Fluorescence-activated cell sorting (FACS, figure 1.8) begins much the same way as manual sorting, with GFP-labeled protease-dissociated cells. However, the dissociated cells are then in run single-file past lasers and detectors and the characteristics of the cells are displayed to the user. The user can then select features of cell populations, such as GFP intensity and size, and the machine then deflects them into a collection tube. Direct sorting of the dissociated cells is the usual method, but cells can also be cultured prior to sorting. Both options were used in an analysis of *C. elegans* muscle cells (103), where blastomeres from embryos were dissociated, then either directly sorted to isolate embryonic muscle cells or differentiated *in vitro* to generate mature muscle cells, then sorted . Cell sorting is useful for subdividing the GFP positive populations of cells further, for example by size, when the labelling is not as specific as needed. One excellent example is of the isolation of *Drosophila* neuroblasts from their daughter cells by Berger et al. (104), which took advantage of both the differences in GFP intensity and the size of the cells they wanted to isolate. This method requires nicely separated cells, preferably spherical, to work effectively. One rather large downside to this technique is the equipment itself, which is both very expensive and technically difficult to operate.

Laser-capture micro-dissection

Laser-capture-microdissection (LCM, figure 1.8) is a precise dissection technique that circumvents the need for cell dissociation. To isolate the cells, whole tissue samples are flash frozen and then cryo-sectioned. Different approaches have been applied to visualise the desired cell types, such as rapid immuno-staining, or GFP expression. There are several methods for cutting out and collecting the cells for analysis, each of which have differing advantages and disadvantages (105, 106). For example, Leica platforms use UV lasers to cut around the cell and simply let the cell fall into a collection tube by gravity. Whereas platforms from Arcturus Biosciences, such as the ArcturusXT, use UV lasers to cut around the cell, followed by an IR laser to heat-adhere the cell to a membrane, which can be lifted away from the remaining tissue.

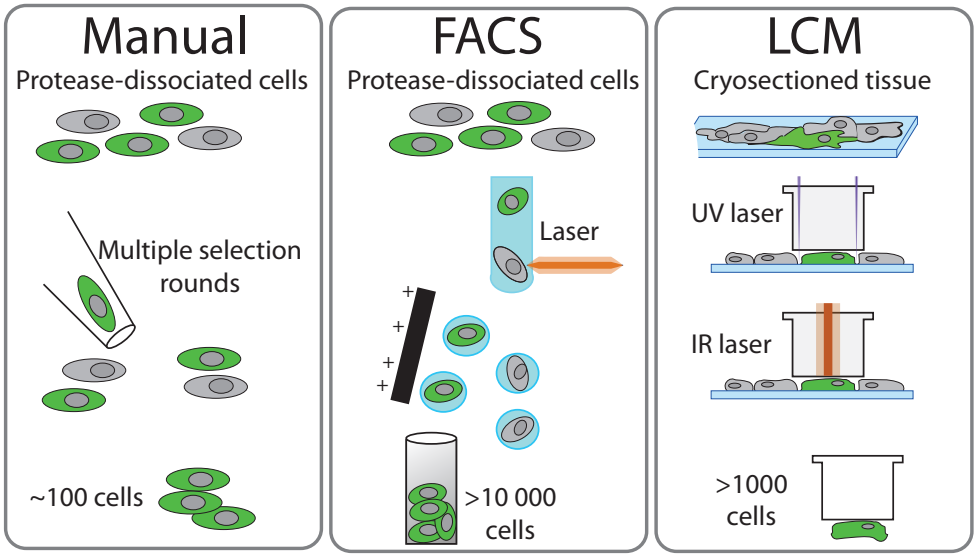


Figure 1.8: Whole-cell isolation methods - Methods for isolating whole cells from a complex tissue composition include manual isolation, fluorescence-activated cell sorting (FACS) and laser-capture microdissection (LCM). Manual sorting involves multiple round of manual selection of GFP-positive, protease-dissociated cells; typical yield is low. FACS involves running protease-dissociated cells in buffer, single-file past a detector; the buffer stream is electronically controlled to break off phased-droplets. The detector measures when a GFP-positive cell is present in the flow, and charges the droplet that contains the positive cell. Number of cells collected can be in the high thousands. A magnetic charge then directs the charged droplet towards a collection tube. LCM involves cryosectioning tissue samples and identifying the labelled cells of interest with a microscope. A UV laser is used to cut out the cell from the other tissue, and an infrared laser is used to adhere the cell to a special cap so that the cell can be removed. Thousands of cells can be collected.

Comparing the whole-cell isolation techniques

FACS, LCM, and manual isolation approaches have both advantages and disadvantages to consider. Firstly, there are differences in how the cells are treated. Both manual sorting and FACS use protease-treated cells, a potential stress, and the cells are intact during isolation, which may result in gene expression changes due to the changed environment. Because the cells are dissociated, non-specific RNAs from lysed cells may stick to the specific cells, contaminating the samples. This problem is overcome in LCM since the tissue is intact and fixed to a slide for the entire protocol. However, although LCM systems can reach the accuracy of one micron, contamination by neighbouring cells or missing segments (e.g. axon branches) in tightly packed tissues it is to be expected (for a comparison see (107)). Manual isolation has the advantage over the other techniques in that there is a relatively simple setup and no special equipment required, providing significant cost benefits. FACS and LCM require equipment that is expensive, often overbooked, and requires specialist training to use.

1.3.3 Nuclei isolation for transcript and chromatin analysis

The whole-cell transcriptome provides an overview of the steady state mRNA levels of the cell. However, by analysing the newly synthesised, unprocessed transcripts within the nucleus, the transcriptional rate of gene expression can be obtained. Analysing the chromatin state, and chromatin-bound proteins can provide information about how genes are regulated in specific cell types. However, the current genome-wide methods for chromatin analysis are not as sensitive as RNA analysis, and require comparatively more material. Nuclei isolation is a more reliable and efficient way to isolate chromatin from specific cell types than isolating whole cells.

Fluorescence-activated nuclei sorting

FACS is frequently used to sort nuclei, where it is aptly named fluorescence-activated nuclei-sorting (FANS). Nuclei are relatively easier to sort compared to whole cells, perhaps because nuclei are more uniform in size and shape than cells. The approaches used to label the nuclei of interest include either expressing a fluorescent protein such as GFP, nls::GFP, and H2B-GFP (108, 109), or immunofluorescence staining of a tag or a nuclear-localised endogenous protein e.g. SBP-H2B, or NeuN (110, 111) (table 1.1). Depending on the intended use of the sorted nuclei, they can either be cross-linked with formaldehyde (for ChIP) or left in the native state (for native-ChIP or RNA isolation). Unfixed nuclei lose their integrity more easily, whereas fixed nuclei are

prone to clumping. It is crucial to avoid nuclei clumping during sorting, as this reduces efficiency, results in impure samples and can clog the machines.

Isolation of nuclei tagged in specific cell types

Isolation of nuclei tagged in specific cell types (INTACT) is an alternative method for isolating nuclei that eliminates some of the limitations of FACS (figure 1.9). INTACT was originally developed in *Arabidopsis* (112), then quickly adapted to *C. elegans* and *D. melanogaster* (113, 114). INTACT relies on expressing a tag localised to the outer surface of nuclei, and two different tagging approaches have been published (table 1.1). The original approach used a more complicated construct containing a WPP domain, which inserts into the outer surface of the nuclei, GFP for detection, and a BRLP. The BRLP is biotinylated *in vivo* by co-expressing BirA. The second, more straightforward tagging approach used Unc84 to insert into the nuclear envelope, and two GFPs in tandem. The tagged nuclei are immuno-purified using anti-tag magnetic beads, and the non-specific populations are washed away.

Table 1.1: Transgenic proteins for nuclei isolation techniques

FANS	INTACT
NLS-GFP	WPP-GFP-BLRP
H2B-GFP	Unc84-2xGFP
H2B-SBP	
NeuN-staining	

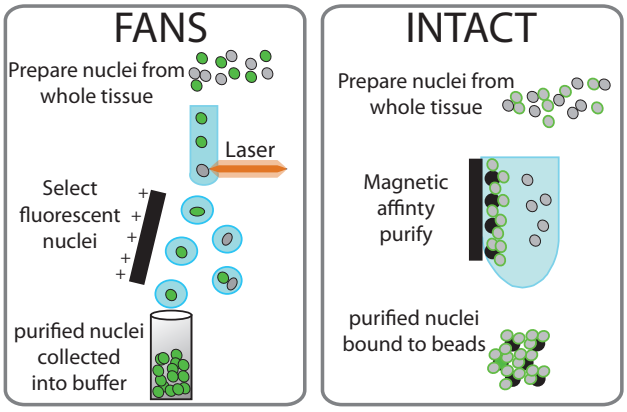


Figure 1.9: Methods for isolating nuclei from specific cell types - Fluorescence-activated nuclei sorting (FANS) and isolation of nuclei tagged in specific cell types (INTACT) are the two current methods for isolating nuclei of specific cell types. FANS follows the same principals of FACS, except that nuclei are prepared for sorting. INTACT involves immunopurifying nuclei labelled with an outer-nuclear envelope construct.

Comparing the nuclei-isolation techniques

There is little difference between these nuclei-isolation methods in how the nuclei are prepared, just whether the material is cross-linked or not. FANS generally uses cross-linking, whereas INTACT is generally carried out under native, non-crosslinked conditions, with the exception of *C.elegans* (115). The major difference between FANS and INTACT is the amount of material and time required. Just like FACS of whole cells, sorting nuclei is a time-consuming and inefficient process, which needs high amounts of starting material. INTACT is comparatively much faster than sorting nuclei and a higher yield of nuclei can be isolated from lower starting material. INTACT is particularly appealing when considering the long time-frames for sorting enough material for a ChIP-seq experiment (∼8 hours for 40 x 106 nuclei (110)) and the efficiency of sorting is relatively low (∼ 50 % loss (116)). On the other hand, sorting nuclei with the cell type of interest marked by staining against an endogenous protein circumvents the need to express a possibly toxic transgene.

1.3.4 Biochemical methods for transcription and chromatin profiling

One of the biggest challenges in obtaining cell-type-specific information is collecting sufficient material for the experiment. Biochemical methods have a big advantage over other isolation techniques in that they are rapid, simple and highly efficient. The high efficiency of biochemical techniques means that each experiment requires far less starting material than other techniques. There are several options for both RNA and chromatin analyses that enable profiling of small populations of cells without demanding high amounts of starting material.

Direct tagging and affintiy purification of RNA-binding machinery

The two most commonly used targets for this experimental approach are polyA-binding proteins (PAB, (117)) or the ribosome (TRAP or ribo-tag (118, 119)), shown in figure 1.10(top left). The PAB-based method was first developed for profiling *C. elegans* muscle cells (117). The technique is relatively straightforward: Flag-PAB1 is expressed using a muscle-specific promoter, the worms cross-linked, the Flag-PAB-1 affinity purified, and the associated RNAs purified. This approach was later applied to profiling *Drosophila* photoreceptor cells (120), *C. elegans* neurons and then different cell types and developmental stages in *C. elegans* (121, 122, 123, 124). No further PAB studies have been published for *Drosophila*, possibly due to the lethality of the transgene when expressed under the control of various Gal4 drivers.

Translating Ribosome Affinity Purification (TRAP)has been the method of choice for RNA profiling in mouse (118, 125, 126, 127). In this method, the ribosomal subunit L10A is tagged with GFP and expressed in the cell-type of interest. A polysome extraction is made from whole tissue, and the tagged ribosomes, along with their mRNA cargo, are immuno-purified. This technique is well established in mouse, and there are several BAC-TRAP lines available for specific cell types. TRAP was also used in *Drosophila* to profile a population of $\sim$ 200 cells in the pars-intercerebralis .

TU-Tagging

Originally developed for *Drosophila* (128), and recently transferred to mouse studies (129), TU-tagging is a method to analyse newly synthesised and non-coding RNAs in specific cell types (figure 1.10, top right). This method works by expressing the *Toxoplasma gondii* uracil phosphoribosyltransferase (UPTR) in the cell type of interest. UPTR acts to couple ribose-5-phosphate to uracil, generating a uridine monophosphate that is incorporated into RNA (130). Feeding flies or mice the uracil analogue 4-thiouracil leads to thio-labelled RNA in the UPTR-expressing cell type. The thio-labelled RNA are then coupled to biotin, and the biotin-thio-RNA is purified using streptavidin.

Chromatin profiling without cell isolation

There are two recently published techniques, Chromatin Affinity purification from Specific cell Types (CAST-ChIP) and TargetedDamID (TaDa) (figure 1.10, lower panels), that can be used to profile protein-DNA interactions in specific cell types without the need for isolating cells or nuclei. CAST-ChIP is a method to directly pull-down chromatin-bound proteins from specific cell types, using a modified ChIP assay (34). A tagged form of the desired protein is expressed in the cell type of interest. Then, the whole tissue is cross-linked and fragmented to release the chromatin, followed by ChIP-seq using an antibody against the tag. This technique has been published for Pol II (eGFP-RPB3) and H2A.Z-GFP profiling of neurons, glia and fatbody in the *Drosophila* head.

TaDa is a method developed to profile Pol II binding in *Drosophila* neural stem cells with no cell isolation, crosslinking, or immuno-precipitation (131). TaDa is based on DamID technology (132), and cleverly solves a major issue with expressing Dam-fusion proteins in specific cell types with Gal4 drivers: that Dam-fusion proteins are toxic when expressed at high levels. Southall et.al.,(2013) (131) resolved this issue by

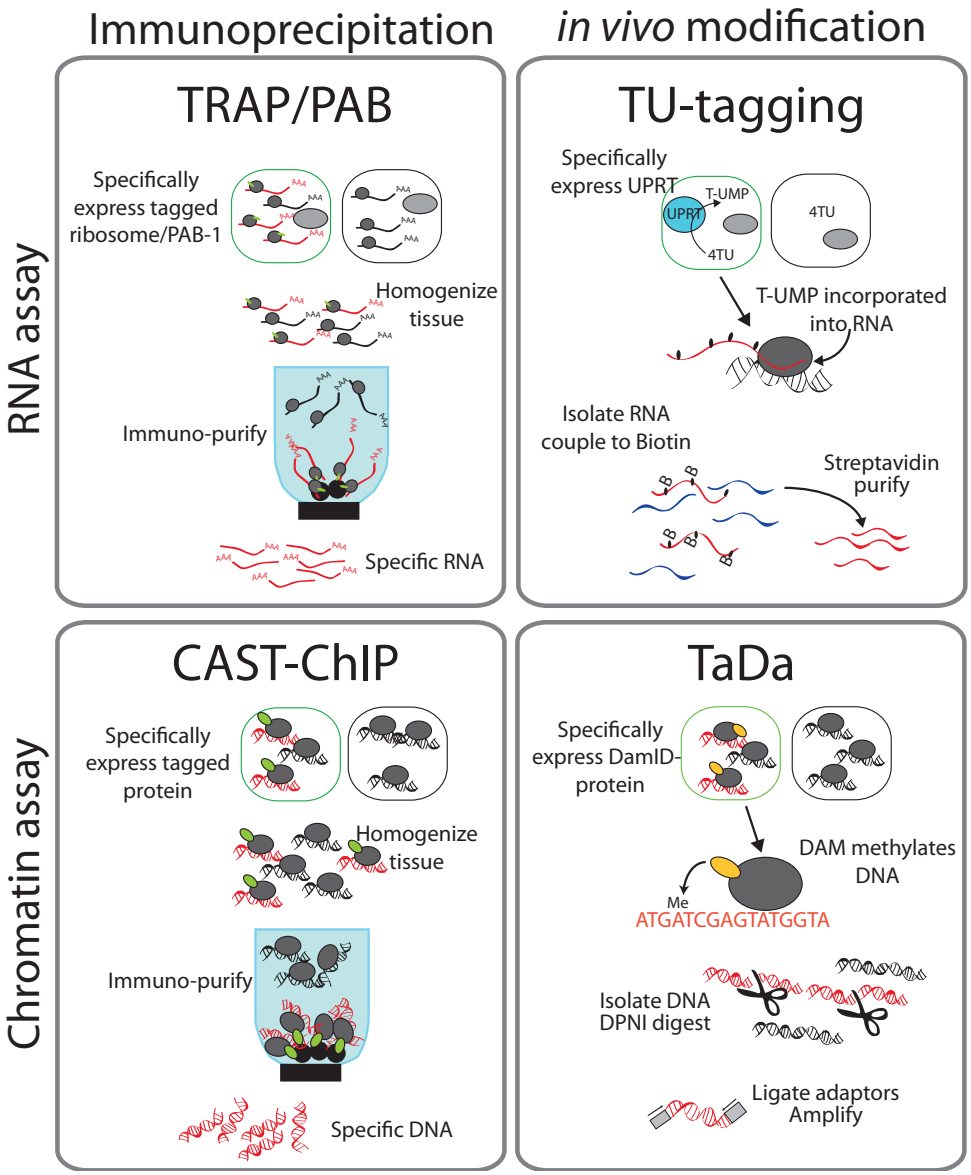


Figure 1.10: Biochemical-based methods for RNA and chromatin analysis - Translating ribosome affinity purification (TRAP) and PolyA-binding protein affinity purification (PAB): Tagged RNA-binding proteins are expressed in a specific cell type. After the tissue is homogenized, the tagged RNA-binding proteins and the associated RNAs are immunopurified. A related approach is used for chromatin analysis using Chromatin affinity purification from specific cell types (CAST-ChIP), except that the tagged proteins bind chromatin and the associated DNA is isolated. TU-tagging allows cell-type-specific RNAs to be directly isolated. By expressing the protozoan enzyme UPRT in specific cells, 4TU is incorporated into RNA. The tissue is homogenized, the RNAs isolated, then the 4TU-labeled RNA is then biotinylated and streptavidin-purified. In targeted DamID (TaDa), a bacterial sequence-specific DNA methylase (Dam) fused to a Pol II subunit is expressed cell-type-specifically. The Pol II-Dam fusion methylates the DNA near where Pol II binds, the chromatin is then isolated, and digested with the methylation-specific restriction enzyme *Dpn1*. The methylated DNA is then amplified and sequenced, providing an indirect measurement of where Pol II was associated.

developing a bi-cistronic construct containing an mCherry primary ORF and the Dam fusion in the secondary ORF. When cell-type-specific Gal4 is active, the construct is transcribed at high levels, but the translation of the DamID fusion is low and non-toxic. The DamID fusion methylates the DNA around its binding site, and the genomic DNA is then extracted and digested with the methyl-specific restriction enzyme Dpn1. The cut (methylated) DNA is ligated to sequencing adapters and amplified by PCR for microarray or sequencing analysis.

Comparing the biochemical methods

The main concern for the biochemical methods is that tagging and over-expressing the protein of interest perturbs the function of that protein. Therefore it is essential to ensure the tagged form of the protein is functioning in the same way as the endogenous protein (for instance, by performing rescue experiments). Another concern is that differences in transgene expression between the cell populations differ, which could skew the results. For example, if there is much higher expression of the fusion protein in one cell type, then more regions are likely to be identified purely because they are more occupied by the fusion protein in that cell type. Of the RNA profiling techniques, TRAP has been the most widely published, possibly due to concerns over the toxicity of tagged-PAB or 4-thiouracil (133). However, TRAP is only for isolating mRNAs that are bound to the ribosome, and cannot give information about non-coding RNA expression or microRNAs. Because the ribosome-bound RNAs are many RNA-processing steps away from active transcription, it may be difficult to compare chromatin features with gene expression levels using TRAP data. For chromatin analysis, CAST-ChIP and TaDa are limited to profiling proteins that can be functionally tagged and over-expressed, and cannot be used for mapping post-translational modifications e.g histone marks. TaDa has the benefit over CAST-ChIP of not needing immunoprecipitation, as a very good antibody to the tag is necessary for CAST-ChIP to work effectively.

1.3.5 Bioinformatics analysis

In order to identify differences between cell types, the genomic data needs to be analysed with specifically designed statistical software. There are several software packages available that compare data counts at features such as ChIP-seq peaks or RNA-seq transcript annotations. These include Cuffdiff (134), baySeq (135), edgeR (136), DeSeq (137), diffreps (138), and BitSeq (139) to name a few (see reviews (140, 141). Tools of the Bioconductor project, and most other tools, are publicly available (see

www.bioconductor.org). Analysing datasets from specific cell types can follow two approaches: either comparing one cell type to the "input" or"total" tissue, or comparing cell types to each other. The"input/total" sample may be an easily dissectible body part, organ or tissue, for example whole head, whole brain, or the striatum of the mouse brain. Finding significant enrichment over the input is difficult in cases where a high percentage of the labeled cell type is part of the whole, such as neurons in the fly brain, because the two profiles may be too similar. In my work I will identify differences between cell types and cell-type-specific enrichments by using pair-wise comparisons. The output of these analysis is a list of genes that is specific to each cell type, and the knowledge of how these cells may perform and maintain their function.

1.3.6 Validation of cell-type specificity

Once a gene is identified as cell type specific, the specificity should be validated to ensure that the measurement is not an artefact. The two commonly used options to evaluate the specificity of results found by ChIP or RNA profiling, are computationally and with fluorescent imaging. Most studies use a computational validation such as gene ontology (GO) analysis, which is a relatively quick way to determine the characteristics of the genes identified in each dataset. This type of analysis evaluates whether the expected characteristics are enriched in the dataset, for example ion-channel activity being enriched in neuron-specific datasets (34, 114). However, one caution that comes with GO-analysis is that the annotation is far from complete, and will be more difficult for less well-studied cell types. Another validation approach is to compare the cell-type-specific datasets to existing data from dissected organs or tissues. In *Drosophila*, there is a extensive resource of microarray data from dissected tissues at different developmental stages known as FlyAtlas (142). Although the dissected tissue may be quite heterogeneous, it can provide an estimate of whether a gene should be enriched in a particular tissue. Immunohistochemistry followed by fluorescence microsopy is often used to complement the computational validation. Depending on the resources available for the gene of interest, several approaches can be take. For example, Bonn et al. used fluorescent in situ hybridization to test co-localisation of a Drosophila mesoderm marker (Mef2) with the RNA of the mesoderm-specific genes they identified (81). Another approach would be direct staining of the gene product, or driving a reporter gene with the gene of interest promoter region (such as enhancer-TRAP). Confirming co-localisation of the newly identified cell-type-specific gene with known markers from that cell type is currently the best method for visualy validating the results from the genome-wide analysis.

Chapter 2

Aims of the project

The overall aim of this research is to understand how chromatin structure regulates transcriptional acitivity in specific cell types of the *Drosophila* brain.

2.1 First aim

Develop a tool to map histone modifcations, nucleosome occupancy and expression level in *Drosophila* neurons and glia

Analysis of chromatin and gene regulation in specific cell types requires experimental tools. As discussed in the introduction, there are several available options for isolating and analysing chromatin and RNA from specific cell types. As I aim to map post-translational modifications of histones as well as measure transcription, I need to use a method that is suitable for these analyses. I chose to use Fluorescence-activated nuclei sorting (FANS) for my experiments, as FANS is a well established method, in terms of available literature, for isolating both chromatin and nuclear RNAs of specific cell types. One key drawback of using FANS is that it is inefficient. A lot of input material and time is required to isolate enough material for genome-wide analysis. Therefore, I aim to thoroughly optimize and stream-line the genome-wide analysis of specific cell types to generate a FANS-based protocol that is highly efficient with both time and material. This optimization is essential in order to make the other aims of this project achievable.

2.2 Second aim

Determine how comparable the CAST-CHIP Pol II, TRAP and FANS RNA-seq methods are in identifying cell-type-specific genes

With the sequencing of the nucRNA from FANS-isolated neurons and glia complete, we will have three available datasets that measure the gene activity in neurons and glia. These three methods: CAST-ChIP of RPB3 (Pol II), FANS-nucRNA and TRAP all measure different steps in the gene-expression cycle: Pol II recruitment, elongation and translation. I aim to compare these methods in terms of the genes identified, to determine how well the gene calls overlap between the methods. Also, if there are any differences, whether using a combination of techniques will give us a better insight into how specific gene groups are regulated.

2.3 Third aim

Analyze the chromatin state of neuronal-enriched, glia-enriched and invariant genes

Many features of chromatin states have been mapped out using large-scale datasets from cell cultures, however these models of chromatin states have not been tested *in vivo*. Whether the cell culture model of gene regulation exists within the organism is a key question from these studies. Using neurons and glia are ideal for addressing gene regulation *in vivo* as they are well characterised, so we already know much about the function of them, and they are post-mitotic, thus removing any complicating factors due to DNA replication. Using this system to study cell-type-specific gene regulation *in vivo* we have already shown that the histone variant H2A.Z marks cell-type-invariant genes as identified by Pol II binding(34). This observation highlights the benefits of refining the resolution at which we study gene regulation mechanisms. In chapter 3, I will determine the chromatin state of different sets of genes, identified as cell-type-specific or invariant from different types of gene classification: by Pol II binidng or nucRNA analyses. These analysis may reveal unique chromatin features that demarcate the specific expression pattern of the gene groups. In chapter 4, I will explore the relationship between nucleosome architecture, H2A.Z deposition and Pol II regulation.

Chapter 3

Measuring gene expression and chromatin states in specific cell types

3.1 Summary

In this chapter, I demonstrate the development of a fluorescence-activated nuclei sorting (FANS)-based method for mapping histone modifications, nucleosome positioning and total RNA in neurons and glia. I have optimised a protocol that minimises the number of input nuclei required for each assay, greatly improving the applicability of the tool for other applications. I compare the gene calls identified as neuron-enriched or glia-enriched with gene groups previously identified by the Chromatin affinity purification of specific cell types (CAST)-ChIP Pol II binding, and Translating Ribosome Affinity Purification (TRAP). I show that there is more overlap of gene calls that are neuron-enriched than those that are glia-enriched. The differences in genes identified between the three techniques reveal unique enrichments of gene functions, indicating that the specific gene classes may be regulated at different stages of the gene-expression process. Comparing a range of cell-type-specific tools, which measure different stages of gene expression, is thus useful in further teasing-out the complexities of how neurons and glia achieve their specialised functions.

3.2 Introduction

CAST-ChIP is a method that allows genome-wide profiling of chromatin-bound factors in specific cell types. This method was used successfully for mapping Pol II binding and the histone variant H2AZ in neurons, glia, and fatbody of the adult *Drosophila*

head (34). Some of the major questions arising from the CAST-ChIP analysis were how the Pol II recruitment to the neuronal and glial-enriched genes was regulated, and what the activity status of those Pol II-bound genes was. To address these questions, the chromatin state of neurons and glia would need to be measured, including assaying histone modifications. CAST-ChIP is not a suitable tool for such analysis as it cannot be used for assaying post-translational modifications such as histone modifications or the phosphorylation status of RNA Polymerase II. Therefore, a technique capable of measuring a range of chromatin features would be necessary. Isolating nuclei of specific cell types is the simplest method for obtaining cell-type-specific chromatin. There are currently two techniques for isolating nuclei from specific cell types, FANS and Isolation of Nuclei Labelled in Specific Cell Types (INTACT), which are discussed in detail in the introduction 1.3.3. Both the INTACT method (data not shown) and FANS method were tested for this work, however, FANS was able to be established more rapidly. Therefore, FANS-isolation was used as the method of choice for this work.

One of the major benefits of isolating nuclei from the cell type of interest, instead of a biochemical approach such as CAST-ChIP, is that any histone mark, transcription factor, or chromatin-binding protein can be assayed within the same genetic background. The type of chromatin feature assayed is only limited by the availability of antibodies to that feature, so many more different types of assay are possible with minimal adjustments to the basic nuclei isolation protocol. Thus the effects of different genetic backgrounds, or differential expression of tagged reporter genes, is greatly diminished for these approaches compared to biochemical approaches. In this work, I will assay three different histone modifications, map the occupancy of nucleosomes, and measure the RNA transcript levels within neurons and glia. These analysis will address how neuronal and glial genes are differentially regulated.

Nucleosomes govern access to the genome. By measuring the occupancy of nucleosomes within the different cell types, I hope to gain a better understanding of how the neuron and glia genes are specifically regulated. For example, differences in nucleosome occupancy between the two cell types could indicate regions that act as enhancers in one cell type but not the other. Determining the underlying transcription factor motifs at these nucleosome-differential regions could provide vast information about what transcription factors could regulate specific gene sets. The occupancy of nucleosomes across the promoter is also known to be different between specifically regulated and constitutively expressed genes. Comparing nucleosome positioning and occupancy between neurons and glia at specific gene subsets could provide a better understanding

of the mechanisms governing the specialisation of gene expression in each cell type. Measuring the histone modifications and nucRNA levels of neurons and glia will allow me to assess the activity state of the genes in each cell type. The histone marks H3K27ac and H3K36me3 are two general markers for active genes. H3K27ac is also used as a proxy for active enhancers, thus measuring the binding pattern of this modification and comparing it to the differential nucleosome regions would aid in identifying active enhancers that are specific in each cell type. The histone modification H3K36me3 is an indicator of transcriptional elongation, however the role of this modification in transcriptional regulation is complex. This mark is established during transcriptional elongation, but acts as a repressive mark by recruiting a histone deacetylase complex and reducing nucleosome turnover. Interestingly, this mark is absent from some specifically regulated genes, even when those genes are active (76). Thus, the role of H3K36me3 and how this mark, or absence of the mark, affects the gene expression program within specific cell types is not clear. As most previous studies into the biochemistry of H3K36me3 have been in yeast, and genomic studies in *Drosophila* have been in cell cultures, studying how this histone modification marks genes within specific cell types will be highly interesting.

The CAST-ChIP data revealed sets of genes with Pol II recruitment to the promoter that was specific to one cell type. Yet, the binding of Pol II to a promoter does not necessarily mean that the gene is transcriptionally active. The Pol II could be bound to the gene and maintained in a "paused" state (see 1.1.1), awaiting a specific signal before transcriptional elongation commences, such as for the hsp70 gene. Alternatively, the Pol II could be transcribing the gene at very high levels. Thus, a complementary method for determining the gene activity would be beneficial. An alternative method, Translating Ribosome Affinity Purification (TRAP), has also been used in my host lab to map mRNAs bound to ribosomes in neurons and glia (T. Schauer, LMU *unpublished*). This method revealed many more genes enriched specifically in each cell type compared to CAST-ChIP of Pol II. However, this method is also not a good measure of the actual transcriptional activity of genes within neurons and glia. Those mRNAs bound to ribosomes have been completely processed, and may have been bound to the ribosome long after the transcription of the gene has stopped. Therefore an additional dataset, which measures the transcript levels of genes within the nucleus, will provide a better approximation of the actual transcription status of genes within neurons and glia.

Having the CAST-ChIP, TRAP, and nucRNA datasets provides the opportunity to assess how comparable three state-of-the-art techniques are for measuring gene activity

in specific cell types. All techniques measure different steps of gene regulation, so comparing all three could give an insight into which genes are regulated at what stage of gene expression. For example, genes that have specific Pol II recruitment but are not identified in the RNA-seq methods could indicate genes that are regulated by Pol II pausing, and a difference in RNA would only be seen under the right environmental conditions. Defining the genes that are cell-type-specific in all methods, or only in a single method could aid in discerning the different regulatory pathways that lead to the specific expression of that gene in that cell type.

Aims

- Develop and optimize FANS-based approaches for MNase-seq, ChIP-seq, and RNA-seq of neurons and glia.
- Identify genes specifically enriched in neurons or glia using nucRNA-seq analysis.
- Map the nucleosome occupancy in neurons and glia using MNase-seq.
- Assay histone modifications H3K27ac, H2K36me3, and H3K27me3 in neurons and glia using ChIP-seq.
- Compare neuron-enriched or glia-enriched gene calls between three cell-type-specific tools: nucRNA-seq, CAST-ChIP, and TRAP.

3.3 ChIP-seq, MNase-seq and nucRNA-seq using FANS

3.3.1 Generating reporter lines for isolating nuclei

The initial challenge of this project was to develop a method for analysing gene activity and chromatin state within specific cell types of the *Drosophila* head. Although the CAST-ChIP method was already developed in my host lab, this method does not enable analysis of post-translational modifications, and requires an epitope tagged transgene for each type of protein that is assayed. Because my aims included assaying histone modifications, as well as comparing multiple chromatin features and RNA expression between the cell types, I chose to develop a Fluorescence-Activated Nuclei Sorting (FANS) protocol. A FANS method has the benefit that it provides the flexibility to perform ChIP-seq, MNase-seq and RNA-seq analysis within the same genotype for each cell type, and with the same isolation method. Thus, experimental variation such as transgene expression level and efficiency of immunoprecipitation between the experiment types was not relevant. This is because each experiment, whether ChIP-seq, MNase-seq or nucRNA-seq, uses the exact same fly line and selection criteria for

each cell type assayed. Because the amount of labelling does not affect the isolation efficiency and the number of nuclei collected can be precisely counted, the experimental errors between cell types is also minimised. For all experiments carried out in this work, there are only three genetic backgrounds used: wild-type, *elav*::H2B-GFP (for labelling neurons with H2B-GFP), and *repo*::H2B-GFP (for labelling glia cells with H2B-GFP). Elav is a transcription factor that is required for neuronal cell fate determination. The *elav* promoter is commonly used to drive Gal4 expression specifically in most neurons of the adult *Drosophila* brain (143). Repo is a transcription factor that regulates glial differentiation, and maintains glial identity (144). The *repo* promoter is used to drive transgene expression across all glia subtypes. All datasets generated in this work, including datasets not used for the analysis, are listed in table 3.1. Fatbody data sets, which are not part of the analysis presented in this thesis, were collected under the same conditions as neuron and glia datasets, except using the *take-out* Gal4 driving H2B-GFP, to specifically label the fatbody nuclei.

Table 3.1: Datasets produced in this work

Assay	Cell type	replicates
H3K36me3	neurons	2
	glia	2
	fatbody	2
H3K27ac	neurons	2
	glia	2
H3K27me3	neurons	2
	glia	2
MNase-seq	neurons	2
	glia	2
	fatbody	2
	head	1
nucRNA-seq	neurons	2
	glia	2
H3K4me3	head	2

3.3.2 Isolating nuclei from *Drosophila* neurons and glia

To mark the neuronal and glial nuclei for FACS isolation, I chose to transgenically express H2B-GFP (figure 3.1, A). This tagged protein results in very highly fluorescent nuclei in the cell type it is expressed in, which is an important criteria for FANS. To express H2B-GFP in neurons, I used a fly line homozygous for both the *elav*::Gal4 driver construct and the UAS::H2B-GFP reporter gene. Gal4 expression under the control

of the *repo* promoter (*repo*::Gal4) is slightly toxic and, while heterozygous viable, the line is homozygous lethal, therefore making lines homozygous for both Gal4 driver and H2B-GFP is not possible. To overcome this shortfall, I generated a fly line with H2B-GFP directly under the control of the *repo* promoter (described in 6.2).To ensure that the reporter gene expression is restricted to glia, I performed immunohistochemistry on brains dissected from the *repo*::H2B-GFP line (figure 3.1, B, 6.4), staining for GFP, elav protein, and repo protein. This shows that the H2B::GFP expression co-localises with *repo*-positive nuclei, and does not co-localise with *elav*-positive nuclei, confirming that the expression of H2B-GFP is limited to glia. Neuronal-specific expression of H2B-GFP was confirmed previously using immunohistochemistry to ensure H2B-GFP was not expressed in *repo*-positive cells (34).

One concern is that the tagged-H2B may interfere with chromatin structure if H2B-GFP expression is too high relative to endogenous H2B levels. To check the relative amounts of endogenous H2B to H2B-GFP expression I performed a western blot using α-H2B antibody (figure 3.1, C, methods 6.2.3). I compared whole head levels, as well as the relative amounts in FANS-purified nuclei and it is clear that H2B-GFP expression is far lower than the endogenous H2B. The H2B-GFP protein is barely visible compared to the large amount of endogenous H2B-GFP. There is also no difference observed in nucleosome positioning in metagene analysis between wild-type chromatin and H2B-GFP containing chromatin, (section 3.3.4) further demonstrating that H2B-GFP does not interfere with the chromatin structure in this system.

To isolate the specifically labeled nuclei, I used a FACS-ARIA III cell sorter (methods 6.3.3). Preparation of the nuclei for sorting was optimised to increase the efficiency, so as to maximise the yield of specific nuclei from the input. Separating the sample into single nuclei is highly important for obtaining optimal yield. Thus, re-suspending the nuclei in Nuclei-Purification buffer (NPB, methods 6.6) with BSA and passing the nuclei gently through a 22G needle, then a 25G needle, was necessary to optimally separate the nuclei prior to sorting. The appropriate dilution and sorting speed was established with each preparation, with a sorting speed chosen to be less than 10 000 events per second to ensure nuclei were not subject to very high pressures. I used qualitative measures, such as the speed of re-clumping during sorting, and what proportion of events are excluded from the first selection step (figure 3.2, A, gate P1). I used three basic parameters to isolate the GFP positive population: Side scatter (SSC, the density and granularity of the nuclei), forward scatter (FSC, size) and GFP intensity (figure 3.2). SSC and FSC (gate P1) were used to remove debris and very large

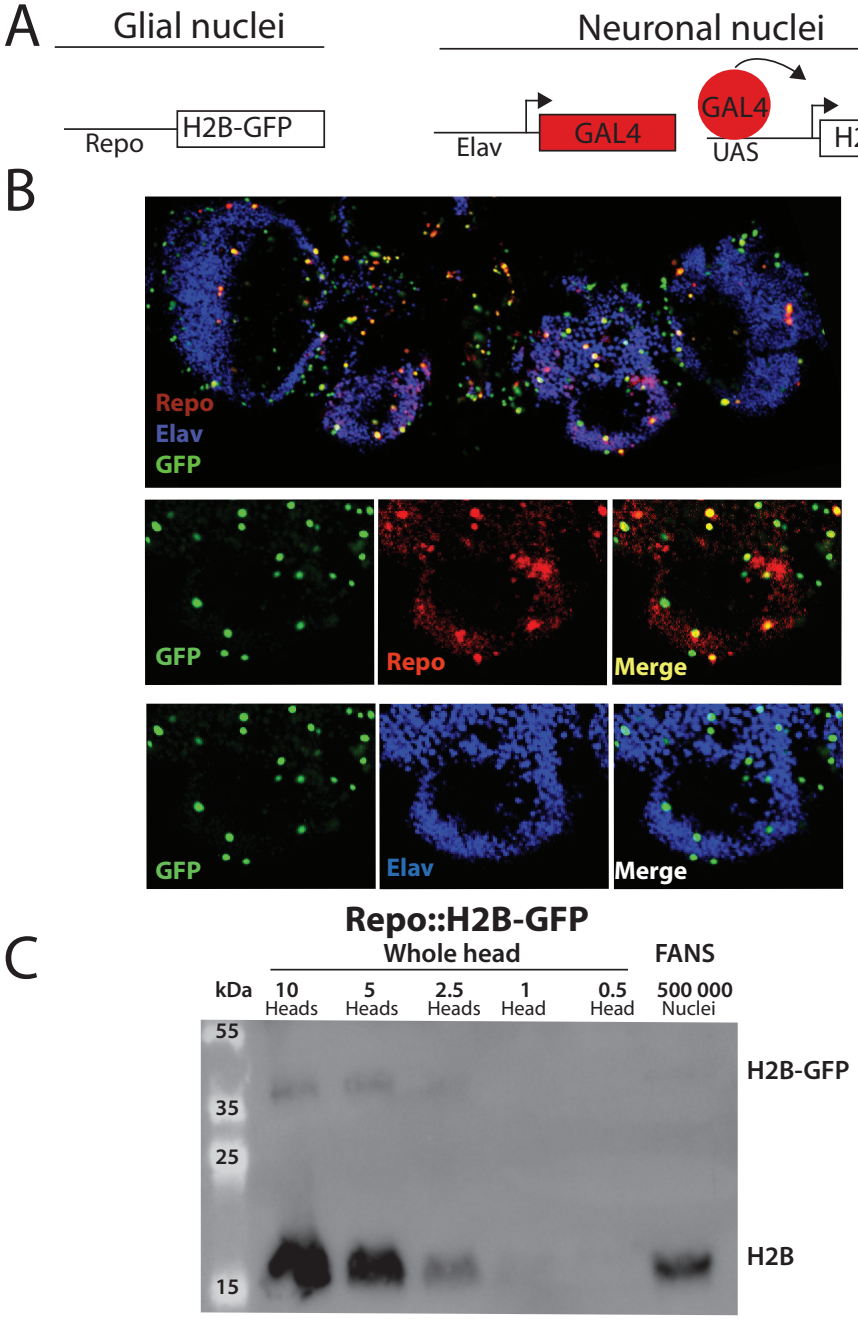


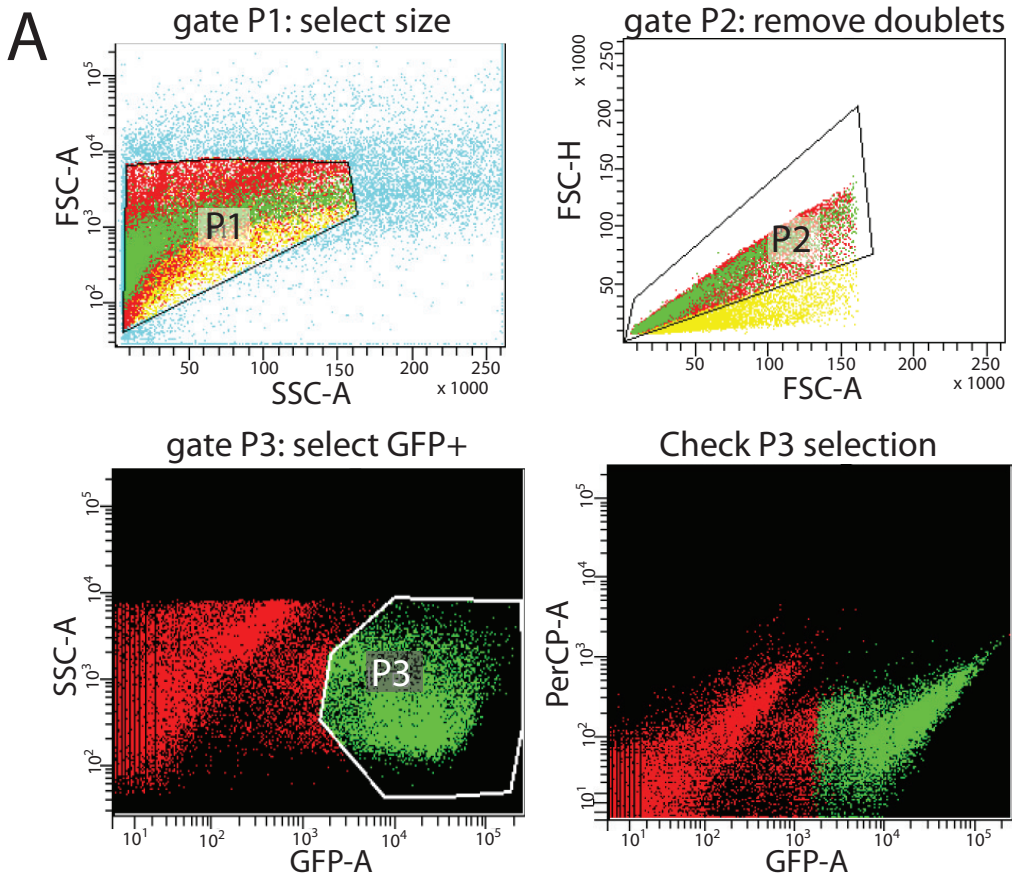
Figure 3.1: Labelling specific cell types in the *Drosophila* brain - A) Schematic of transgenes used to label each cell type with H2B-GFP. For glia-specific H2B-GFP expression, a transgenic fly was generated with H2B-GFP directly under control of the repo promoter. B) Confocal microscopy using immunohistochemistry against GFP, Elav, and Repo, to test specificity of H2B-GFP expression in the repo::H2B-GFP line. C) Western blot showing that the GFP-tagged form of H2B is far less abundant than the endogenous H2B protein.

particles (clumps of nuclei); the second gate compared two different FSC parameters, FSC-H (height of FSC intensity) and FSC-A (area of FSC intensity), to select only those events of single nuclei and eliminate doublets. Finally, SSC and GFP intensity were used to select the GFP positive population of nuclei. An additional plot comparing GFP fluorescence with autofluorescence was used to confirm correct selection of GFP positive nuclei.

At the beginning of each sorting session, a small sample of nuclei ($\sim$ 2000) were sorted and then run through the sorter a second time to measure the accuracy of the sorting. This second sort shows the proportion of non-specific nuclei that are being collected with the GFP positive nuclei, for instance doublets of positive and negative nuclei that were within the size cutoff range (figure 3.2, C). If the proportion of the positive population was less than 90 %, then the gates were adjusted, or nuclei diluted further until the purity of the population was acceptable (over 90 % purity). After collection, the nuclei were also checked with a fluorescence microscope to confirm the GFP fluorescence and also to check the structure of the nuclei were still circular and not destroyed from the sorting process.

3.3.3 Establishing genomic assays for FANS-isolated nuclei

FANS enabled me to isolate populations of neuronal and glial nuclei with high purity. However, one major drawback of this isolation method is that it is time consuming and inefficient. For example, around 35 % of the sample is lost from the first gate, due to clumping nuclei and other debris, and the low proportion of glia in the whole head population (3 %). As a consequence, one day of sorting would be needed to acquire enough material to perform a single standard ChIP assay for neurons, and considerably more time would be needed for analysing glia. To overcome this, I optimised all of the assays for using a limited number of nuclei. I chose to use 1 x10^6 nuclei for each assay, as this was a reasonable number of nuclei to collect, including several technical replicates, during a single day of sorting of glia nuclei. Each assay (MNase-seq, RNA-seq, and ChIP-seq) required different optimisation, which is described in the corresponding sections of each analysis in this chapter; also in the methods (6). This optimisation allowed me to assay many more chromatin features than would be possible following the standard protocols for chromatin and transcription analysis.



B

Elav::H2B-GFP

Population	#Events	%Parent	%Total
All Events	50,000	----------	100
P1	32,203	64.4	64.4
P2	26,025	80.8	52.0
P3	9,630	37.0	19.3

Repo::H2B-GFP

Population	#Events	%Parent	%Total
All Events	50,000	----------	100
P1	32,816	65.6	65.6
P2	28,502	86.9	57.0
P3	869	3.0	1.7

C

Checking purity

Population	#Events	%Parent	%Total
All Events	878	----------	100
P1	874	99.5	99.5
P2	874	100	100
P3	848	97.0	96.6

Figure 3.2: FACS isolation of nuclei - A) FACS-Aria (FACS-diva software) analysis of gating. Gate P1 removes the majority of debris from the sample, gate P2 removes doublets, and gate P3 selects for GFP-positive nuclei. B) Population statistics of sorting. There are a far larger number of neuronal nuclei than glia nuclei in the populations: 37 % compared to 3 % of parent population. C) Control check of sorted material showing that most GFP positive nuclei are truely GFP positive, thus contamination of non-GFP positive nuclei is minimal.

3.3.4 Assaying nucleosome organisation in neurons and glia

One of the key aims of my research was to better understand how neuron-specific and glia-specific genes are regulated. The competition between nucleosomes and effector proteins for binding the DNA means that when and where a nucleosome is present on the DNA has important implications on gene expression. Nucleosome-free regions may contain transcription factor binding motifs that are involved in regulating nearby genes. Identifying the regions that are differentially nucleosome-free between different cell types would enable us to assess if there is differential enrichment of transcription factor binding sites between the two cell types. The differences in transcription factor binding sites could provide clues about which transcription factors are involved in the different regulatory networks involved in establishing and maintaining neuronal or glial cell fate and function.

To map the chromatin structure in neurons and glia, I performed MNase-seq analysis. I developed an MNase-seq protocol that used 1 x10^6 nuclei per sample (methods 6.3.6. Because of the limited material, I did not perform an H3-ChIP after MNase digestion, instead opting to electrophorese the MNase-digest material and purifying the mono-nucleosome band from the gel. The major step for optimisation was to determine the correct amount of MNase needed to achieve around 80 % mononucleosomes (figure 3.3, A and B). Determining the ideal electrophoresis conditions to separate the MNase-digested material was also critical (3 % agarose gel at very low voltage for 5-6 hours). This approach ensured significant separation of the mononucleosome band from the di-nucleosome band, greatly reducing the risk of cross contamination. Two biological replicates of neurons and glia were made, each consisting of two pooled MNase-digestion replicates to ensure the MNase digestion levels were as similar as possible between samples. The mono-nucleosomal DNA fragements for each sample were sequenced using an Illumina hiseq 2000 (EMBL genecore). The neuronal and glial MNase samples were sequenced across a total of 1.5 sequencing lanes each, making a total of approximately 500 million reads for each cell type. Very deep sequencing of the MNase samples was necessary to identify the nucleosome-occupancy differences between the cell types with high confidence (Pawel Bednarz, University Warsaw, data not shown). The sequencing data was mapped to the *Drosophila* genome (dm3) by Pawel Bednarz using Bowtie (145), and the data were processed and analysed using the MNase-analysis tool DANPOS (146).

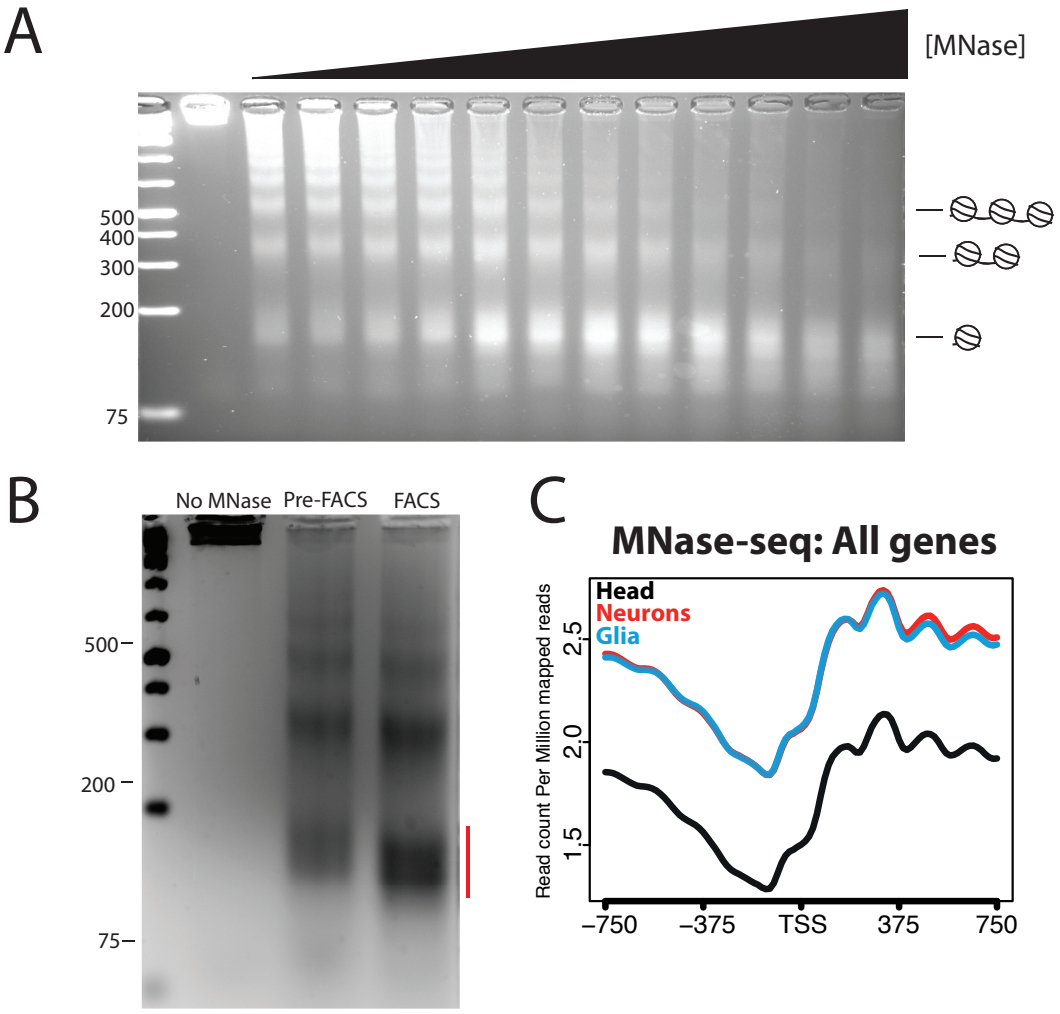


Figure 3.3: MNase-seq to analyse nucleosome occupancy - A) MNase titration to achieve correct mononucleosome enrichment, visualised using agarose electrophoresis and ethidium bromide staining. A two-fold dilution series was used to find the amount of MNase-enzyme to achieve approximately 80 % mononucleosome release. B) MNase digestion of Input and post-FANS samples. The red line indicates the region of the gel that was removed for DNA purification. C) Average profile plots of all genes, using the MNase-seq data aligned to the TSS. The read counts are higher for neurons and glia because additional sequencing was performed and the biological replicates were merged. The pattern between all samples is highly similar.

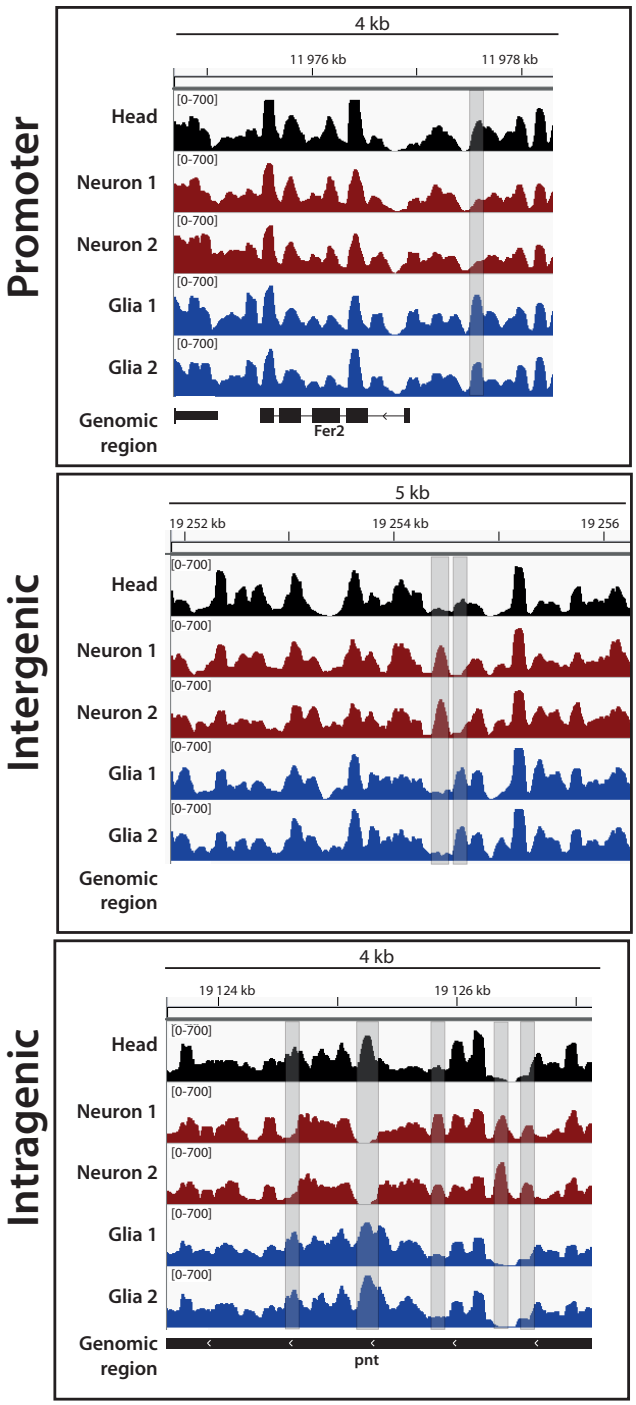


Figure 3.4: Highly specific differences in nucleosome occupancy between cell types - Genome-browser snapshots of MNase-seq data showing that single nucleosome differences can be detected between neurons and glia. The differences in nucleosome occupancy occur in many genomic regions, and can be found near promoters, as well as at intergenic and intragenic loci. The two biological replicates for neurons and glia are shown, demonstrating the highly similar MNase-digestion pattern between the biological replicates. MNase data was processed and smoothed using DANPOS 6.5

One advantage of using FANS to isolate the nuclei is that the number of nuclei in each sample can be precisely counted, thus the level of MNase-digestion between samples is highly similar (figure 3.3, C, methods 6.5). Although the average profile of the MNase-seq data across all genes is highly similar , there are many regions across the genome that have distinctly different nucleosome occupancy between the cell types (figure 3.4, shaded boxes). These differences can be found across all genomic regions, not only near the promoters of genes. In many cases, there appears to be a single nucleosome difference between the two cell types, with the neighbouring regions having highly similar nucleosome occupancy. These differences were quantified bioinformatically using DANPOS (146) by our collaborators Bartek Wilczynski and Pawel Bednarz (University of Warsaw), who are using the cell-type-specific nucleosome-free regions to build a predictive model of specific gene activity neurons and glia. This modelling will be the basis for the PhD thesis of Pawel Bednardz, and so will not be discussed further here.

3.3.5 Assaying histone modifications in neurons and glia

The major reason for using a FANS-based approach in these analyses was to be able to assess the post-translational modifications of histones. Assessing the signature of histone marks across different types of genes may provide information about how those genes are regulated. I assayed two "active" histone modifications, H3K27ac and H3K36me3, and the polycomb-repressed mark H3K27me3 (figure 3.5, A) to assess the chromatin state of genes between neurons and glia. Details of the known functions of these histone modifications are described in the introduction (1.1.3). Before performing the ChIP-seq analysis with antibodies against the histone marks, I generated a general ChIP-seq protocol optimised for 1 x 10^6 nuclei (methods 6.3.4) using an antibody against the Pol II subunit RPB3 (T.Schauer, LMU, *unpublished*). The RPB3-antibody was ideal for establishing the FANS-ChIP protcol as the optimal conditions for this antibody using the standard ChIP protocol, as well as good positive and negative controls, were known. The major contributing factor to the success of this protocol was altering the sonication, specifically changing from a Branson sonifier 250 to the covaris S220 sonicator. By performing a time-course sonication with the covaris sonicator, using aliquots of 1 x 10^6 nuclei, I was able to determine the best sonication conditions for the ChIP analysis was six minutes of sonication time (figure 3.5, B). The immunoprecipitation conditions were then optimised separately for each individual histone mark antibody , by adjusting bead and antibody levels (methods 6.3.5). The conditions were considered optimal if (1) the ratio of positive control to negative

control was greater than 5-fold (methods 6.3.1), and (2) the yield of DNA after purification was high enough for sequencing from a maximum of four pooled technical ChIP replicates. Firstly this ensures that the signal to noise background in the sequencing is optimal. Secondly, this was to reduce the amount of nuclear sorting for each biological replicate as much as possible, to ideally allow completion in a single sorting session. In some cases of glia sorting, one sorting day was not sufficient to collect enough nuclei, so some ChIPs were performed using nuclei that had been sorted on different days.

After optimising the ChIP protocol for each antibody, I performed ChIP-seq analysis of all three histone marks on FANS-isolated neuron and glia nuclei (methods 6.3.5). Each ChIP-seq was performed with two biological replicates, and the input chromatin was sequenced for all H3K36me3 replicates, as a control for back ground signal. Inputs were not sequenced from the other ChIP-seq samples as this is assumed to be identical in all samples. The EMBL-genecore facility performed sequencing library preparations and sequencing (Illumina hiseq 2000). Sequencing data was mapped to the *Drosophila* genome (dm3), and peaks were called using MACS by Pawel Bednarz. A method for identifying differences between ChIP-seq signals between cell types, known as Diff-reps, was also tested, however the output of those analyses has not been optimized. Section C of figure 3.5 shows the average profile plots of each biological replicate for the ChIP-seq data (methods 6.5). Each graph displays the average coverage of all annotated genes (ensembl) across the *Drosophila* genome in the two biological replicates of the two cell types. All replicates for each histone modification looks similar, with the background levels being slightly different between samples (look at flanking regions before TSS and after TES, figure 3.5, C). The average shape of each plot reflects the expected pattern of these histone marks according to the current models. For the active marks, H3K27ac was more highly enriched at the $5'$ ends of genes and H3K36me3 was more enriched at the $3'$ ends of genes. These are the expected patterns since H3K27ac marks active genes (by CBP/P-300, see 1.1.3), and H3K36me3 is placed by Set2 which interacts with the elongating form of the polymerase and thus is restricted to the $3'$ end of genes. The H3K27ac is depleted from the $3'$ end of the gene because H3K36me3 mediates recruitment of the histone deacteylase RPD3, which deacetylates histones, including H3K27ac (see 1.1.3). Thus H3K27ac is restricted to the $5'$ ends of genes, where RPD3 does not de-actylate the histones. The repressive mark H3K27me3 shows an even distribution across the gene body, and forms large domains that can be observed by viewing the data on the genome browser (data not shown).

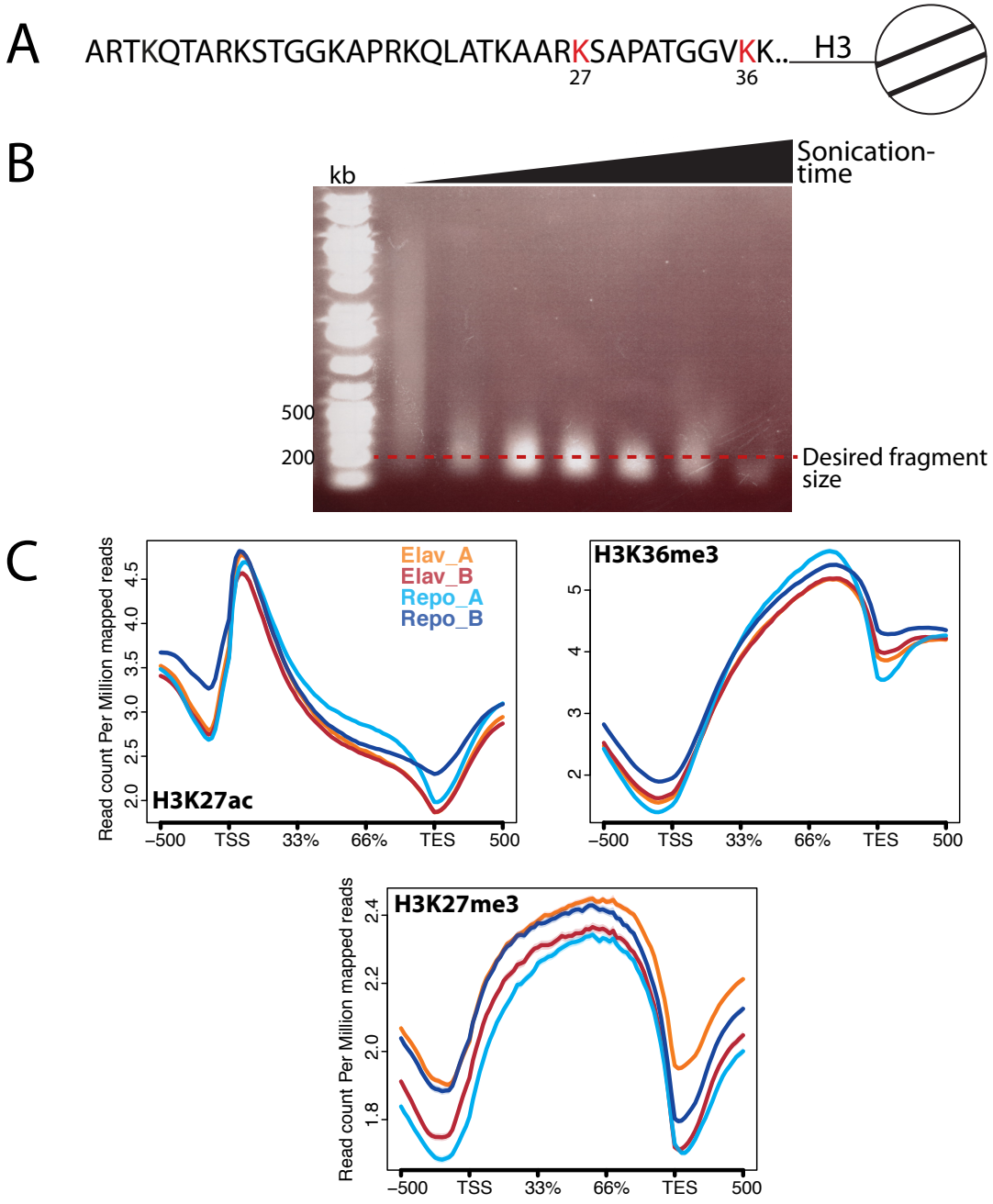


Figure 3.5: ChIP-seq of Histone H3 modifications - A) Schematic of histone H3 N-terminal tail. Histone modifications assayed in this work are H3K27ac, H3K27me3, and H3K36me3. The sites for these modifications are highlighted in red. B) A time course of covaris sonication on 1 x 10^6 nuclei (4, 10, 20, and 40 minutes sonication time) was used to assess the optimal conditions for generating 200 bp chromatin fragments. C) Average profile plots of the H3K27ac, H3K36me3, and H3K27me3 data in neurons and glia, across all genes. The pattern and levels of ChIP-seq enrichment are highly similar between samples.

3.3.6 Nuclear RNA-seq to identify cell-type-specific genes

Before comparing the chromatin states between neurons and glia, I first needed to identify genes that are expressed differentially in each cell type. This would enable me to split genes into different classifications, such as neuron-enriched and glia-enriched, and observe what happens at these genes at the chromatin level. There are already two datasets available from our research group which can be used to identify neuron- and glia-specific genes. These datasets were generated by a previous PhD student in the group, Tamas Schauer. The first dataset contains Pol II binding sites in neurons and glia (34), obtained by a method termed CAST-ChIP described in the introduction (section 1.3.4). The second dataset (T. Schauer, LMU, *unpublished*) was generated using TRAP (section 1.3.4) and provides information about ribosome-bound mRNAs (polyA-selected) in neurons and glia.

Each of these methods can aid in identifying specific gene activity between the cell type, however, they do not measure transcriptional activity directly. The binding of RNA polymerase II at gene promoters does not directly imply gene expression, and cannot be used to directly determine the amount of gene expression. The RNA polymerase may be recruited there and stalled, generating no transcript at all. TRAP analysis is at the other extreme, measuring only those mRNAs that are bound to the ribosome. Theoretically, TRAP should provide a closer estimation of what the proteome of each cell type may consist of, but for measuring gene expression in the nucleus it may be less adequate. Indeed, some genes identified by TRAP could be highly stable mRNAs that are no longer transcribed at all. In this case, the chromatin state of these genes would probably not give any insights into how those genes are specifically regulated, since this would likely be a post-transcriptional process.

To obtain an overview of currently transcribed genes, I aimed to generate a dataset that would more directly reflect the actual transcriptional rate of genes within neurons and glia. I achieved this by isolating and sequencing total RNA from FANS-isolated neuron and glia nuclei (methods 6.3.8. This nuclear-RNA (nucRNA) dataset would also be complementary to the CAST-ChIP and TRAP datasets in that it measures an intermediate step between Pol II binding and translation by the ribosome. The nucRNA data should contain both fully processed and non-processed mRNAs and non-coding RNAs that are greater than 75 bp. I used a library preparation protocol (Nugen, methods 6.3.8) that maintained the directionality of each RNA fragment sequenced, meaning that transcription of non-coding RNAs that occur in the antisense direction

of coding RNAs can be distinguished from sense transcription. Each sample was produced in duplicate, with a corresponding "input" sample. The samples were sequenced on an Illumina hiseq 2000 (EMBL genecore), on a single lane for each sample. The input sample was produced from the same nuclei preparation for each specific dataset. However, the nuclei were run through the FACS machine but the final FACS gate for selecting GFP-positive nuclei was not used. Thus, the input nuclei were not selected for fluorescence but were treated in the same conditions as the specific nuclei, and represent the "whole head" sample for each biological replicate. Mapping of each nucRNA sample to the *Drosophila* genome (dm3) and the DESeq2 analysis to identify enriched genes, was performed by Pawel Bednardz (University of Warsaw).

The mapped nucRNA-seq data was assessed visually using the genome browser to obtain an overview of the quality of the data, and if gene that are expected to be enriched in each dataset have the expected expression pattern. An example of two well characterised genes is shown in figure 3.6. Two transcription factors are generated from the *pointed* (*pnt*) that are soley expressed in glia cells, and are important for establishing glial identity (147). As expected, *pnt* had higher expression in the glial nucRNA datasets than the neuron datasets (figure 3.6, top panel). Neuronal Synaptobrevin (*n-syb*) is a neuron-specific gene ivolved in vesicle fusion at synapses. The nucRNA analysis reveals that *n-syb* has enriched expression in the neuronal data set and depleted expression from the glial dataset, as expected (figure 3.6, bottom panel).

From the read counts I generated a cluster plot using the DESeq2 package to determine the variation between samples (figure 3.7, A). This plot shows that the neuron-specific samples cluster together with the neuron A input sample, and that the remaining samples cluster together. As another test to see if the biological replicates are similar to each other, I generated a principle component plot (figure 3.7, B). The principal component plot shows that the neuron-specific replicates and the glia-specific replicates form discrete clusters. The inputs would be expected to cluster together, however as can be observed in figure 3.7, A, the neuron-A-input sample clusters more closely with the neuron-specific samples. The other input samples cluster closely to each other. As the neuron-A-input sample is more similar to the neuron-specific samples than to the other input samples, I would suggest that this input sample had some selection for GFP fluorescence during the nuclei purification. The purpose of the inputs was to determine if there are any major difference in gene expression between the two transgenic fly lines used in this study. As the neuron-A-input is not suitable for

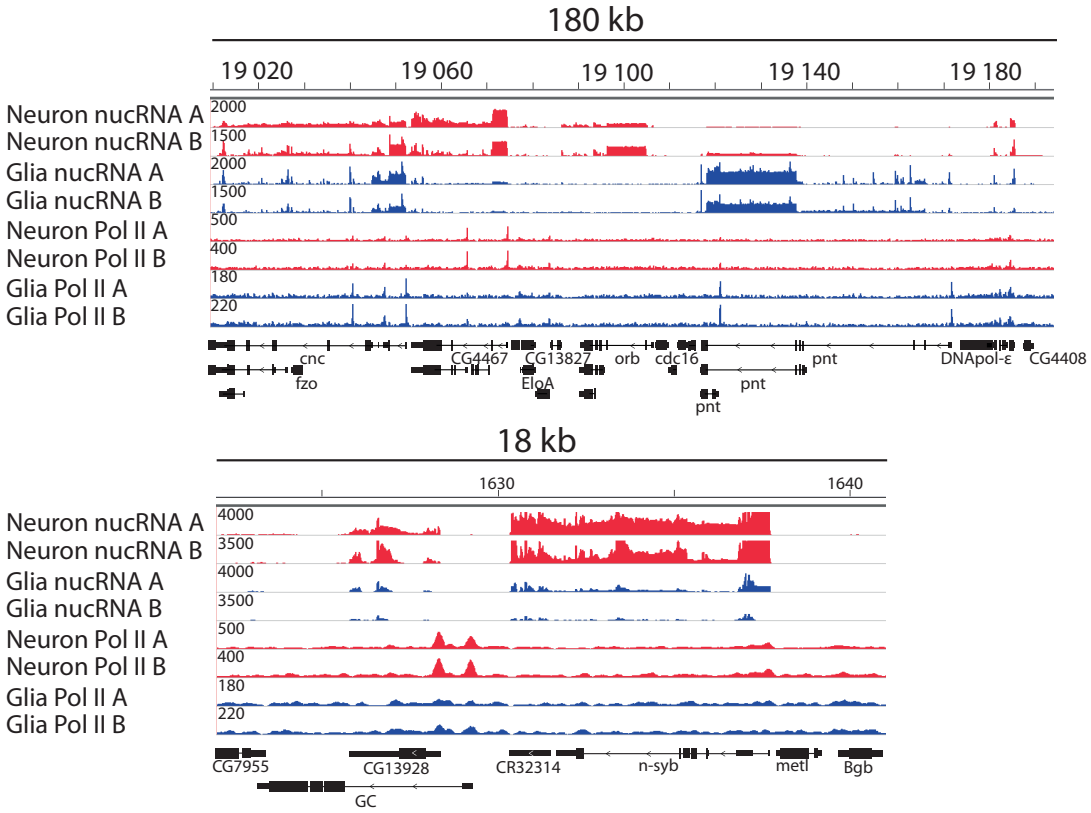


Figure 3.6: Well characterised genes show the expected expression patterns in the nucRNA data - Genome-browser snapshots (IGV) were taken of the nucRNA and CAST-ChIP Pol II datasets at two genes that have been well characterised as specific to each cell type. Pointed (*pnt*) is a known glia-specific gene, neuronal-synaptobrevin (*n-syb*) is a well known neuron-specific gene.

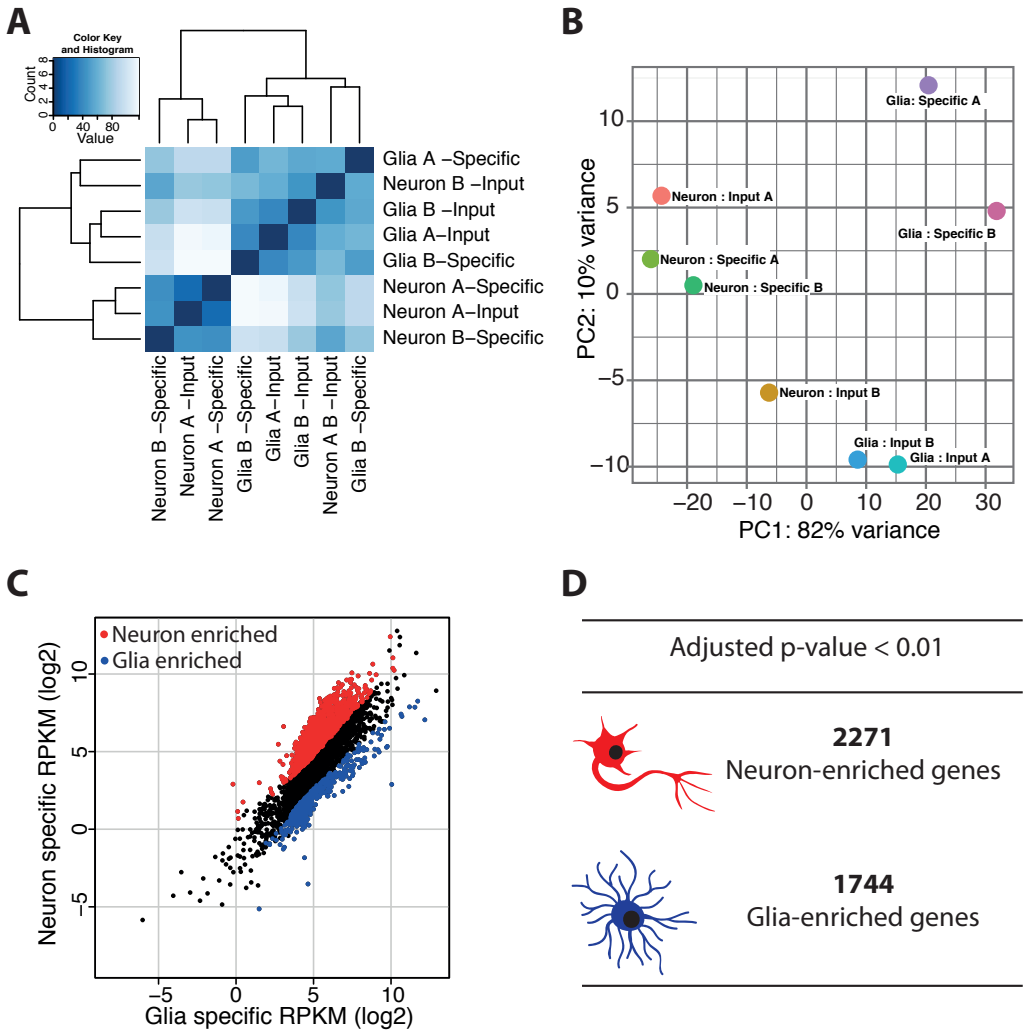


Figure 3.7: DEseq analysis of neuronal and glial nucRNA - A) Clustered heatmap showing the similarity between all nucRNA samples. B) Principle component analysis showing that the cell-type-specific nucRNA measurements are more similar between replicates than between cell types. Neuron input A is more similar to the Neuron-specific samples, indicating a fault with the isolation process. C) Scatterplot of RPKMs comparing the neuronal dataset with the glial dataset. Those genes identified as cell-type-specific from the DESeq analysis are indicated in red (neuron-enriched) and blue (glia-enriched), numbers of genes shown in D. Genes were identified as enriched for each cell type based on the p-value cutoff of less than 0.01 in the DESeq analysis.

this purpose, I did not use these input data and continue the analysis of the cell-type-specific data. To compare the differences in nucRNA transcription between the two transgenic lines, I aim to isolate nuclei from these lines and purify nucRNA without FACS isolation, to eliminate any inadvertent selection of fluorescent nuclei (not shown).

The genes that have differential expression between the two cell types were selected from the DESeq2 output for having an adjusted p-value of 0.01 (Wald-test implemented in the DESeq2 package). These are the standard cutoff criteria, and the same criteria that was used for the analysis of the TRAP dataset, thus I chose the same parameters for my nucRNA analysis. The analysis was performed between the neuron-specific and glia-specific datasets, and the input datasets were not taken into consideration. The 0.01 p-value cutoff resulted in 2271 genes being called neuron specific, and 1744 genes to be glia specific (figure 3.7C and D). To check the overall level of expression of these two gene classes, I plotted the RPKM (reads per kb per million mapped reads, produced by Pawel Bednarz) of each dataset as a boxplot (figure 3.8). This shows the distribution of RPKM between the two cell types, across the neuron-specific and glia-specific genes. Note that the y-axis is different for each plot. As can be seen with these plots, the neuron-specific genes generally have a higher expression level than the glia-specific genes (mean RPKM: 6.55 for neurons and 5.45 for glia).

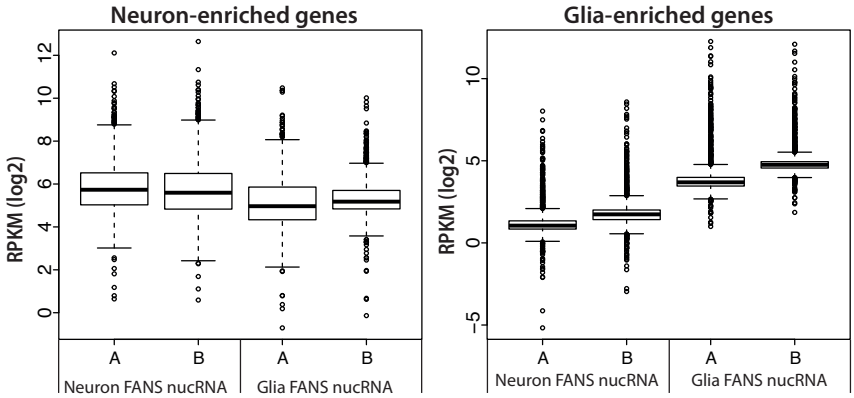


Figure 3.8: Normalised expression at cell-type-specific genes - boxplots of RPKMs (log2) at neuron- and glia-enriched genes (calculated using DESeq2) Neuron-enriched genes have generally higher expression level than the glia-enriched genes. A and B indicate the biological replicates for the nucRNA-sequencing for each cell type.

3.4 Assessing the nucRNA gene calls

Before analysing the chromatin state of neuronal and glial-specific genes, I first needed to assess the genes identified as cell-type-specific from the nucRNA data. I assessed the gene sets computationally, using publicly available resources to determine if the genes identified as neuronal-enriched or glia-enriched were consistent with what is known about the gene functions (gene ontology, GO) and expression pattern (FlyAtlas).

3.4.1 Characteristics of the cell-type-specific genes

To assess if the expected types of genes were enriched from DESeq2 analysis of the two cell types, I used the online *Drosophila* genomic database FlyMine (see http://www.flymine.org/). FlyMine performs GO-term enrichment analysis on the gene lists, and the results of these analyses can be seen in tables 3.2 and 3.3. The neuron-specific gene set contained many GO terms that are indeed expected for neurons. Signalling and cell communication are the most highly enriched, and many other GO terms such as neurogenesis, axon-guidance, behaviour, neurotransmitter secretion, and cognition are very significantly enriched in the neuronal gene set. Many more significant GO-terms were found for the neuronal genes than shown in table 3.2, but the majority of overlapping terms were removed and I selected those terms with the highest p-value that could be displayed in a single table. The glia-specific gene set has far less enriched GO terms than the neuronal specific gene set (table 3.3). This may be due in part to less being known about glia cell function than neurons, and thus there are less GO terms that would match the glia gene-set significantly.

An alternative analysis that is also computed through FlyMine is an overview of how each gene set is represented in the FlyAtlas database (148). The FlyAtlas database was generated by dissecting tissues from adult flies or larvae, isolating the mRNA and measuring gene expression levels using a microarray that covers the majority of genes in the *Drosophila* genome. In this dataset, all genes were identified as having higher expression (up) or lower expression (down) relative to the average of the whole fly. The neuronal-enriched genes are in agreement with what is observed in the FlyAtlas (figure 3.9, neuron-enriched). The neuronal-enriched genes are found up-regulated in the FlyAtlas tissues that are composed of neuronal cell types; the head, brain, eye, larval CNS, and thoracoabdominal ganglion. The glia-enriched genes are found to be down-regulated in most tissues, with the exception of testis (figure 3.9, glia-enriched). The lack of up-regulation of glia genes in the head and brain makes sense in that these tissues are largely made up of neurons, with glia contributing to only a few percent of

Table 3.2: Neuron-enriched genes GO analysis

Biological process	***p*-value**
Cell communication	2.40×10^{-47}
Biological regulation	2.45×10^{-47}
Signalling	2.97×10^{-47}
Response to stimulus	5.92×10^{-44}
Neuron differentiation	5.94×10^{-32}
Neuron development	9.36×10^{-30}
Synaptic transmission	8.47×10^{-27}
Nervous system development	1.88×10^{-20}
Locomotion	2.29×10^{-20}
Axonogenesis	2.47×10^{-19}
Localisation	9.54×10^{-19}
Taxis	2.52×10^{-18}
Mitotic spindle elongation	2.33×10^{-18}
Intracellular signal transduction	2.86×10^{-17}
Synapse organisation	2.37×10^{-16}
Single-organism behaviour	3.56×10^{-16}
Behaviour	3.18×10^{-15}
Neurogenesis	9.71×10^{-15}
Axon guidance	2.60×10^{-14}
Photoreceptor cell development	4.63×10^{-14}
Response to chemical	5.42×10^{-14}
Chemotaxis	5.47×10^{-14}
Cognition	6.01×10^{-14}
Neuron projection guidance	7.60×10^{-14}
Transport	1.37×10^{-13}
Cytoskeleton organisation	1.47×10^{-13}
Synapse assembly	2.42×10^{-13}
Growth	2.28×10^{-12}
Regulation of neurotransmitter levels	2.39×10^{-12}
Eye development	1.15×10^{-11}
Neurotransmitter secretion	7.78×10^{-11}
Neuron recognition	1.87×10^{-10}
Phosphorus metabolic process	8.23×10^{-10}
Memory	3.95×10^{-09}
Vesicle localisation	6.16×10^{-09}
Secretion	2.04×10^{-08}
Metamorphosis	9.05×10^{-08}
Cell adhesion	1.28×10^{-07}
Learning	1.71×10^{-07}
Circadian rhythm	1.82×10^{-07}

Table 3.3: Glial-enriched genes GO analysis

Biological process	*p*-value
Proteolysis	3.70×10^{-09}
Detection of chemical stimulus	1.32×10^{-08}
Transmembrane transport	7.51×10^{-08}
Sensory perception	1.81×10^{-06}
Chitin metabolic process	2.59×10^{-05}
Glucosamine-containing compound metabolic process	1.49×10^{-04}
Amino sugar metabolic process	1.83×10^{-04}
Aminoglycan metabolic process	2.36×10^{-04}
Sensory perception of taste	3.75×10^{-03}
Body morphogenesis	5.62×10^{-03}
Chitin-based cuticle development	1.23×10^{-02}
Neurological system process	3.42×10^{-02}

the cell numbers (figure 3.2, B: Repo::H2B-GFP). Why glia-enriched genes are enriched in the testis is unclear. One example from mouse found a homeo-box gene Gtx that is expressed in glia and testis (149), which would regulate many downstream target genes that would also be the same. Because glia are such a minority cell population in the *Drosophila* brain, very little is known about what genes are expressed in them. This exemplifies how the cell-type-specific analysis performed in this research can lead to a better understanding glia cell biology.

3.4.2 Comparing tools for identifying cell-type-specific genes

The nucRNA dataset for neurons and glia is distinct, yet complementary to the CAST-ChIP Pol II binding and TRAP-analysis obtained previously in the lab. All three of these tools measure gene activity in neurons and glia, using the same cell-type-specific promoters (elav or repo) to label the cell type of interest. Therefore, it would be expected that many of the genes identified as cell-type-specific by one method would also be identified in the other methods. Genes that are identified by only a single method are also expected, since each method measures a distinct step in the gene expression process and genes may be regulated at different points of gene expression. I wanted to determine how well the three methods for cell-type-specific gene identification agreed with each other. Because the data types are not easily comparable (i.e. ChIP-seq versus RNA-seq), I instead focused on comparing those genes that were called as specific within each method, then compared the gene lists to each other.

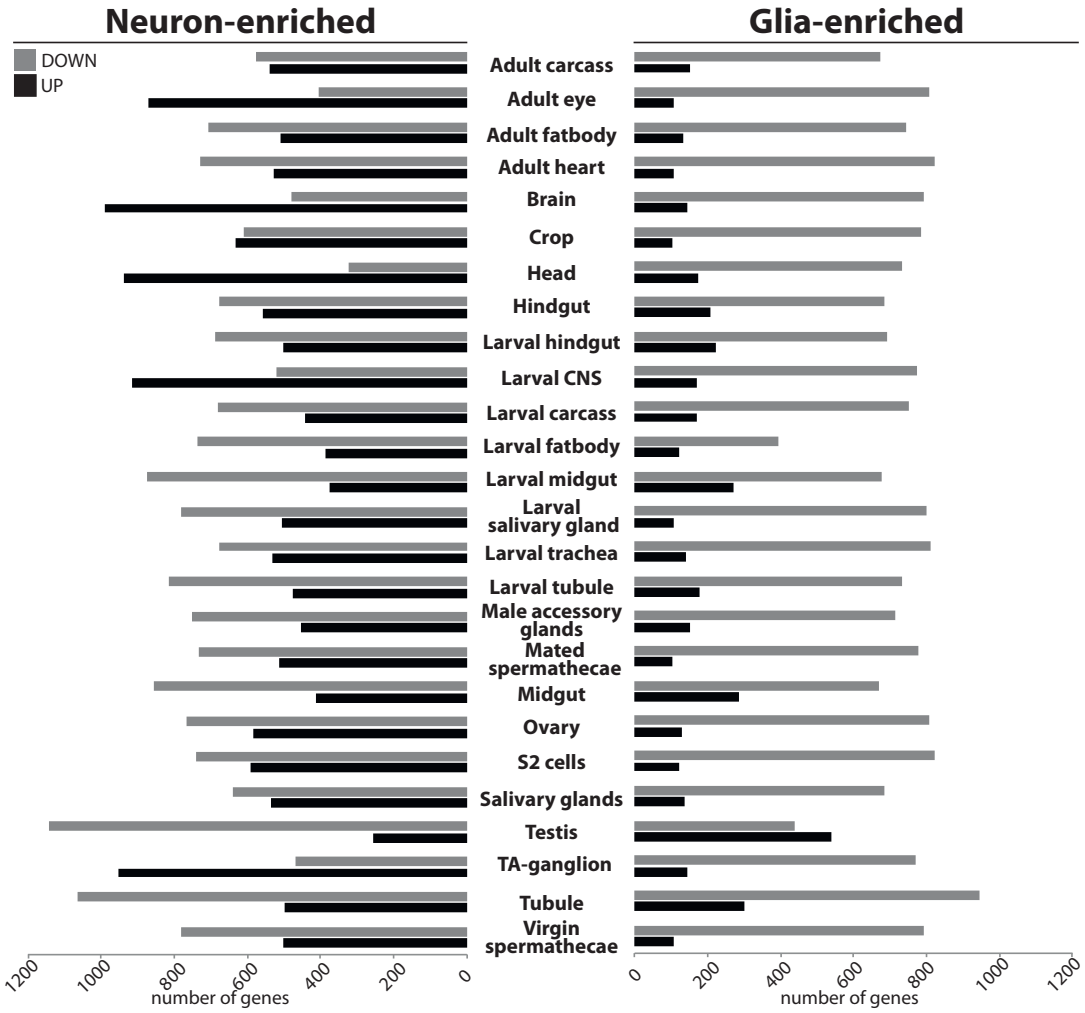


Figure 3.9: FlyAtlas analysis of neuron- and glia-enriched genes - FlyAtlas-shows that the neuron-enriched genes are enriched for those known to be expressed in the head, brain and CNS. The glia-enriched genes are generally not over-expressed in any tissue, with the exception of the testis. TA-ganglion=Thoraco-abdominal ganglion

3.4.3 Different cell-type-specific techniques call different gene sets

To determine how different these techniques are in terms of the genes that were called cell-type-specific, I overlapped the gene sets using a weighted venn package, Vennerable, in R (figure 3.10). There is surprisingly little overlap between the groups, particularly for the glia-specific genes (figure 3.10), where only 60 genes intersect between the nucRNA and CAST-ChIP gene lists. There is more overlap between the neuronal gene sets, with 117 genes intersecting between the nucRNA and CAST-ChIP gene calls. There is much more overlap between the TRAP and nucRNA gene sets for each cell type, which may be because both methods measure RNA and are more experimentally comparable. I also plotted the gene expression level in the nucRNA of those genes identified as cell-type-specific in the CAST-CHIP or TRAP analysis, with a control comparison to the nucRNA gene lists (figure 3.11). Since one possibility for the lack of overlap between the nucRNA and Pol II binding is if the gene expression is very low at the Pol II-identified genes, then the genes would not be called as statistically significant between the cell types. This appears not to be the case, as the level of nucRNA expression at the CAST-ChIP and TRAP genes is as high as the average level of the nucRNA-identified genes (figure 3.11, A).

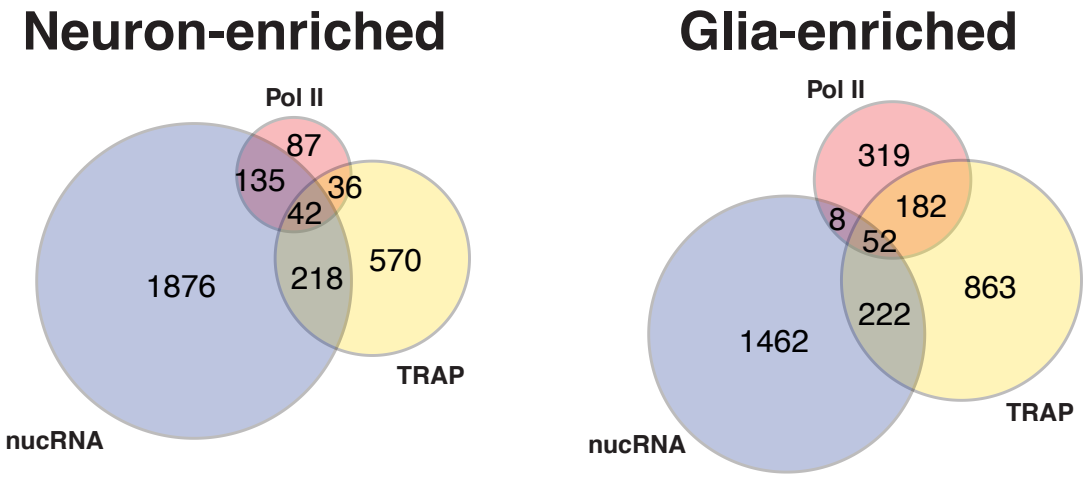


Figure 3.10: Cell-type-specific genes defined by nucRNA are not reflected by Pol II binding or TRAP - Weighted Venn diagrams showing the overlap of cell-type-specific genes called in each method. There is more overlap between the three cell-type-specific methods for the neuron-enriched genes than the glia-enriched genes.

When looking at the scatterplots of the CAST-ChIP and TRAP genes at the nucRNA expression levels (figure,3.11,B and C), it is clear that these genes are in the central region of the nucRNAs expression range. It can also be observed from the scatterplots that some of the genes that are specific for one cell type in the CAST-ChIP or TRAP analysis actually have higher nucRNA expression in the other cell type. A comparison of the level of Pol II binding at the transcription start site (TSS) and the nucRNA RPKM indicates that there is a positive correlation between the level of Pol II binding and expression (appendix A.1). Thus I conclude that it is not a complete disagreement within the total data that is contributing to the differences in gene calls, but likely the subtle differences between the methods in how genes are called as cell-type-enriched that is contributing to the major differences.

There are several technical reasons why the genes called with each technique are so different. First, there are vast differences in the dynamic range between ChIP-seq data and RNA-seq data. The ChIP-seq data typically has a range of 0–1000 reads, whereas RNA-seq can have a range of 0–10 000 reads, or more. Background binding for ChIP-seq experiments also means that peaks need to reach coverage above a threshold before they are considered as true peaks, which also reduces the dynamic range. First a region needs to be identified as a peak, and then the level of reads across this region needs to be significantly different between the cell types. Therefore peaks that have low levels of transcription may fall below detectable levels in the ChIP-seq analysis, but can be measurable in the RNA-seq analysis.

A biological reason for the disparity between gene groups is that these three techniques measure different stages of gene expression. The CAST-ChIP method measures the binding of Pol II, but cannot determine whether the bound Pol II is actively transcribing or not. The nucRNA measures processed and un-processed transcripts that were within the nucleus at the time of isolation. NucRNA analysis therefore provides better information about what is transcribed, yet lacks information about the level of mature mRNAs produced from each gene. TRAP measures the abundance of mRNAs at the final stage of gene expression. Some mRNAs identified in TRAP may be long-lived, and transcription no longer occurring. Some genes may also have the same transcription level in all tissue-types, but the stability of the mRNA is different between the cell types. This would lead to a gene being identified as specific in TRAP, but not in the nucRNA. Each step of regulation could differ in one cell type compared to another cell type. The challenge will be to tease apart what is true biological regulation from the noise and error each technique introduces.

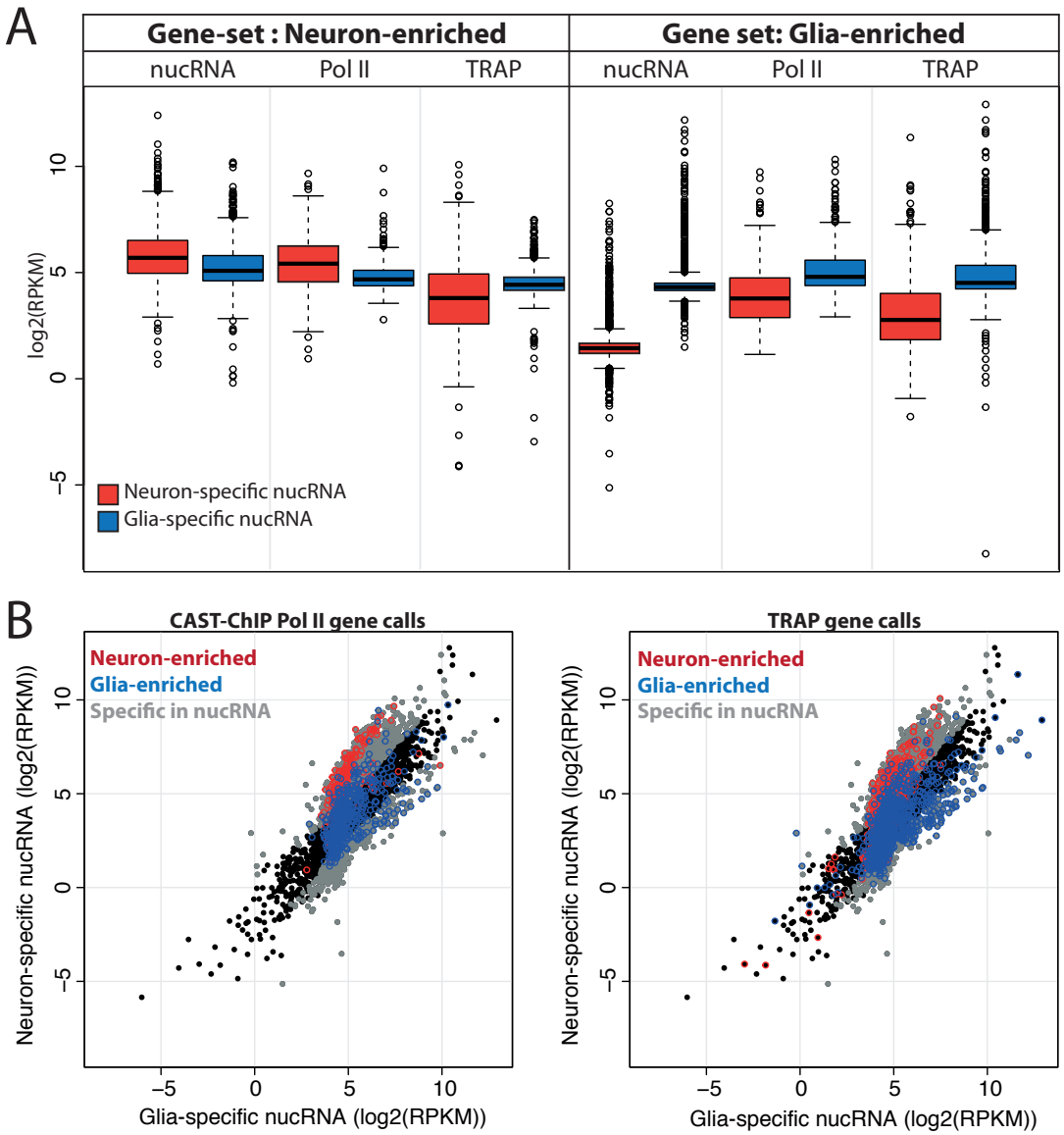


Figure 3.11: Comparison of different cell-type-specific gene calls - Comparing the level of expression in the nucRNA dataset in the three cell-type-specific gene classifications. A) Boxplots showing the RPKMs of the nucRNA datasets of those genes identified as cell-type-specific by either nucRNA, Pol II binding, or TRAP analysis. B) Scatterplots showing the nucRNA expression levels of cell-type-specific genes identified by either CAST-ChIP Pol II binding, or TRAP. The genes identified as cell-type-specific for CAST-ChIP-Pol II and TRAP gene calls are highlighted in red for neuron-enriched and blue for glia-enriched, gene calls identified as cell-type-specific from the nucRNA data analysis are highlighted in grey.

3.4.4 Characteristics of genes identified from a single method

Do the genes that are identified as neuronal-enriched or glia-enriched in only one method have specific characteristics? For instance, if the genes identified as cell-type-specific in the TRAP analysis only are indeed regulated at the translational level, then a specific type of gene may be regulated in this way. Thus by comparing and contrasting the three cell-type-specific techniques, which type of gene is regulated at what level of gene expression may be identified. Using the statistical program R, I generated lists of genes corresponding to the non-overlapping regions of the Venn digram in figure 3.10. I then entered these gene lists into the Flymine database, to identify enriched GO-terms for each gene set. I manually refined each GO-analysis to remove redundant and non-sensical GO-terms, for example both signalling and single organism signalling were present in the nucRNA-only neuronal GO analysis but I selected only signalling. The results of the refined GO analysis are presented in tables 3.4 and 3.5 for neuron-enriched genes, and table 3.6 for glia-enriched genes.

Comparing the results for the Neuronal-enriched genes hows there is a clear difference in the enriched GO-terms between the three methods. Some GO-terms are shared between the genes called by two methods, such as signal transduction found in nucRNA-only and Pol II-only or G-protein coupled receptor signalling pathway between TRAP and CAST-ChIP. However, many of the GO-terms are unique to each method. GO-terms for behaviour, learning, memory, and cognition are uniquely enriched in the nucRNA-only gene set (table 3.4). GO-terms enriched in generating neurons and their connections, e.g. neurogenesis, axonogenesis, synapse assembly and axon guidance, are uniquely enriched in the nucRNA-only gene set. Cell communication and signalling are the main enrichment for CAST-ChIP (Pol II)-only genes. There is an unexpected enrichment in the TRAP-only gene set for cilium assembly and organisation. TRAP-only genes are also enriched for functions in transcriptional regulation, although the p-value for these is much lower. Within the transcriptional-regulation GO term, the most common protein domain is homeobox domain (43 % of cases); the next most commonly enriched protein domain is zinc-finger (13 % of cases).

Table 3.4: GO analysis of nucRNA-only neuron-enriched genes

Biological process	p-value
Signalling	7.65×10^{-38}
Response to stimulus	3.38×10^{-36}
Synaptic transmission	4.17×10^{-25}
Neuron differentiation	9.02×10^{-25}
Generation of neurons	5.61×10^{-24}
Localisation	3.47×10^{-20}
Spindle elongation	5.95×10^{-17}
Locomotion	5.55×10^{-16}
Transport	1.09×10^{-15}
Cytoskeleton organisation	1.29×10^{-15}
Synapse organisation	2.59×10^{-15}
Axonogenesis	5.99×10^{-15}
Growth	5.51×10^{-14}
Taxis	9.13×10^{-13}
Behaviour	9.63×10^{-13}
Neurogenesis	4.81×10^{-12}
Neurotransmitter secretion	6.25×10^{-12}
Synapse assembly	8.16×10^{-12}
Axon guidance	3.78×10^{-11}
Metamorphosis	5.15×10^{-11}
Cell differentiation	3.49×10^{-09}
Regulation of synapse assembly	5.21×10^{-09}
Neurotransmitter transport	1.04×10^{-08}
Learning or memory	4.67×10^{-08}
Cognition	4.67×10^{-08}
Secretion	4.92×10^{-08}
Phosphorylation	7.59×10^{-08}
Endocytosis	3.26×10^{-07}
Spindle organisation	1.17×10^{-06}
Memory	1.62×10^{-06}
Centrosome duplication	2.42×10^{-06}
Long-term memory	1.66×10^{-06}
Mitotic cell cycle	2.39×10^{-05}
Cell adhesion	3.26×10^{-05}
Regulation of gene expression	4.81×10^{-05}

Table 3.5: GO analysis of neuron-enriched Pol II only and TRAP only genes

CAST-ChIP-only neuron-enriched genes	
Biological process	*p*-value
G-protein-coupled receptor signalling pathway	1.28x 10^{-16}
Cell communication	1.20 x 10^{-12}
Neuropeptide signalling pathway	4.23 x 10^{-08}
Signal transduction	1.08x 10^{-07}
Synaptic transmission	3.18 x 10^{-05}
Neurotransmitter transport	1.6 x 10^{-02}

TRAP-only neuron-enriched genes	
Biological process	*p*-value
Cilium assembly	6.27 x 10^{-12}
Cilium organisation	9.27 x 10^{-12}
Sensory perception of taste	2.77 x 10^{-10}
Neuropeptide signalling pathway	2.02 x 10^{-09}
G-protein-coupled receptor signalling pathway	2.93 x 10^{-06}
Detection of stimulus	0.011
Regulation of transcription, DNA-templated	0.015
Regulation of RNA biosynthetic process	0.015
Transcription, DNA-templated	0.030
RNA biosynthetic process	0.033

The GO-analysis comparisons for the glia-enriched genes are shown in table 3.6. As with the neuronal-enriched genes, the GO-terms are unique for each method class. Those genes identified only by the nucRNA analysis are enriched for functions in proteolysis and sensing perception (chemical stimulus or taste). In contrast to the neuron-enriched genes, where regulation of transcription is enriched in the TRAP-only class, this GO term is enriched at the Pol II only class. Genes involved in metamorphosis are also significantly enriched for the Pol II-only gene class. Those genes in the TRAP-only class are enriched for metabolic functions, and female mating behaviour and receptivity.

Table 3.6: GO analysis of glia-enriched genes identified by a single method

nucRNA-only glia-enriched genes	
Biological process	p-value
Proteolysis	3.37x 10^{-11}
Sensory perception of chemical stimulus	2.24 x 10^{-09}
Detection of stimulus	2.90 x 10^{-06}
Sensory perception of taste	2.60 x 10^{-04}
Body morphogenesis	0.013
Microtubule-based movement	0.044

CAST-ChIP-only glia-enriched genes	
Biological process	***p*-value**
Post-embryonic development	2.50 x 10^{-05}
Metamorphosis	2.68 x 10^{-04}
Negative regulation of transcription from RNA polymerase II promoter	2.23 x 10^{-03}
Regulation of transcription from RNA polymerase II promoter	3.61 x 10^{-03}
Small molecule metabolic process	8.02 x 10^{-03}

TRAP-only glia-enriched genes	
Biological process	***p*-value**
Glucosamine-containing compound metabolic process	2.11 x 10^{-09}
Chitin metabolic process	1.31 x 10^{-08}
Cuticle development	2.13 x 10^{-08}
Septate junction assembly	6.93 x 10^{-02}
Regulation of female receptivity	6.93 x 10^{-02}
Toll signalling pathway	0.022
Negative regulation of endopeptidase activity	0.047
Female mating behaviour	0.048

3.4.5 Characteristics of genes identified with all methods

The genes identified by all three methods are likely the most extreme examples of cell-type-specific gene expression; they have the most robust differences in transcript and Pol II levels (figure 3.12. These are the genes which would fit into the expected model, of differential Pol II recruitment leading to differential mRNA expression, and finally to differential protein levels in the two cell types. However, there are very few of these genes, with only 42 neuronal and 52 glial genes (figure 3.10). As expected, these genes are highly enriched for expression in head and nervous-system tissues in the FlyAtlas datasets (not shown). The neuron-enriched genes have no significant GO enrichment, likely due to the low number of genes, however they have significant enrichment of the immunoglobulin-like fold protein domain (p-value 1.8 x 10^{-08}). This protein domain is found in genes such as the *beaten-path* family of cell-adhesion molecules, which are involved in axon guidance (150). The glia-enriched genes are significantly enriched for two GO-terms: localisation and transmembrane transport (p-values 7.95 x 10^{-07} and 8.0 x 10^{-06}, respectively).

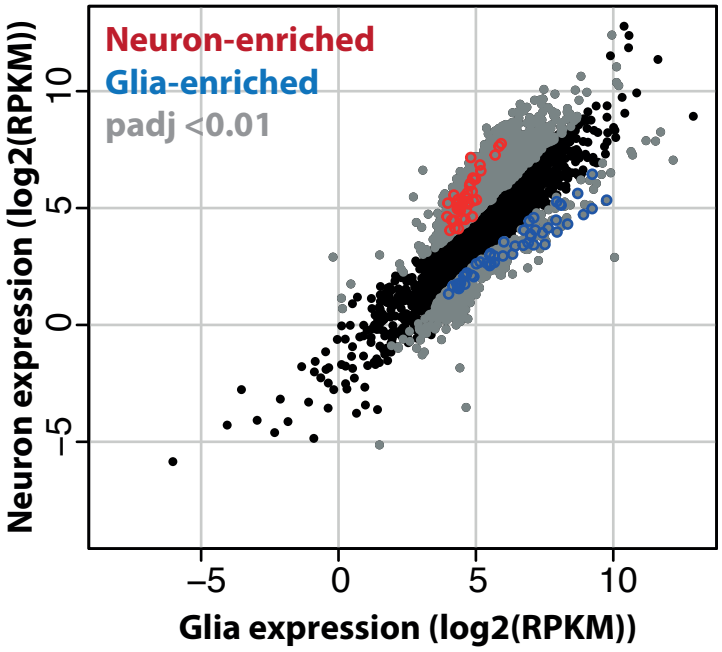


Figure 3.12: Genes called in all three methods - Scatterplot showing the expression level in the nucRNA dataset of those genes identified as specific in all methods.Those genes identified in all methods are not necessarily those with the highest expression, or largest difference in expression.

3.5 Discussion and future directions

This chapter highlights the challenges that come from analysing the chromatin state and gene expression in specific cell types. It is necessary to push the limits of the current methods for assaying chromatin structure and gene expression, so that we can reach an ever finer resolution. Here, I have established methods for RNA-seq, MNase-seq, and ChIP-seq from a small number of nuclei that made it possible to assay small populations of cells within a complex organism. While the limits of these techniques have by no means been reached, the advantage of my methods is that no additional amplification processes are required, apart from the usual sequencing library preparation. However, many amplification methods are being developed and once it is established that they introduce no bias to certain DNA/RNA signals, then the size of the cell population being studied may become extremely small. It may be possible to assay a single neuron per brain hemisphere.

The finding that there are loci with single nucleosome differences between the neurons and glia is striking (section 3.3.4). The depth of sequencing reads (approximately 500 million reads per sample) has no doubt contributed to observing such a phenomenon. Whether these single nucleosome differences are true or some artefact of the MNase-seq assay is yet to be established. Some alternative methods are possible: Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) (151) and the commonly used DNAse-I foot-printing assay (152). These two methods are used to identify open chromatin regions, which is the opposite of the MNase-protection assay, and would thus be a good complement to the MNase-seq data. Indeed I have also established the FAIRE-seq assay in the lab, and it is now possible to perform this assay on FANS-isolated nuclei. A more recent technology, named ATAC-seq (assay for transposase-accessable chromatin sequencing) (153), is also a good alternative method to develop and would require far less starting material, thus would be applicable to smaller cell populations. Once it has been established that these nucleosome-free regions are real, then dertermining the functional relevance of such regions would be highly interesting. Identifying underlying transcription factor motifs would be the first step. However, these regions could have many different functions, so separating them into distinct regions such as within intron-exon boundaries, or within the first intron, would be useful in determining the functional role of these regions.

The direct comparison of different cell-type-specific methods is extremely difficult. There are different efficiencies of enrichment, and multiple variables for each different method, such that even comparison between two cell types is not without difficulty. Thus, many further experiments are required before we can conclude from these comparisons that the methods indeed identify gene sets that are regulated at specific stages of gene expression. The conclusions drawn from these analyses underline the importance of keeping the shortcomings of an approach in mind when interpreting the data. Although there may be some biological significance to the findings here, this only aids in developing directed hypotheses, and more directed experimental work is needed to follow up and test these findings.

A major technical reason why the different methods have such disparate findings is the method of isolation for each technique. Both CAST-ChIP and TRAP use immunoprecipitation to enrich for the cell-type-specific material, whereas the nucRNA dataset comes from FANS-isolated nuclei. Immunoprecipitation can result in higher background noise by non-specific antibody binding and non-specific binding to the beads. This reduces the dynamic range of the technique and could hide real differences in tissue-specific expression if that expression is normally low.

The nucRNA dataset may suffer from different types of background, firstly genomic DNA contamination, and secondly non-specific RNAs from lysed cells binding to the outer membrane of the nuclei. Although the RNA isolated from nuclei was treated with DNAse I (methods 6.3.8), there is still the possibility that some of the reads in these samples are derived from DNA and not RNA transcripts. However, the extent of any contamination can be seen by mapping the RNA reads to the genome, without any annotation guidance. If there is genomic DNA contamination, then there will be a low level of reads evenly spread across the genome. An example of this type of contamination can be seen for one sample of the nucRNA (appendix A.2), where the glia-specific samples have a very slight coverage over the entire genome, indicating that there is genomic DNA contamination. However, the level of contamination is extremely low, and would contribute very little to the reads compared to the RNA.

Testing the kinetics of the differences in Pol II recruitment, elongation, or translational control would require more quantitative approaches to measure the gene-expression pattern at each stage. For example, using methods for measuring transcriptional kinetics, such as a pulse-chase experiment with radioactive nucleotides, would

tease out the differences of expression at the level of elongation. These types of methods could be used not only to identify nascent transcription, but also as a measure of the lifetime of mRNAs. Such experiments would be useful in determining whether the results from comparisons between Pol II binding, nucRNA, and TRAP make biological sense. Testing for kinetic differences in Pol II recruitment will be a difficult challenge, as ChIP-seq has far less dynamic range and is less quantitative than the RNA-seq techniques. Using an alternative method such as NET-seq (154), which uses the very last incorporated nucleotide as a proxy for where Pol II was bound, could be a good, quantitative alternative to ChIP-seq of Pol II.

One important point in these analyses is that the CAST-ChIP dataset was extensively validated experimentally. The genes identified as neuron- and glia-enriched were validated using enhancer-TRAP lines where the insertion site was near the identified cell-type-specific Pol II peak. Using these lines to drive H2B-GFP followed by immunohistochemistry of dissected brains, confirmed that the CAST-ChIP analysis identified genes with the correct cell-type-specificity. The TRAP and nucRNA datasets have not been validated experimentally in this way so far. Experimental validation of the cell-type-specific findings will be a necessary step, to ensure that the genomic data have not produced artefacts. Immunohistochemistry, or RNA FISH on dissected *Drosophila* brains would be the ideal method for testing cell-type-specific expression. RNA-FISH against target genes found only in the TRAP dataset would also be extremely interesting to see if there is different sub-cellular localisation of ribosome-bound RNAs in the neurons, for example. ChIP-qPCR, and RT-qPCR of several target genes would also be useful for assessing if the sequencing data reveals the true levels, and is not an amplification artefact. Using an alternative nuclei isolation method, such as INTACT, for both Pol II ChIP-seq and RNA seq would be a good way to determine how much the methodology affects the genes identified. For example, does Pol II ChIP against endogenous Pol II from isolated nuclei show the same binding patterns as the CAST-CHIP? This experimental validation will be a necessary next step to progress this analysis towards a manuscript that would be highly informative and useful for any researcher performing cell-type-specific genomics experiments. However, the computational analysis performed thus far has revealed highly interesting differences in how different gene groups may be regulated differentially to achieve the required cell-type-specificity.

Chapter 4

Mechanisms of cell-type-specific gene regulation in the *Drosophila* head

4.1 Summary

In this chapter, I describe two major mechanisms that govern cell-type-specific gene expression. One mechanism is regulated through the specific recruitment of RNA polymerase (Pol II) to the gene promoters, and the other mechanism is hypothesised to be regulated by cell-type-specific modulation of transcriptional elongation.

The genes regulated by Pol II recruitment show differential nucleosome occupancy in the promoter regions, and an unusual chromatin signature. An interesting feature of genes that are regulated through specific Pol II recruitment is lack of the "active" histone modification H3K36me3, despite high level of transcription. These active genes, with no H3K36me3, have H3K27ac broadly spread across the gene body. Neuron-enriched genes that are regulated at the level of Pol II recruitment, are predicted to be bound by the insulator protein suppressor of hairy wing Su(Hw) in non-neuronal cell types. Thus, binding of Su(Hw) outside of neurons is likely a mechanism to achieve repression of a subset of neuronal genes outside of neurons.

Genes regulated at the level of transcriptional elongation present highly similar promoter architecture, chromatin state and level of Pol II recruitment between different cell types, yet have higher levels of transcription in one cell type. How the higher expression in one cell type is achieved is unclear, but likely depends on distal regulatory elements, rather than sequences within the promoter.

4.2 Introduction

Chromatin networks of histone modifications, histone variants and multitudes of interacting proteins deliver the precise spatial and temporal patterns that control gene expression to establish a complex organism. However, it is now evident that even in the brain, where the majority of cells are post-mitotic and their fates have long been established, the chromatin landscape appears to maintain a high level of plasticity. Without this plasticity, no molecular adaptation to changing environments would be possible. The chromatin structure within the brain, which is assumed to be established at the final stages of differentiation, is not as tightly set as previously imagined.

Fly learning behaviour requires modulation of chromatin structure. For example, geneome-wide analysis of transcription in brain tissues of mated and virgin female *Drosophila* revealed an up-regulation of genes encoding chromatin remodellers 24 hours after mating takes place (155). However, the available experimental evidence from such studies only give limited insight into cell-type-specific functions. The key question is whether the majority of genes within these post-mitotic cells are confined within a stable chromatin structure, or can they have altered expression and be transformed into a new chromatin state in specific cells?

Measuring gene activity, and identifying how activity differs between two closely related cell types, is not a simple task. Recruitment of Pol II to a gene's promoter does not necessarily reflect the transcriptional output from that gene. It is also not the case that genes with high levels of expression would have high amounts of detectable Pol II binding. Dynamic interactions between chromatin and the transcriptional machinery influence and modulate every step of gene expression. One example of this is promoter proximal pausing, characterised by Pol II pausing 30–50 nucleotides into the gene. Pausing acts as a regulatory check-point before Pol II progresses to transcriptional elongation, with suggested functions to anticipate future gene expression requirements, and modulate the transcriptional output (154, 156). Gene regulatory mechanisms, such as Pol II pausing, occur in the context of a highly complex chromatin structure. The nucleosome plays a repressive role in transcriptional elongation, as well as governs the access of regulatory proteins to binding sites. Modifying nucleosomes, such as acetylation or methylation of the histone tails, contributes to a complex "code" that modulates gene activity (157). However, the term "histone code" presents a vastly oversimplified function of histone modifications, and the contribution histone modifications make to gene regulation is far from clear. A prime example of this complexity is described in

this chapter; the histone modification H3K36me3 is associated with active genes, but is not correlated with gene expression level and some active genes do not have this mark at all (76).

In the previous chapter, I demonstrated that the different techniques for measuring cell-type-specific gene activity, RPB3-CAST-ChIP and FANS-RNA-seq, resulted in vastly different gene groups being identified (section 3.4.3). The question arises as to whether these genes are regulated by different mechanisms. In this chapter, I focus on the comparison of chromatin states between those genes identified using the RPB3-CAST-CHIP method (Pol II binding) and those identified from the analysis of nucRNAs (nascent transcripts). I chose to investigate these two techniques as they are expected to more closely represent the actual state of gene activity at the level of chromatin, whereas the TRAP is expected to be far removed from what is happening at the chromatin level.

Aims

- Assess nucleosome architecture across cell-type-specific and invariant promoters.
- Compare H2AZ, H3K27ac, and H3K36me3 between cell-type-specific and invariant genes.
- Assess the characteristics of gene groups that possess different chromatin states.

4.3 Activity of cell-type-specific and invariant genes

In the majority of the results for this chapter I will be using metagene analysis (using NGS-plot, (158)) to look at the average profiles of genomic data over subsets of genes defined as neuron-enriched, glia-enriched, or invariant as defined by two different methods; CAST-ChIP or nucRNA analysis (section 3.4.3). Before analysing the chromatin state of the different cell-type-specific gene groups, I wanted to assess the nucRNA levels of the CAST-ChIP gene sets, and conversely the Pol II binding levels of the nucRNA gene sets. Thus, I performed metagene analysis of the CAST-ChIP and nucRNA data split into the different gene sets, which have two levels of categorisation (table 4.1). The first level is whether the genes are from the CAST-ChIP or the nucRNA analysis; the second level is whether the genes are classified as neuron-enriched, glia-enriched, or invariant. The specific questions addressed here are: what is the Pol II binding profile of those genes called as specific or invariant from the nucRNA-seq analysis, and conversely what is the nucRNA expression profile of genes identified as specific or invariant

by Pol II binding? This analysis will provide the basis for interpreting the chromatin profiles of the different gene groups.

Table 4.1: Gene groups used in these analyses

	Neuron-enriched	**Glia-enriched**	**Invariant**
Identified by Pol II	300	561	1880
Identified by nucRNA	2271	1744	4670

I first analysed the Pol II binding and nucRNA expression levels of those genes identified as neuron- or glia-enriched based on the CAST-ChIP Pol II analysis. I generated average profile plots (NGS plot) of the CAST-ChIP and nucRNA data in both neurons and glia, spliting the data by the different gene categories (figure 4.1). For the Pol II analysis, I chose to align the data to the TSS, rather than scaling the data across the entire gene, because the majority of detectable Pol II enrichment for the CAST-ChIP Pol II data is at the TSS, with very little enrichment over the gene body. The upper two panels of figure 4.1 show that the genes called as cell-type-specific by the CAST-ChIP method do indeed have cell-type-specific Pol II binding. For example, genes called as neuronal do indeed have Pol II enrichment in neurons, and little Pol II enrichment in glia. An important observation to note is the difference in height of the Pol II peaks in the different cell types. Neuronal genes in neurons have a much higher level of Pol II binding than glial genes do in glia. Comparing the peaks relative to invariant genes is useful, since by definition the Pol II binding is the same in both cell types. In this way it can be seen that neuron-enriched genes in neurons have a level of Pol II binding similar to that of invariant genes, and glia-enriched genes in glia have Pol II binding that is much lower than invariant genes.

Differences identified by Pol II binding correlate with differences in nucRNA expression

The lower panel of (figure 4.1) shows the nucRNA level of those genes identified as specific or invariant in the CAST-ChIP assay. In this analysis, the data are scaled across the entire gene length between the transcription start site (TSS) and the transcription end site (TES), so that the average nucRNA level across the entire gene body can be observed. As might be expected, the RNA levels reflect the Pol II binding profiles. In neurons, the neuron-enriched genes have an equivalent level of RNA expression as invariant genes, just as the Pol II peaks were equivalent in the top left panel of figure 4.1. However, the neuronal genes have significantly more $3'$ reads than the invariant

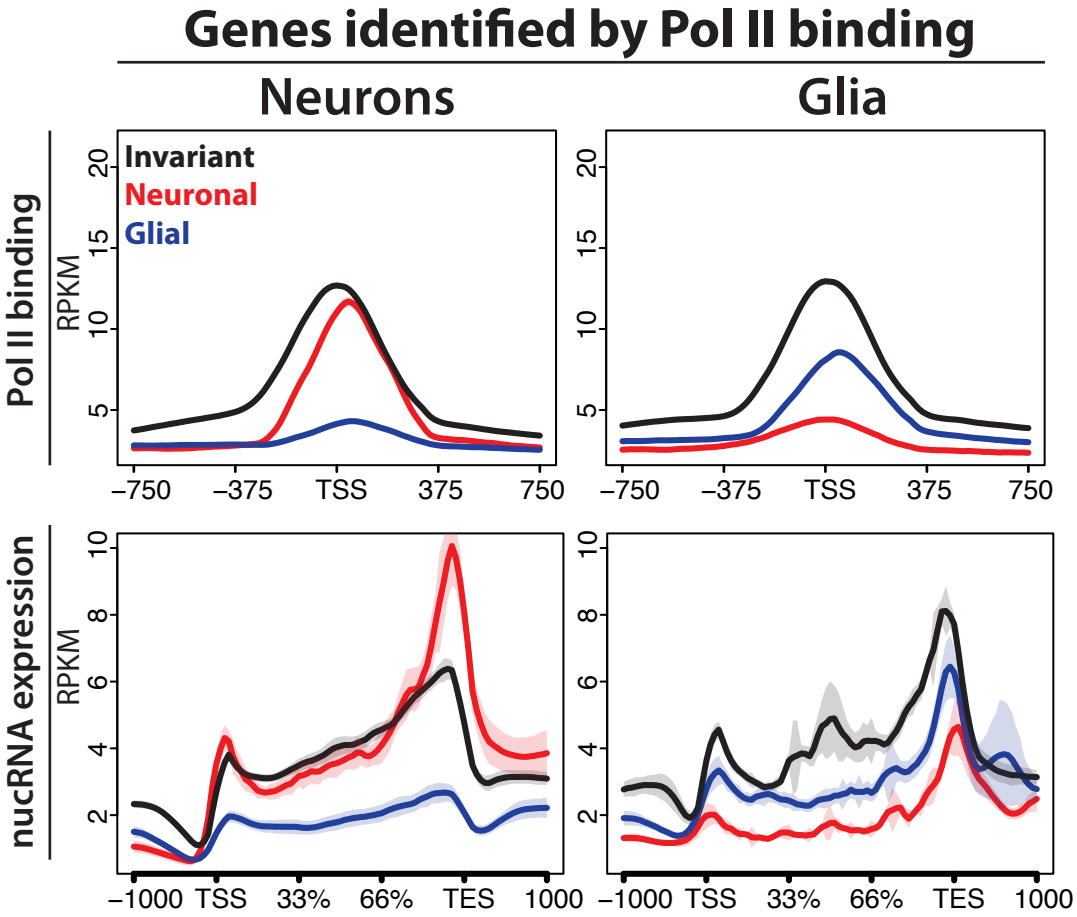


Figure 4.1: Average Pol II and nucRNA profiles of Pol II-identified genes - Metagene analysis of the CAST-ChIP RPB3 data (upper panel) and the nucRNA data (lower panel) from neurons and glia. Data was divided into three gene classes as identified from the Pol II binding analysis: invariant, neuron-enriched, or glia-enriched (black, red, and blue, respectively). RPKM stands for read-count per million mapped reads. For the RPB3 data, the data are centred around the transcription start site(TSS). For the nucRNA data, the data are scaled across the gene body.

genes. This may be an artefact of library preparation, since neuronal genes are generally much longer than invariant genes. The library preparation uses a mixture of polyA selection and random hexamers, thus longer genes may have less enrichment towards the 5′ end because the cDNA synthesis at the 3′ end is more efficient. Glia-enriched genes, as identified from the Pol II binding, have a low level of nucRNA expression in neurons, which is also comparable to the level of Pol II binding within this cell type. Gene expression within glia cells also reflects the Pol II binding profile, the expression level of glia-enriched genes within glia is lower than that of invariant genes.

The expression pattern of cell-type-specific genes identified by Pol II binding showed the expected trends. However it is easy to see how the majority of glia-enriched genes identified by Pol II binding were not called as specific from the nucRNA analysis (section 3.4.3). Genes with more Pol II in glia than in neurons (blue line, bottom panels, figure 4.1) have similar levels of nucRNA in both cell types, which is generally very low. This indicates that Pol II binding analysis identified genes that show the desired expression pattern within a cell type, i.e within glia cells, genes with glia-enriched Pol II binding have higher expression than genes with neuron-enriched Pol II. However the Pol II binding did not necessarily identify genes that have significantly different RNA expression levels between cell types. Pol II binding correlates to the RNA expression level to a certain extent. Genes with a low level of Pol II binding have lower expression, and genes with a high amount of Pol II binding have higher RNA expression. But this relationship is not a perfect correlation (section A.1), which could explain why the genes called as significant differences between the two techniques can be in disagreement (section 3.4.3).

Genes identified as different by nucRNA have invariant Pol II binding

I next analysed the Pol II binding and nucRNA expression levels of those gene identified by the nucRNA DESeq analysis (figure 4.2). The analysis was performed using the same parameters as those of figure 4.1. When observing the Pol II binding at the nucRNA-identified genes, it is clear that there is little difference in the Pol II profiles between the cell types (upper panels, figure 4.2). Here, the neuron-enriched genes have a high level of Pol II binding in both neurons and glia. Conversely, glia-enriched genes have no Pol II binding in neurons and only a very small amount more in glia cells. Thus, specificity of expression at the RNA level is not reflected in the Pol II binding profile of those genes.

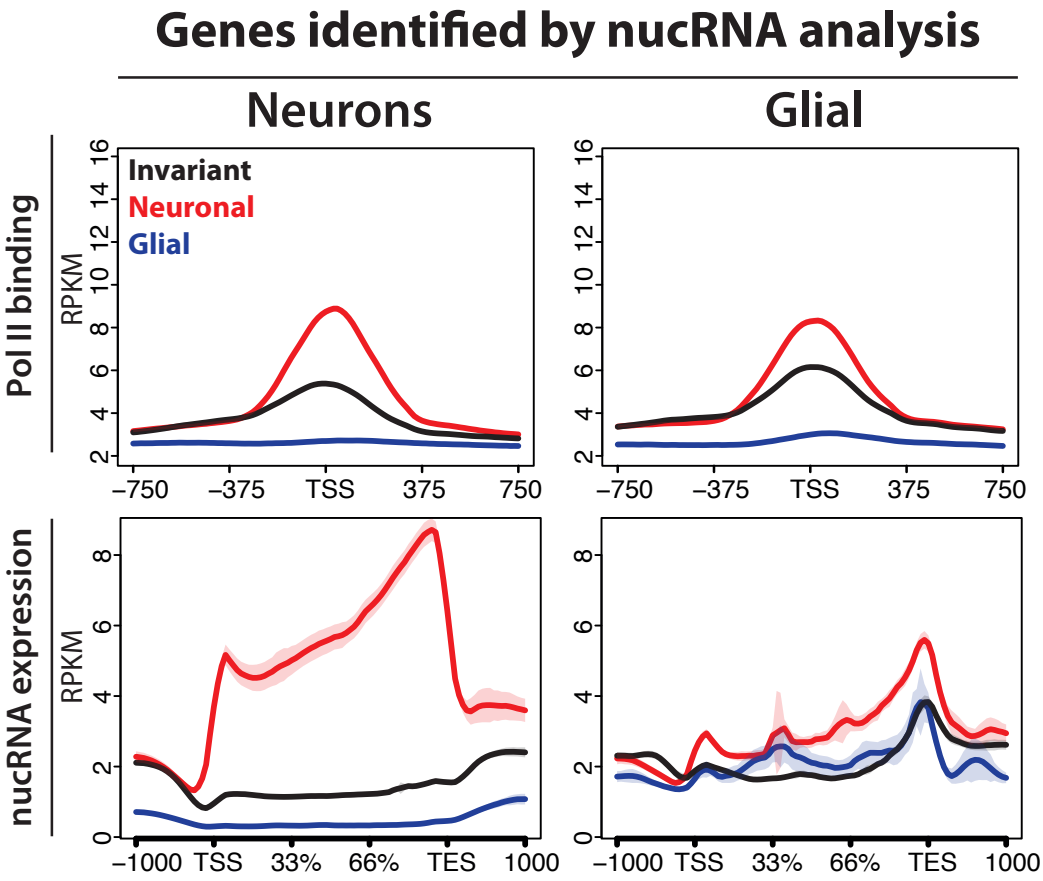


Figure 4.2: Genes identified by nucRNA have invariant Pol II binding - Meta-gene analysis of the CAST-ChIP RPB3 data (upper panel) and the nucRNA data (lower panel) from neurons and glia. Data was divided into three gene classes as identified from the nucRNA analysis: invariant, neuron-enriched, or glia-enriched (black, red, and blue, respectively). RPKM stands for read-count per million mapped reads. For the RPB3 data, the data are centred around the transcription start site(TSS). For the nucRNA data, the data are scaled across the gene body)

Visualising the average profile plots of the nucRNA expression levels for those gene identified by the nucRNA DESeq analysis reveals some intriguing results (bottom panels, figure 4.2). Here, the neuron-enriched genes have a very high level of expression in neurons and mid-range expression in glia. It is immediately clear that neuron-enriched gene expression within glia cells is higher than glia-enriched gene expression in glia cells (bottome right, figure 4.2). Indeed, glia-enriched gene expression is relatively low in glia, but in neurons glia genes have nucRNA levels that are extremely low, being nearly zero reads (blue line, figure 4.2). This is initially counterintuitive, as it might be natural to expect that the glia-enriched genes should have higher expression than neuron-enriched genes in glia cells. However, this pattern occurs because DESeq is comparing between cell types, and not gene expression of groups within the same cell type. Thus there is certainly significantly more expression of the neuron-enriched genes in neurons compared to glia, but this does not mean that there is no expression of these genes within glia.

Different cell-type-specific methods uncovered different mechanisms of gene regulation

The observation that the glia-enriched genes are very lowly expressed within glia cells, and have even lower expression in neurons is intriguing. Two possible explanations for this observation are: 1) that glia-specific genes are strongly repressed by default in all cell types and specifically de-repressed in glia cells, or 2) glial genes have a low basal expression level in all cell types and are specifically repressed in neurons. It is likely that both possibilities occur, and further analysis is necessary to determine the most prominent scenario. One method to differentiate between neuron-specific repression of the glia genes or glia-specific up-regulation would be to compare the cell-type-specific data to the average expression level of the whole fly, or other non-neuron related tissues. This comparison would provide an indication as to whether the glia-enriched gene expression is at the same level in glia as in other tissues, indicating neuron-specific repression. Conversely, if the glia-enriched genes are expressed in the other tissues to the same level as that found in neurons, then this would indicate glia-specific up-regulation.

Most of the genes defined as specific from Pol II binding are different from those genes defined as specific from nucRNA expression analysis (figure 3.10). Both methods identify genes that are cell-type-specific, and both ways of calling genes were consistent with current knowledge and public data sets (section 3.4). However, the experimental bias of each method, discussed in section 3.5, may have incidentally identified gene groups that are regulated by different mechanisms. Very generally, these analyses

indicate two methods of achieving proper spatial expression of a gene: 1) recruitment of Pol II or 2) modulating transcriptional elongation. For example, genes that have differing amounts of Pol II across the gene body between cell types, but no difference at the promoter, would not be identified from the Pol II analysis because the sensitivity of the CAST-ChIP assay is too low across the genic regions. Thus genes that have differences in the recruitment of Pol II to the promoter were selected by the CAST-ChIP assay. In contrast to this, the neuronal genes identified by the nucRNA analysis may have an equivalent level of Pol II binding in both neurons and glia, but far more RNA expression in neurons. This suggests that the average speed or amount of elongation at these genes is far higher in neurons than glia, but the recruitment of Pol II to the promoter is the same. Another possibility is that these genes have differential mRNA stability or export from the nucleus between the cell types. The chromatin features that facilitate cell-type-specific Pol II recruitment may be different from the chromatin features that govern cell-type-specific elongation. I therefore wanted to analyse the chromatin state at both types of gene set, since this may provide information about how the different types of regulation are achieved.

4.4 Nucleosome architecture at specific promoters

Is there a difference in chromatin architecture at promoters between the different cell-type-specific and invariant genes? Do genes that are regulated at the level of Pol II recruitment have a different nucleosome composition than genes regulated at the elongation level? To address these questions, I generated average profile plots using the neuron and glia MNase-seq datasets, splitting the data into neuron-enriched, glia-enriched and invariant genes that were called either from Pol II binding (CAST-ChIP) or nucRNA analysis (table 4.1). I expected to observe lower nucleosome occupancy at the nucleosome-free region (NFR) and higher nucleosome order in the cell type where the genes are active (ordered architecture, figure 4.3, A), and expected a covering-up of the NFR in the cell type where the genes are not active (fuzzy architecture, figure 4.3, A). However, this quality was not observed (figure 4.3).

When looking at the gene sets identified by Pol II binding, there was a weak correlation between Pol II binding and depletion of nucleosome at the nucleosome-free region (NFR) (top panels, figure 4.3, B). A high level of Pol II binding at cell-type-specific genes corresponded to only a small depletion of nucleosomes in the cell type where the Pol II was binding. This relationship was more defined for neuron-enriched genes than for glia-enriched genes. This again demonstrates a relationship between Pol II

binding and NFR formation, since the neuron-enriched genes in neurons have a much higher enrichment of Pol II binding compared to invariant than the glia-enriched genes in glia. However, high amounts of Pol II binding was not sufficient to completely clear the NFR of nucleosomes (on average), since the cell-type-specific genes were far more fuzzy than invariant genes. For example, neuron-enriched genes had comparable Pol II binding to invariant genes (compare to figure 4.1, top left), yet the specific genes had little NFR formation or ordered nucleosome arrays compared to the invariant genes.

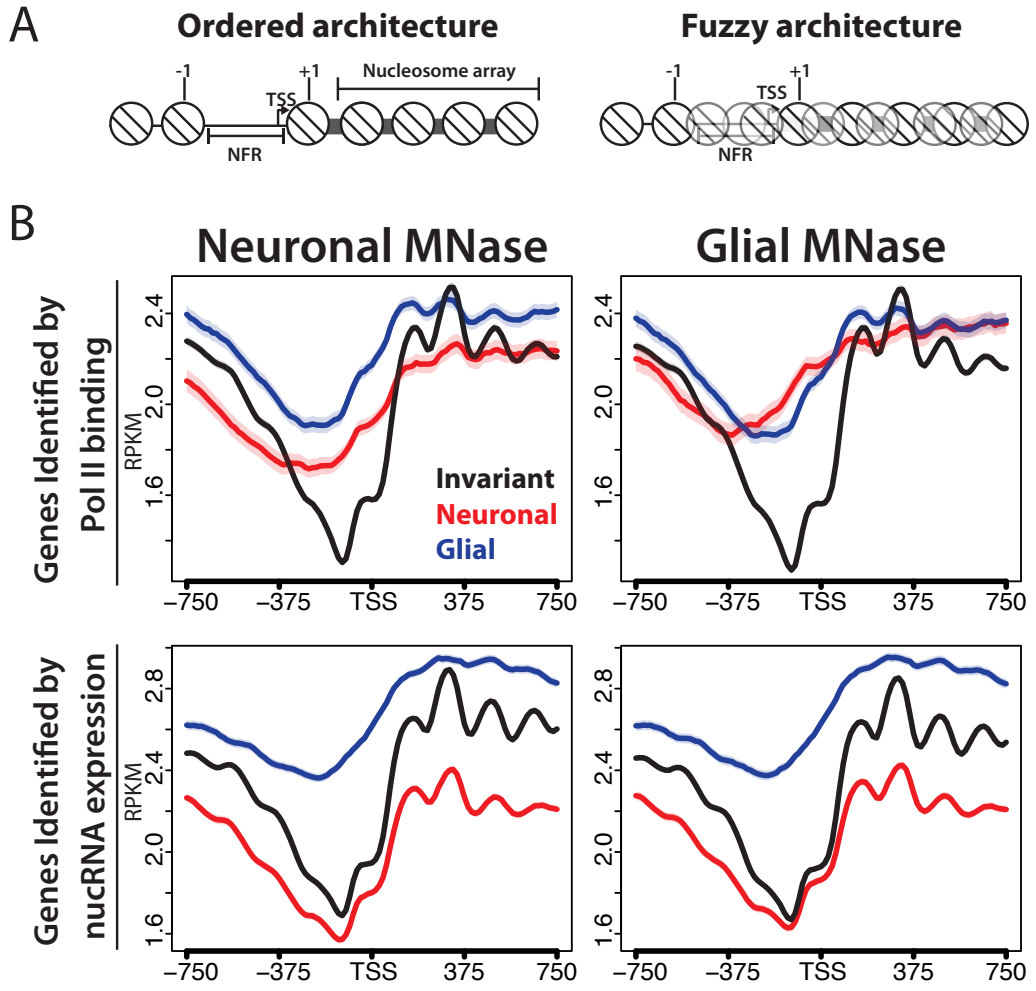


Figure 4.3: Nucleosome architecture at cell-type-specific genes - Metagene analysis of the neuron and glia MNase-seq data. Data were divided by the type of analysis, either by Pol II binding (upper panels) or nucRNA analysis (lower panels), and further divided into three gene classes: invariant, neuron-enriched, or glia-enriched (black, red, and blue, respectively).

The nucleosome pattern of genes identified by nucRNA-seq show a completely different pattern than those genes identified by Pol II binding (bottom panels figure 4.3). It is immediately obvious that the nucleosome patterns in the two cell types are nearly identical, even though there were significant differences in nucRNA expression level between neurons and glia. The neuron-enriched genes had the lowest nucleosome occupancy over the gene body, consistent with higher levels of expression. The neuron-enriched genes also had a deep NFR and an ordered nucleosome array. This pattern is highly similar to that of the invariant genes, although invariant genes has far higher nucleosome occupancy over the gene body, likely because the invariant genes were not as highly expressed as the neuronal genes, leading to less clearance of nucleosomes from the gene body. The similarity in the nucleosome pattern at invariant and neuronal genes makes sense in that these genes were expressed to some level in both neurons and glia, and had equivalent levels of Pol II binding at the promoter in both cell types. The amount of expression of the neuron-enriched genes was much higher in neurons than glia. Thus, an open "active" promoter architecture would be expected, particularly since the level of Pol II binding influences nucleosome occupancy over the NFR more than expression level (chapter 5). The glia-enriched genes showed the opposite pattern to the neuronal and invariant genes. Glia-enriched genes had high nucleosome occupancy, and no detectable nucleosome positioning, in either neuron or glia cells. Both Pol II binding and gene expression levels of the glia-enriched genes was very low, even in glia cells, consistent with the observed nucleosome pattern.

4.5 H2AZ correlates with invariant Pol II binding

We have previously shown that genes with cell-type-specific Pol II binding lack the histone variant H2AZ (34). The question is whether this observation still holds for the neuronal and glial enriched genes that were identified by nucRNA analysis. To address this question, I generated average profile plots of the head H2AZ data from (34), aligned to the TSS of the different gene classes (figure 4.4). Confirming our previous observation, the genes that are cell-type-specific based on Pol II binding lacked H2AZ enrichment, whereas the invariant genes had high H2AZ enrichment (figure 4.4, left panel). In contrast to this, genes identified as neuron-enriched in the nucRNA analysis had H2AZ levels nearly equivalent to that of invariant genes. Yet, genes identified as glia-enriched by the nucRNA analysis had no H2AZ enrichment, and so showed no difference from the Pol II-identified glia genes (figure 4.4, right panel).

Why do neuronal-enriched genes, based on nucRNA analysis, have incorporation of H2AZ, even though they have cell-type-specific expression? The answer may be that the nucRNA-identified neuronal gene have invariant Pol II binding and nucleosome architecture between the cell types (figures 4.3, bottom panels, and 4.2 top panels). The finding that H2AZ is associated with invariant genes, and not cell-type-specific genes, was based on classifying genes using Pol II binding. The neuron-enriched genes have highly ordered nucleosomes, with a deep nucleosome-free region and have the same level of Pol II binding in both cell types. Thus the incorporation of H2AZ at these genes may be influenced more by the cell-type-invariant recruitment of Pol II to the promoter, rather than by the actual transcriptional output of the gene, which is cell-type-specific. This observation would suggest that H2AZ plays a role outside of regulating the rate of transcriptional elongation of the gene. The relationship between gene activity, H2AZ binding and promoter architecture will be explored more deeply in the next chapter (5).

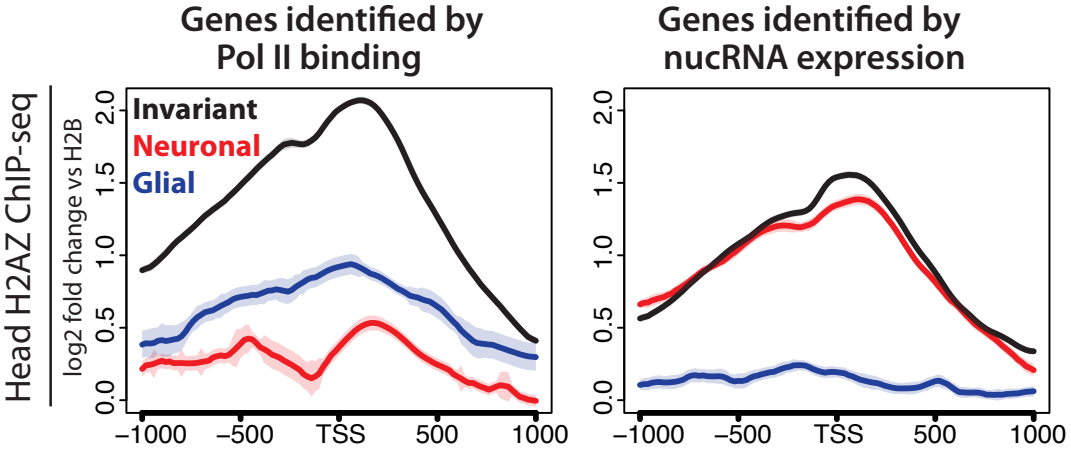


Figure 4.4: H2AZ profile at cell-type-specific genes - Metagene analysis of H2AZ binding in whole head, divided into analysis type: genes called as specific or invariant based on Pol II binding (left panel) or based on nucRNA analysis (right panel), then further divided into invariant, neuron-enriched and glia-enriched gene classes (black, red, blue, respectively) and split.

4.6 Histone modifications at cell-type-specific genes

The previous section indicates that incorporation of H2AZ correlates with invariant Pol II recruitment and ordered nucleosome architecture, rather than correlating with gene expression level. It would be therefore interesting to observe whether the active histone modifications H3K27ac and H3K36me3 follow the same trend as H2AZ, marking a cell-type-invariant chromatin state, or whether they follow the expected patterns, where the marks correlate with transcriptional activity. I generated metagene plots to observe the average trend of H3K27ac and H3K36me3 at the neuron-enriched, glia-enriched, and invariant genes identified by either Pol II binding or nucRNA analysis (figures 4.5, and 4.6). I chose to plot the H3K27ac data across the TSS $\pm$ 1000 bp, since the general enrichment of these data are at the 5′ end of the gene (figure 3.5, C). Conversely, I plotted the H3K36me3 modification centred around the transcriptional end site (TES) $\pm$ 1000 bp, as this modification is generally enriched at the 3′ ends of genes (figure 3.5, C). Each plot is normalised to the level of nucleosomes using cell-type-specific CAST-ChIP H2B-GFP data (T. Schauer, *unpublished*), to ensure high histone modification levels are not due to high nucleosome occupancy over that region.

H3K27ac correlates with gene activity

The active histone modification H3K27ac was observed as expected, and closely reflected the activity of each gene group (figure 4.5). Where there is Pol II binding, there is a corresponding level of H3K27ac (figure 4.5, upper panels), with enrichment of the active mark reflecting also cell-type-specificity of Pol II binding. For example, neuron-enriched genes had higher H3K27ac than glia-enriched genes in neurons (upper panels, figure 4.5). With the nucRNA gene calls, the level of H3K27ac also appeared to correlate with the RNA expression level (bottom panels, figure 4.5. For instance, neuron-enriched genes in neurons had far higher H3K27ac relative to invariant genes, than the neuron-enriched genes in glia did. Thus, H3K27ac showed the expected pattern as a mark of gene activity; where there was gene activity, either by nucRNA or Pol II binding, H3K27ac was present at a comparable level.

H3K36me3 does not correlate with gene activity

The results for the H3K36me3 were not as straightforward to interpret as those of the H3K27ac, in that there does not seem to be a direct correlation between this mark and gene activity (figure 4.6). The most interesting observation comes from the profiles of the Pol II gene calls with the neuron-specific H3K36me3 ChIP-seq data (top left panel, figure 4.6). Here there was no difference in H3K36me3 binding levels between

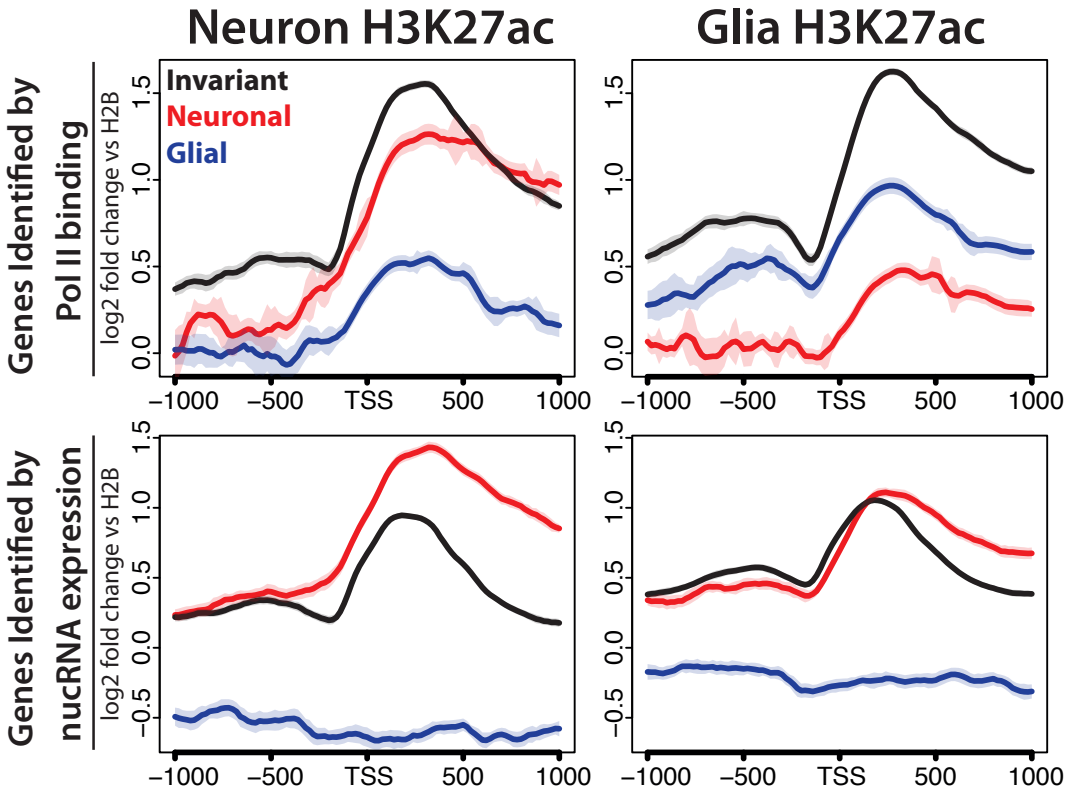


Figure 4.5: H3K27ac at cell-type-specific genes - Metagene analysis of the neuron and glia H3K27ac ChIP-seq data. Data was split by type of analysis, either by Pol II binding (upper panels) or nucRNA analysis (lower panels), and further divided into three gene classes: invariant, neuron-enriched, or glia-enriched (black, red, and blue, respectively).

the neuron and glia enriched genes, even though the neuronal enriched genes have far higher levels of Pol II and nucRNA in this cell type (figure 4.1). The level of H3K36me3 was also far lower at cell-type-specific genes compared to invariant genes in neurons, with genes identified by Pol II binding (upper left panel, figure 4.6). Within glia cells, this histone modification seemed to partially reflect the activity of glia-enriched genes (two right panels, figure 4.6). The glia-enriched genes, by Pol II binding, had higher levels of H3K36me3 than the neuron-enriched genes in glia. Also, those glia-enriched genes identified by nucRNA analysis had a higher level of H3K36me3 in glia cells than in neurons. However, the level of H3K36me3 did not correlate well with the activity level of neuronal nucRNA genes (red line, figure 4.6, bottom panels). With these nucRNA-neuronal genes, the level of H3K36me3 was approximately the same enrichment as invariant genes in both neurons and glia. Thus, there was not the increase in H3K36me3 in neurons compared to glia that would be expected with the increase in transcriptional output of neuron-enriched genes.

4.7 H3K27ac and H3K36me3 in gene activity

The previous observations indicate that H3K27ac correlates with gene activity, whereas H3K36me3 only appears to correlate with gene activity in some cases, such as for the glia-enriched genes. I therefore wanted to look more closely at the relationship between the amount of Pol II binding or nucRNA expression and the level of the histone marks. To do this, I ranked the genes by the level of Pol II binding, based on the read count ± 250 bp around the TSS, or by normalised read counts (RPKM) of the nucRNA data. For ranking the neuron-enriched genes I used the neuronal datasets for the ranking, for the glia-enriched genes I used the glial datasets for ranking, and for the invariant genes I used either whole head Pol II binding data (for Pol II-identified genes), or the average of all nucRNA inputs RPKMs (for nucRNA-identified genes). I then plotted the histone modification data as heatmaps, normalised to the H2B-GFP, ordered from the highest Pol II binding/nucRNA levels at the top to lowest at the bottom (figures 4.7 and 4.8). Heatmaps showing the Pol II data and nucRNA data, ranked in the same way, show that the ranking was performed correctly (figures 5.1 and 5.3).

There is little correlation between Pol II binding and active histone marks

From these analyses, the amount of Pol II bound at the promoter region of cell-type-specific genes did not strongly correlate with the level of H3K27ac (figure 4.7). There appeared to be little correlation between Pol II binding and H3K36me3 levels (figure 4.7). Higher Pol II occupancy at invariant gene promoters seemed to lead to higher

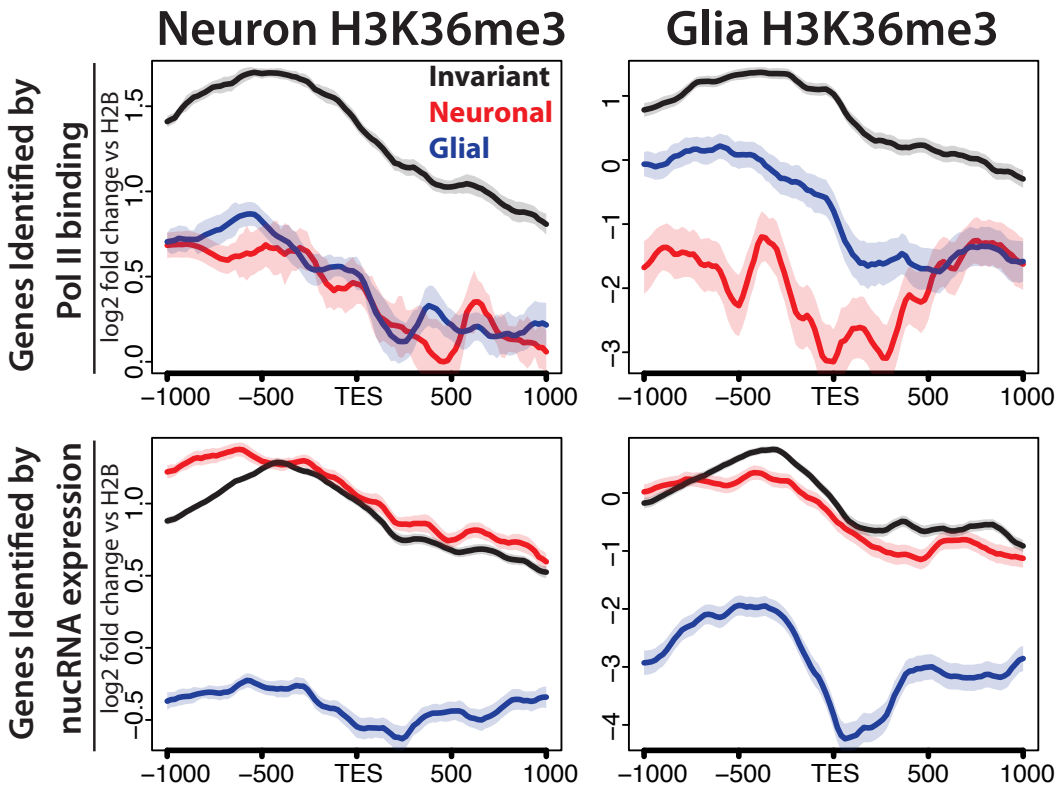


Figure 4.6: H3K36me3 at cell-type-specific genes - Metagene analysis of the neuron and glia H3K36me3 ChIP-seq data. Data was split by type of analysis, either by Pol II binding (upper panels) or nucRNA analysis (lower panels), and further divided into three gene classes: invariant, neuron-enriched, or glia-enriched (black, red, and blue, respectively).

levels of H3K27ac, however the genes with the highest Pol II binding appeared to have less H3K36me3. One explanation for this observation is that the genes with the highest Pol II binding at the promoter are "stalled" or "paused" genes. If these genes are regulated by Pol II pausing they may not actually be transcribed, and since H3K36me3 is deposited in conjunction with the transcribing polymerase, H3K36 would not be methylated at these genes.

H3K27ac correlates with nucRNA expression level

When the genes were ranked by nucRNA expression level, a general trend of H3K27ac could be observed (figure 4.8). The most highly expressed genes had high levels of H3K27ac, with the mark appearing to spread into the gene body. The lower the expression; the lower the H3K27ac level was. It is interesting to note that both the invariant and neuronal gene classes still had H3K27ac even at the lowest gene expression levels, where nucRNA levels were barely detectable (figure 5.3. As discussed previously, the chromatin signature (ordered nucleosomes, Pol II binding, and H2AZ-incorporation) of the neuron-enriched gene class was the same as that of invariant genes (section 4.5), so perhaps presence of some H3K27ac is associated with genes that are maintained in an open/active state. This maintenance of the active state could ensure that the neuronal-enriched and invariant promoters are potentiated for transcription by having open/active promoters, with incorporation of H2AZ and some H3K27ac. Other factors would influence the activation of transcription of these potentiated genes, which would lead to increased H3K27ac levels.

Glia-enriched genes, as identified by nucRNA analysis, showed a different pattern of histone marks from neuronal and invariant genes. High expression of glial genes within glia correlated with high H3K27ac. Glia-enriched genes with low expression had little detectable H3K27ac, and this modification was barely detectable in neuronal cells (figure 4.8). A difference in the type of regulation of the glia-eniriched genes from neuron-enriched genes could explain this observation. If glia-enriched genes are normally repressed in all tissues, the chromatin state would have no active marks and high nucleosome occupancy, to ensure the complete shutdown of these genes. Then the de-repression of these genes in glia would lead to specific H3K27ac in glia. Another interpretation is that if these glia genes are specifically repressed in neurons, then the active marks would be specifically removed in neurons, to ensure that no expression is possible in neurons.

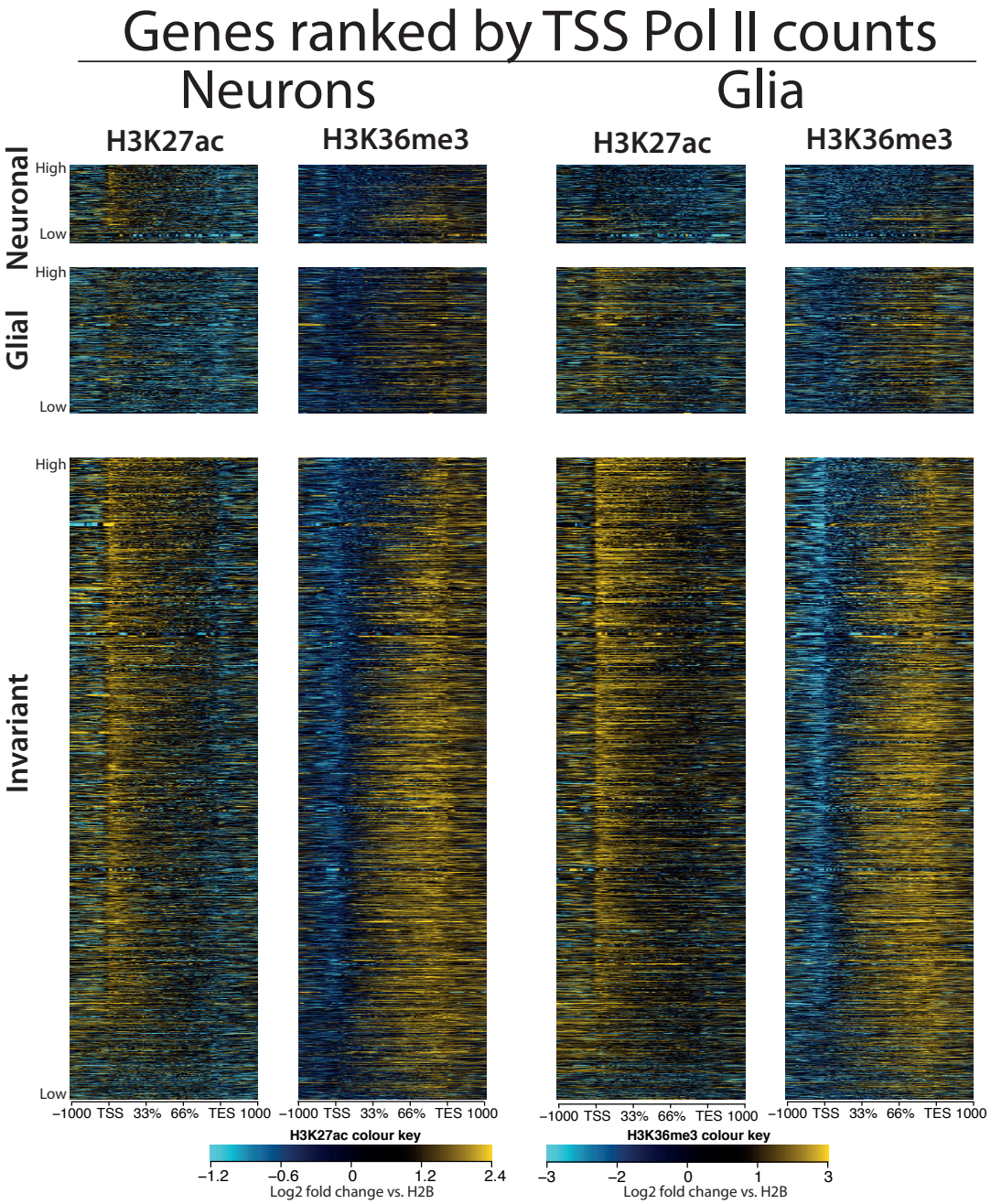


Figure 4.7: Active histone marks ordered by Pol II binding at TSS - Heatmaps of H3K27ac and H3K36me3 data across neuronal, glial and invariant genes. Genes were ordered by calculating number of reads over a 500 bp window across the TSS and ranking genes from highest reads to lowest.

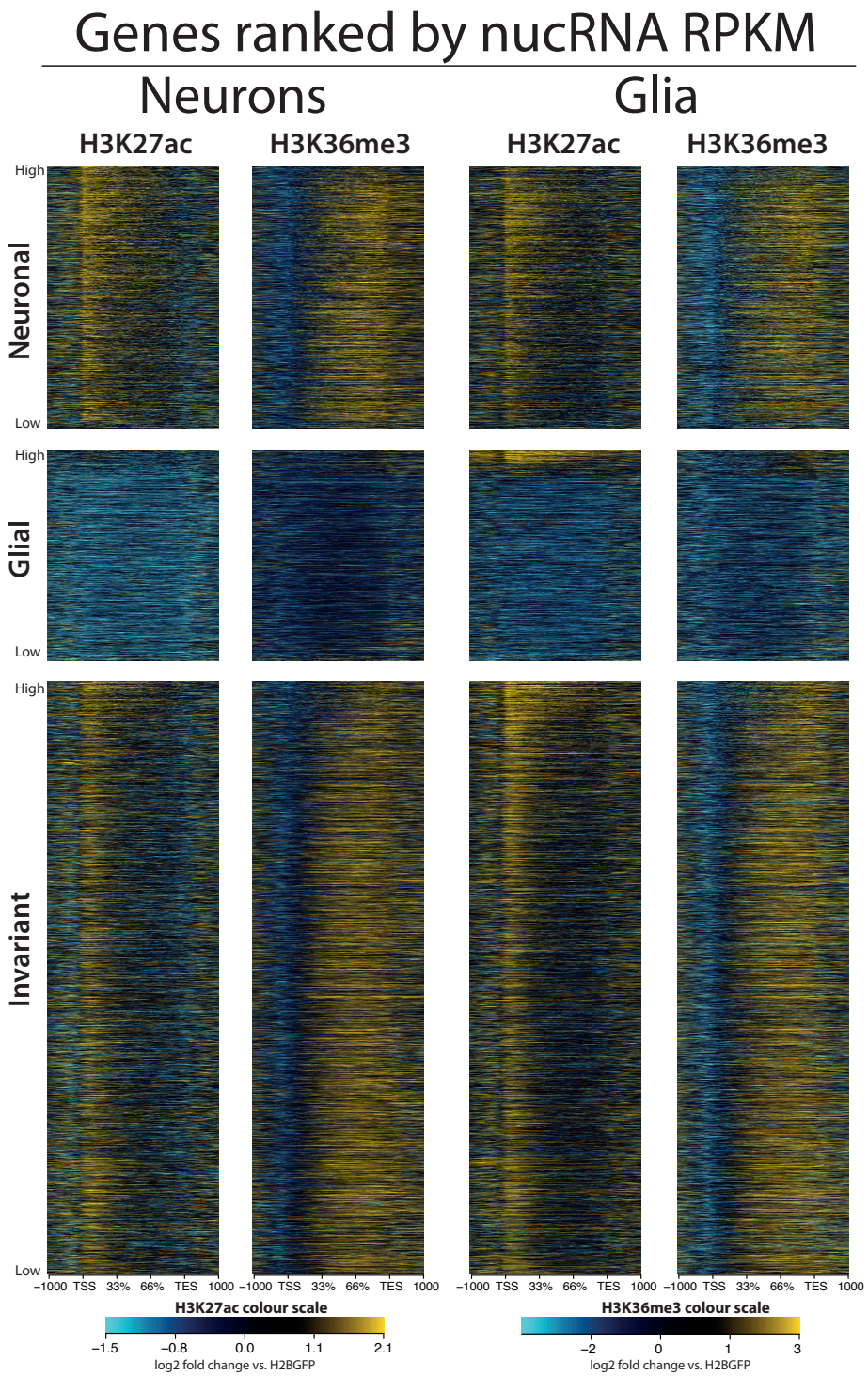


Figure 4.8: Active histone marks ranked by gene expression - Heatmaps of H3K27ac and H3K36me3 across neuronal, glial and invariant genes. Genes are ranked by RPKM level in the cell-type that the gene was called specific in, or the RPKM of the average input samples.

H3K36me3 does not correlate with nucRNA expression levels

Tri-methylation of H3K36 did not generally correlate with nucRNA expression levels (figure 4.8). Indeed, as with the highest level of Pol II binding, the most highly expressed genes had less H3K36me3 across the gene body than the lowest expressed genes. This observation was not unexpected as infrequently transcribed genes were shown to be more dependant on H3K36me3 than highly transcribed genes (159). Only the most highly expressed glia-enriched genes within glia cells appeared to show some enrichment for H3K36me3. These analyses showed that there was little correlation between the level of gene expression and the level of H3K36me3. Those genes at the higher end of the expression spectrum had the same, or less, H3K36me3 than those genes at the lower end of the gene expression spectrum. Only in the very lowest expressed genes, at the very bottom of each heatmap, was there a reduction in H3K36me3 levels. It is likely that these genes are repressed and may have never been transcribed, thus would not have the histone modification. It is also possible that the H3K36me3 is actively removed at these genes (by Kdm4), and so is not detected.

The lack of relationship between gene activity and H3K36me3 likely contributes to the observation that this modification is mostly invariant between cell types. H3K36me3 functions to repress cryptic transcription, and does so by reducing nucleosome turnover (see section 1.1.3). Therefore, any transcription of the gene, whether frequent or infrequent, would acquire and retain this modification. The average accumulation of H3K36me3 would then be expected to be similar between all active genes. For example, if a gene was expressed at four copies per cell in glia and 100 copies per cell in neurons, the level of H3K36me3 across that gene would be the same in both cell types because the modification would be set after the first round of transcription, and would be maintained regardless of how many rounds of transcription there were. Extremely high levels of gene expression that result in massive histone loss, such as the heat shock response, would probably lead to lower H3K36me3 since the histones themselves are no longer there to be modified.

Taken together, the results from my analysis clarify the observations made in section 4.6. Where there is more Pol II binding, or higher nucRNA expression, there are higher levels of H3K27ac, thus this modification is useful as a marker of gene activity. The level of H3K36me3 does not directly correlate with the level of gene activity, and appears to be enriched at any gene that is in an active state, regardless of the level of that activity. Hence, H3K36me3 is not a useful marker of gene activity and would not aid in

identifying gene activity differences between cell types. However, the observation that the neuron-enriched genes, as identified by Pol II binding, have both Pol II binding and nucRNA transcripts but lack H3K36me3 was not explained by these analyses. The lack of H3K36me3 at expressed genes has been observed within the "red" chromatin domains in the five state chromatin model (figure 1.5, (76)). Red chromatin contains those genes which are specifically activated. These specifically regulated genes have active gene expression, yet lack H3K36me3. Interestingly, the neuron-enriched Pol II genes strongly displaced this chromatin state, yet the glia-enriched genes did not. For the genes whose specificity was assessed on the nucRNA level, those neuronal genes had high levels of H3K36me3 in both cell types. Further, in the glia nucRNA genes, H3K36me3 was cell-type-specifically enriched in glia cells compared to neurons. This underscores the complications with assigning combinations of chromatin features with gene-activity features, as clearly being specifically regulated does not necessarily mean that the genes will have the H3K36me3-less chromatin state.

4.8 Active, H3K36me3-less genes have broad H3K27ac

When looking at histone modification data at individual genes (figure 4.9), there were indeed active genes with complete absence of H3K36me3. Genome-browser snapshots in figure 4.9 show examples of genes that were expressed and also had Pol II binding, but no H3K36me3 binding was detected. These types of genes were not always cell-type-specific. In some cases they were neuronal, in others they were glial, but in many cases this chromatin state was found where the gene expression was invariant between the two cell types. One clear feature of these genes was that H3K27ac appeared to spread across the entire gene in a broad peak, sometimes stretching beyond the boundaries of the gene (e.g glec and SoxN, figure 4.9). This spreading of H3K27ac would be consistent with the function of H3K36me3 to recruit the histone deacetylase complex RPD3, which de-acetylates H3K27ac. If H3K36me3 is absent from these genes, then the H3K27ac would not be removed from the $3'$ end of the gene and so remain dispersed across the active gene body. The intriguing question is how do some genes escape regulation by H3K36me3? And what are the consequences of this type of regulation, particularly when H3K36me3 is important for suppressing cryptic transcription and regulating splicing? Could there be a functional consequence of such a chromatin state that was selected for at these genes, such as generation of non-coding regulatory RNAs from within the gene body?

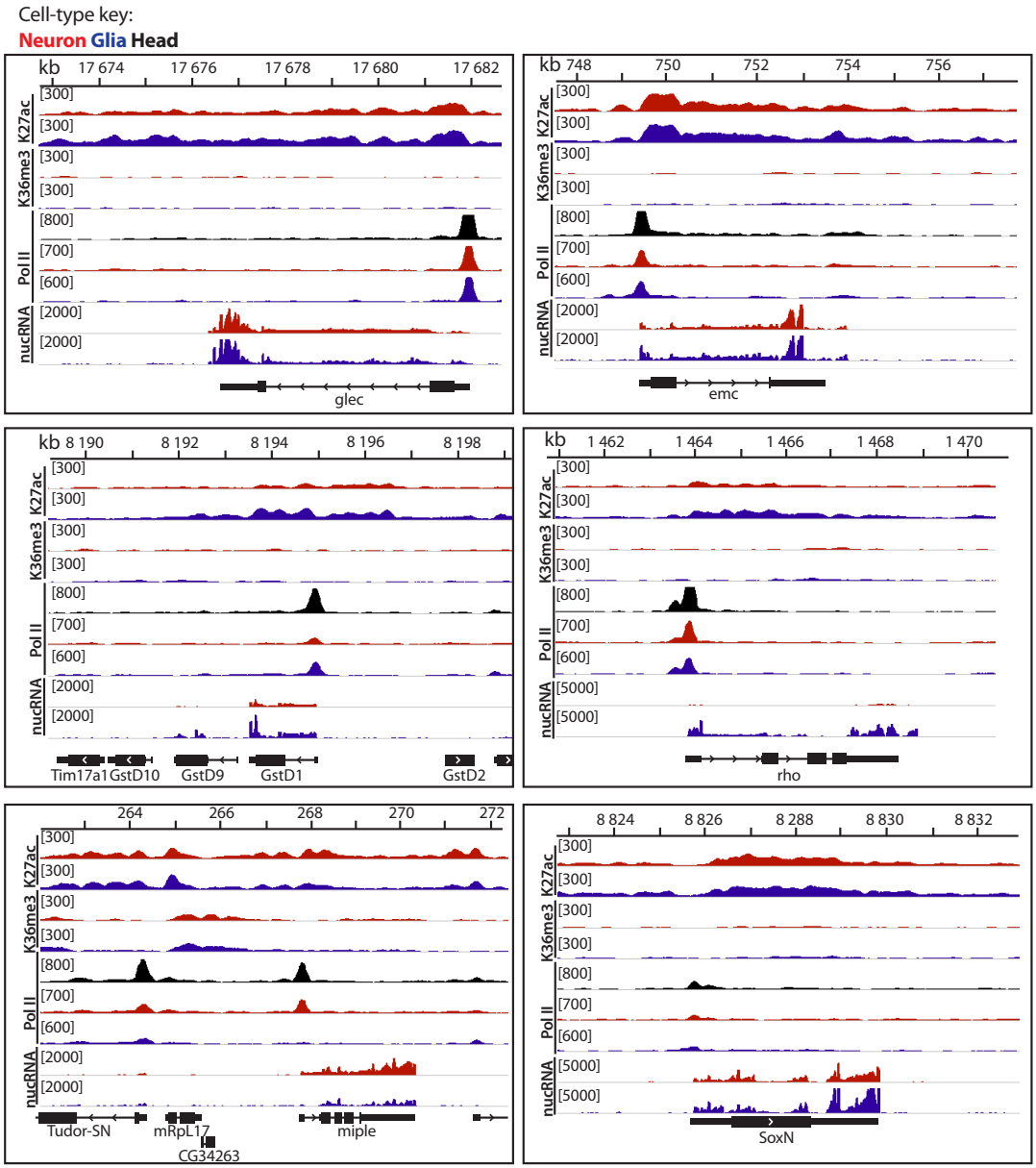


Figure 4.9: Expressed genes with no H3K36me3 have broad H3K27ac binding - Snapshots of the H3K27ac, H3K36me3, Pol II, and nucRNA data at several genomic regions using the IGV-browser. Representative gene regions are shown that are active; have Pol II and nucRNA, and have no H3K36me3. There is a clearly observable spreading of H3K27ac across these regions.There are examples of this chromatin state where the Pol II and nucRNA is invariant (glec, emc, rho, SoxN) or cell-type-specific (GstD1, miple).

Do those genes with medium-high expression and no H3K36me3 show a spreading of H3K27ac across the gene body? To address this question, I split all annotated genes into those having an identified H3K36me3 peak and those that do not. Since the peak calling of H3K36me3 may be different in each cell type, I analysed each cell type separately, producing a list of genes with and without H3K36me3 for each cell type. I then further split these genes into subsets based on the RPKM of the cell type the H3K36me3 peaks were defined in. This produced a list of genes with or with out H3K36me3, each further divided into high expression (RPKM greater than 100), medium expression (RPKM 30–100) or low expression (RPKM lower than 30). The number of genes in each category is shown in table 4.2.

Table 4.2: Splitting based on H3K36me3 and RPKM

	RPKM	**+ H3K36me3**	**-H3K36me3**
Neurons	High	497	143
	Medium	1506	397
	Low	5450	4676
Glia	High	524	155
	Medium	2238	505
	Low	4809	4434

I then assessed the average chromatin profile for the six H3K36me3 gene classes, by generating average profile plots of the H3K27ac and H3K36me3 data from both neurons and glia (figure 4.10). The top-most panels of sections A and B in figure 4.10 is a control plot. These control plots show that the peak calls from the H3K36me3 data do indeed correspond to the actual H3K36me3 enrichment across these gene sets. These average profile plots further demonstrate that the gene expression level did not influence the level of H3K36me3 (top left panel of figure 4.10, parts A and B). Here it could be seen that the lowest expression class had slightly higher levels of H3K36me3 than the medium and high expression groups. The H3K36me3 binding across the genes classified as having no H3K36me3 was very low (upper right panels, figure 4.10, parts A and B). This demonstrated that the peak calling, and subsequent gene classifications was consistent with how the data actually looks on the metagene level.

The bottom panels of figure 4.10, show the average H3K27ac profile across each gene group in neurons (figure 4.10, A) and glia (figure 4.10, B). Of those genes that have H3K36me3, the level of H3K27ac enrichment corresponds to the level of expression. The shape of the H3K27ac profile at H3K36me3-positive genes is also as would be

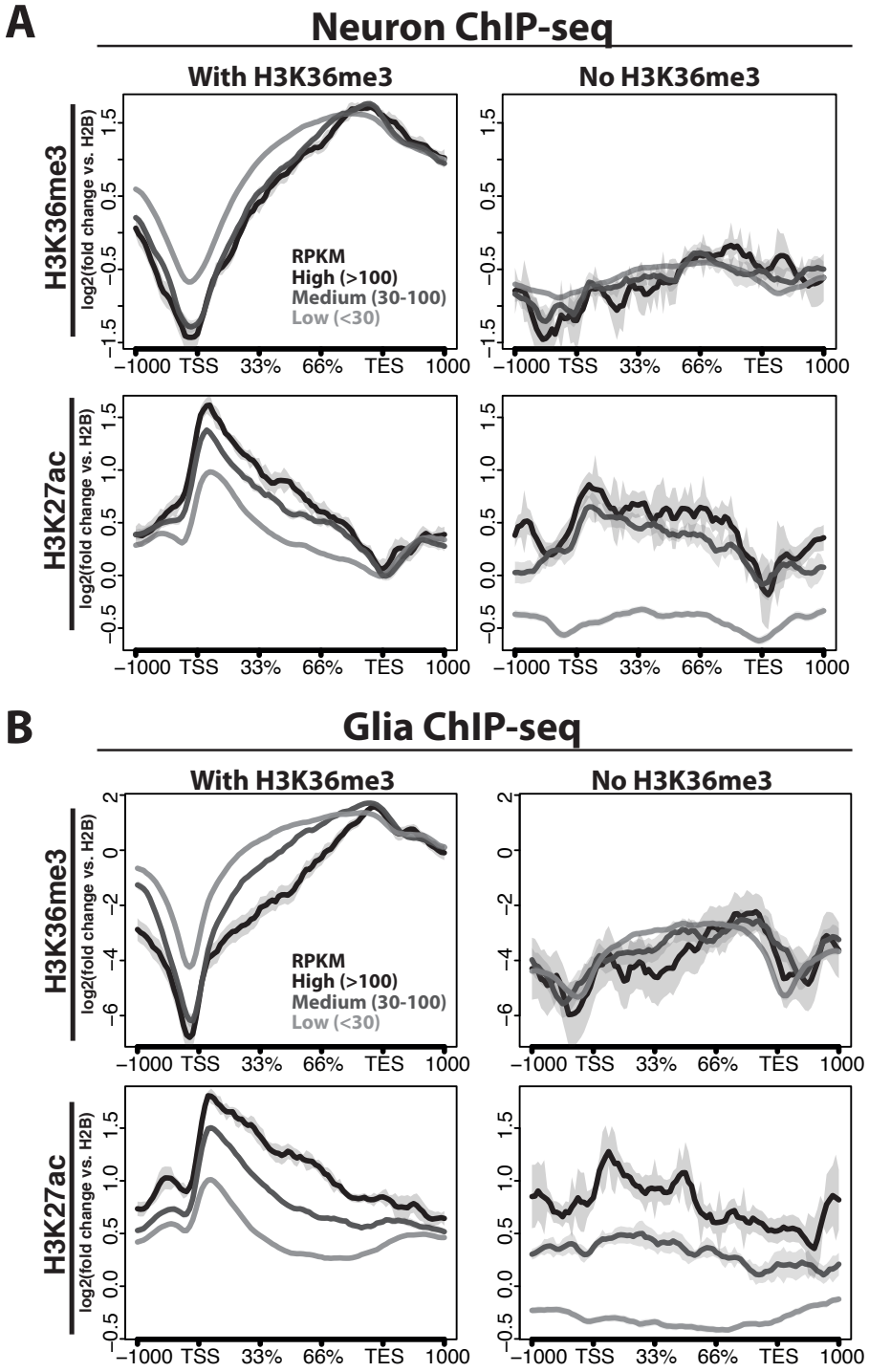


Figure 4.10: Active genes with no H3K36me3 have broad H3K27ac - Genes were first split based on whether an H3K36me3 peak was called or not, then each group was further divided as having high, medium or low expression. A) The average profile analysis in neurons. B) The averge profiles in glia.

expected from this histone modification: there is high enrichment at the 5′ end of the gene that decreases in a slope towards the 3′ end of the gene. At H3K36me3-negative genes, the level of H3K27ac still corresponds to the level of gene expression, with low expressed genes having no enrichment of this modification at all. However, the shape of the H3K27ac profile is vastly different from the H3K36me3-positive genes. As observed in the genome-browser snapshots (figure 4.9), the H3K27ac profile across H3K36me3-negative genes with medium to high expression is flat and broad. There is very little 5′ bias to the enrichment and consequently very little slope to the profile. I would expect that other active histone modifications, particularly histone acetylation, would also be spread across these genes. Indeed an inspection of publicly available data of many histone modifications in the head supports this hypothesis (data from the modencode consortium, http://www.modencode.org).

4.9 Su(Hw) is enriched at H3K36me3-less neuronal genes

It is straightforward to see how the active, H3K36me3-less genes have broad H3K27ac, and likely broad enrichment of many active histone modifications. However, what the functional consequence of such a chromatin state is and why these particular genes have this pattern is unclear. The genes identified as neuron-enriched from the Pol II analysis most strongly display this chromatin state. It has been previously shown that binding of the insulator protein su(Hw) is enriched at genes with neuronal function (93). We have also confirmed the observation that su(Hw) is present in glia, yet is absent from neurons ((93), and Mirjam Appel, *unpublished*). The question arises as to whether all neuron-enriched genes contain su(Hw) sites, or whether there is some preference for particular neuronal genes. I hypothesised that those genes identified as neuron-enriched by Pol II binding, which present the unusual H3K36me3-less chromatin structure, would contain more su(Hw) binding sites than those neuronal genes with a constitutively "active" chromatin structure. If this were true, it would present a very interesting mode of regulating those neuron-enriched genes, where loss of su(Hw) in neurons would lead to a de-repression of those genes, and the subsequent incorporation of broad histone acetylation would push those neuronal genes into a highly active state, specifically in neurons.

Our bioinformatics collaborator Pawel Bednarz analysed the insulator occupancy of the different neuronal-enriched gene groups identified in my analysis (figure 4.11, A). Shown in figure 4.11 (A) are two gene groups that have the two distinct chromatin

states: 1) Those genes identified by Pol II binding, which have fuzzy nucleosome architecture, broad H3K27ac and no H3K36me3. 2) A further refined group of genes identified by the nucRNA analysis, those genes which are higher expressed in neurons compared to whole fly, and have invariant expression between glia and whole fly. This second group has highly ordered nucleosome architecture, H2AZ, and the normal expected patterns of H3K27ac and H3K36me3. The refined group of nucRNA-identified genes was used because the total nucRNA group actually contains some genes with the same chromatin state as identified by the Pol II binding. In the refined group, those genes that are likely regulated by Pol II recruitment, and have the active H3K36me3-less chromatin state are depleted from the gene list. In the analysis, publicly available ChIP-seq data of four insulator binding proteins (suHw, BEAF, CP190, and CTCF) in *Drosophila* Kc167 cells, were used to define insulator binding sites (94).

Our approach was to ask what insulator binding sites were found associated with the genes in each of the three gene groups. Different methods for assessing the relationship between the neuron-enriched genes and insulator binding sites were investigated. The first approach was to look at the nearest insulator flanking the gene (figure 4.11, B, left panel). In this approach the identity of the nearest insulator to the $5'$ and $3'$ end of the gene was assessed. As can be seen in the bar graph (figure 4.11, B, left panel), there is little difference in the distributions of the four insulators between the two gene groups. Thus, there is no enrichment for a particular insulator for either gene group when looking at the nearest flanking insulator. However, the most interesting result came from looking at the insulator binding inside the neuronal gene regions (figure 4.11, B right panel). We decided to analyse the internal sites as we knew neuronal genes are generally much longer than the average gene, and contain very large introns (not shown), thus increasing the possibility for regulatory proteins to bind within the gene body. When identifying the internal insulator binding sites, all insulator binding sites that fell within the annotated gene region were counted, thus each gene could have multiple insulator binding sites. It can be seen in the bar graph (figure 4.11, B right panel) that su(Hw) sites are significantly enriched in the Pol II-identified neuronal gene class, compared to the neuron-enriched genes identified by nucRNA (glia-invariant, neuron-up compared to whole fly). Interestingly, the proportions of CTCF and CP190 binding sites remain relatively constant between the gene groups, but BEAF is increased in proportion in the nucRNA gene group. Thus, when looking at internal insulator binding, those neuron-enriched genes with the fuzzy nucleosome architecture, broad H3K27ac and low H3K36me3 have significantly more su(Hw) binding sites than those neuron-enriched genes with a more ordered, standard chromatin signature.

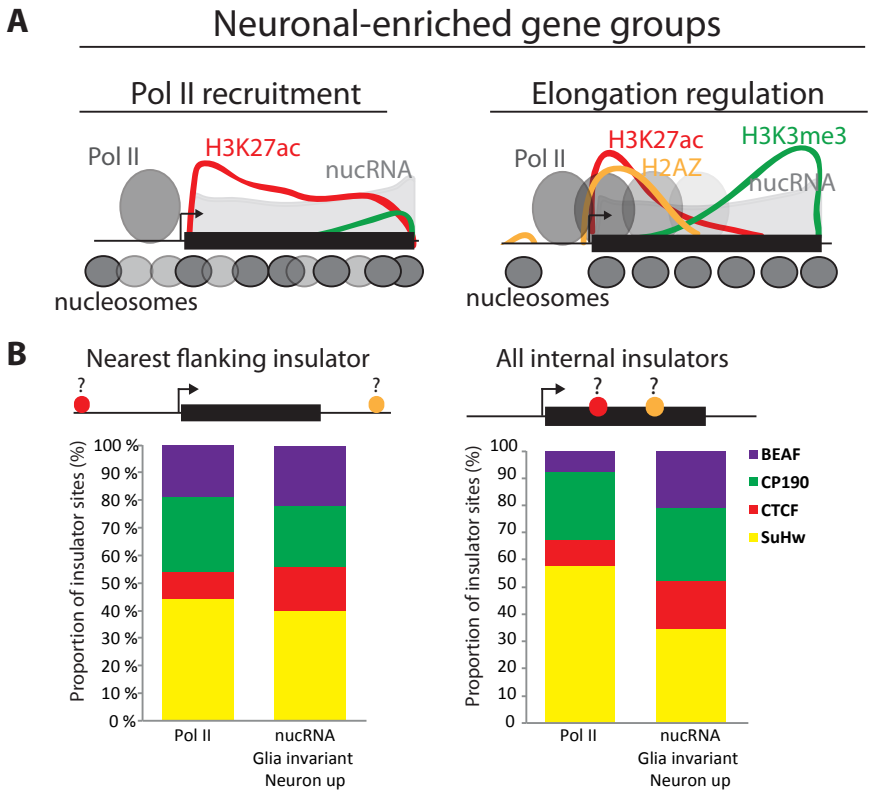


Figure 4.11: SuHw is enriched at fuzzy neuronal genes - A) A graphical summary of the chromatin signature at the neuronal-enriched gene groups identified from Pol II binding, or from a refined nucRNA analysis. The nucRNA analysis genes were refined according to whether the gene expression in glia was equal to whole fly and expression in neurons was up compared to whole fly. B) Publicly available insulator binding data for suHw, BEAF, CP190, CTCF were used to define regions of insulator binding. The number of insulator regions was counted for each type of insulator (red and yellow circles in the models) that was either the nearest flanking insulator to the gene (left), or within the gene body of each gene (right) for the three gene groups. The bar graph shows the proportion that each insulator contributes as a percent of the total number of insulator sites found.

4.10 Discussion

The observations made in this chapter highlight two distinct mechanisms for achieving cell-type-specific gene regulation, summarised in figure 4.12. In the one case, there is a cell-type-specific recruitment of RNA polymerase to the promoter. The chromatin state appears to play a highly important role in this type of regulation. The promoters have a closed, or "fuzzy" nucleosome architecture, which would facilitate highly specific recruitment of Pol II to the gene promoter. To recruit Pol II to these fuzzy promoters would require the concerted effort of multiple transcription factors and chromatin remodellers in order to allow the PIC to compete with the nucleosomes for promoter DNA binding.

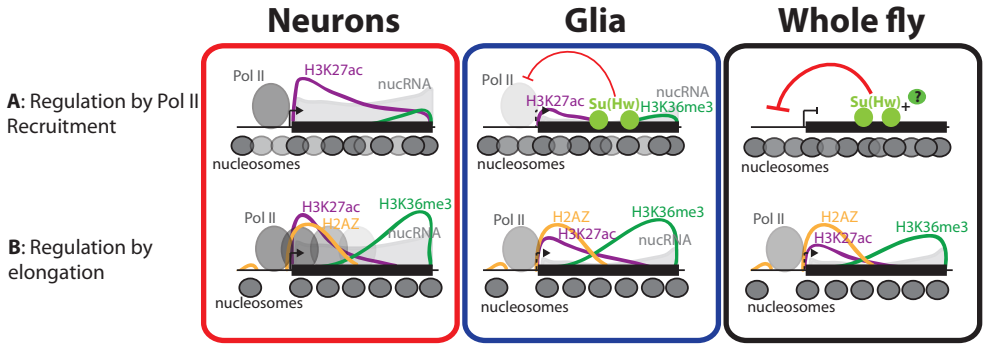


Figure 4.12: Model of the different neuron-specific gene regulation mechanisms - Model of the modes of cell-type-specific regulation of the neuronal-specific genes observed in this chapter. Genes regulated at the level of Pol II recruitment (A) show differences in chromatin state between neurons and glia. These genes are likely to be associated with the insulator protein, Su(Hw) in other tissues. Lack of SuHw and increased histone acetylation in neurons would guide neuron-specific expression.Presence of Su(Hw) in glia may aid in preventing Pol II recruitment to the promoter. Genes regulated by Pol II elongation (B) have a canonical "active" chromatin state, but how the specific modulation of elongation is achieved is not clear.

Neuronal gene regulation through loss of su(Hw)

It is still not clear what the function is for these genes having highly disordered nucleosome occupancy, broad spreading of H2K27ac and lack of H3K36me3. Those genes with this unusual chromatin state are enriched for su(Hw) binding, suggesting an interesting mechanism for regulating some neuron-specific genes illustrated in figure 4.13. In non-neuronal tissues su(Hw) would be present and binding to the regions within the neuronal-enriched genes. This su(Hw) binding would lead to a repression of those neuronal genes in other tissues, but in neurons, where su(Hw) is not present, these genes

would become active. Therefore, the neuron-specificity of these genes is not necessarily caused by neuron-specific enhancer regions or neuron-specific transcription factors, lack of su(Hw) binding could be enough to achieve this specificity. Neurons and glia cells are derived from the same lineage; the "ancestral" cell type is the neuroblast, and neurons and glia can be separated by as little as two cell divisions. Su(Hw) is expressed in the larval neuroblast (104), thus su(Hw) expression must be specifically repressed in neurons after the neuronal and glial lineage separation. Su(Hw) was shown to be more highly enriched at lamin-associated domains (LADs) (89), indicating that the neuronal genes may be associated within the repressive environment of the nuclear lamin in tissues outside of neurons. If it is the case that these neuron-enriched genes are associated with a repressive environment up until the final cell division, then I would expect that these genes would remain within the space of that repressive environment within neurons, since genomic regions do not change much during developmental shifts (160). Thus, when su(Hw) is absent from neurons, much of the repression would be diminished, but it is likely that the gene activity requires additional enhancement to achieve full expression. This additional enhancement could be achieved through broad incorporation of the histone acteylation marks. Having broad domains of H3K27ac (and likely H3K18ac) would aid in "opening" the chromatin, and making it more accessible within the domain, and prevent interaction with the lamin. This would then presumably enable the gene to be expressed even if the gene was near to or within a generally repressive environment such as the nuclear lamin (89).

Regulation of gene expression through modulating elongation

The neuron-specific nucRNA genes had an extremely large difference in expression level between neurons and glia. Yet, there was no difference in nucleosome occupancy, chromatin marks, or Pol II recruitment at these genes between the two cell types. These genes had open promoters in both cell types, thus recruiting Pol II to these genes would not involve competitive binding between the transcription machinery and nucleosomes. It would be expected that these genes would not rely on specific transcription factors or chromatin remodellers for recruitment of Pol II, since Pol II is recruited in both cell types. However, specific inputs would be required to achieve the specific up-regulation of these genes in neurons. I make the assumption that since it is nuclear RNA being measured, and not total RNA, that the longevity of the RNA molecule is not the predominant factor leading to increased gene expression in one cell type. Therefore, the mechanism for the up regulation of these genes in likely lies in modulating the elongation rates of Pol II.

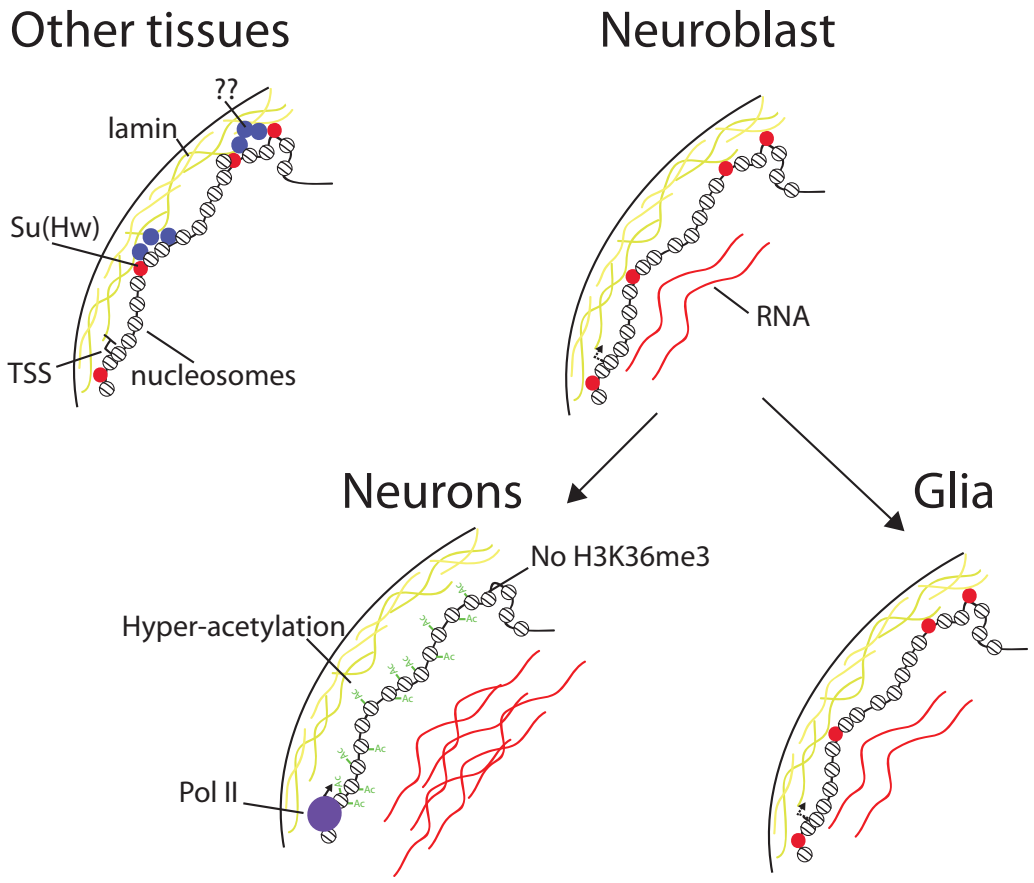


Figure 4.13: Model for achieving neuron-specific expression of su(Hw)-regulated genes - Schematic of how chromatin state and su(Hw) function amy regulate a subset of neuron-specific genes. In other tissues, meaning those tissues with a non-neural lineage, su(Hw) is present. In these cells, su(Hw) and other regulatory factors completely repress the neuronal genes. In the neuroblast, and in differentiated glia, su(Hw) is present but other factors may be absent, resulting in a slight de-repression of the neuronal genes in these cell types. In fully differentiated neurons, su(Hw) is absent. With the cumulative effect resulting from loss of su(Hw) and other repressive factors, the neuronal genes are highly de-repressed. Additionally, hyper-acetylation of these genes forms a highly active chromatin state, thus driving increased expression of these genes in neurons.

The importance of elongation in transcriptional control is only recently being appreciated (see review (161)). Even though the chromatin state across the elongation-regulated genes is invariant between cell types, it is most likely that there is still considerable contribution of chromatin features to the regulation of elongation. The contribution of underlying sequence elements in modulating the elongation rate may come from distal *cis* regulatory modules, rather than being directly at the promoter region. Defining the point(s) of the elongation process where the genes are regulated would be essential to know how the regulation could be achieved. There are many ways in which elongation can be modulated. For instance, the point of regulation could be through the regulated release of Pol II from a paused state to active elongation. Alternatively, the rate at which nucleosomes are removed/remodelled preceding the transcribing Pol II could be a major mechanism for specifically modulating the transcriptional rate. The involvement of non-coding RNAs or an influence by the splicing machinery are also possibilities for mediating the regulation of elongation.

Different methods for assaying specific gene activity reveal different regulatory mechanisms

The analysis here has revealed subsets of neuron-specific genes that are regulated by either Pol II recruitment or elongation. The methods used to identify which genes are cell-type-specific and which are invariant seem to greatly influence the type of cell-type-specific regulation that is observed. With the CAST-ChIP analysis of Pol II binding, there is obviously a bias for identifying genes regulated at the point of Pol II recruitment since differences in elongating Poll II are undetectable with the current data. The nucRNA expression data revealed a different bias, identifying mainly those genes that have significant differences in RNA expression levels without detectable differences in the chromatin state between cell types. By comparing and contrasting both techniques for identifying cell-type-specific genes, some sense might be made out of the seemingly conflicting datasets. These comparisons revealed that instead of looking for a single, unifying mechanism that guides neuron-specific gene expression, a far more complex system with many modes of regulation needs to be considered.

4.11 Future Directions

The observations made from the genome-wide analysis in this chapter lead to many questions and many possible directions for future work. The most intriguing direction would be to identify the mechanisms governing the specific regulation of transcriptional elongation, if indeed this regulation is at the level of transcriptional elongation. Other possibilities for achieving the same outcome could be increased mRNA stability, or decreased mRNA export in the cell type with the higher RNA levels. Thus initial experiments would need to test these possibilities. Such experiments could include "pulse-chase" experiments (TU-tagging, for example), where nascent RNAs can be distinguished from older RNAs. If the regulation is indeed at the level of transcriptional elongation, and not by a post-transcriptional mechanism, then I expect some level of influence of chromatin in this regulation. Identifying computationally any enriched sequence motifs in nearby nucleosome-free regions would be a first step. Using public datasets for other histone modifications (such as data of the modencode consortium, http://www.modencode.org), histone variants, chromatin remodellers, nucleosome-turnover, and insulators would further refine the chromatin state governing the difference between regulation by Pol II recruitment and regulation by Pol II elongation. Further assessment of gene features within each regulation class may help to further understand the mechanism, such as whether there are different numbers of splice sites or intron lengths, or if the genes fall into possible co-regulatory domains.

How the regulation of Pol II recruitment is orchestrated at specific gene types is also an interesting question. Specifically, the involvement of su(Hw) to repress a subset of neuron-specific genes outside of neurons is a highly interesting avenue to pursue. Two relatively straight-forward experiments would be to A: force expression of su(Hw) within neurons, and B: deplete su(Hw) in glia cells and other, non-neuronal, tissues. The read-out of these analyses would be RNA-seq of FANS-isolated neurons and glia (or other tissue). The hypothetical outcome of over-expressing su(Hw) in neurons would be that the neuron-enriched genes would have reduced expression, possibly to equivalent levels as the normal expression of these genes in glia. This would be the case if we assume that su(Hw) alone is sufficient to repress the neuron-enriched genes to the level of expression found in glia. If other factors also play a role in providing additional repression, outside of neurons, then over-expression of su(Hw) in neurons would not have a very big effect on the neuronal gene expression. The predicted outcome of reducing Su(Hw) expression in glia, for example by using an RNAi line, would be that the neuron-enriched genes would have increased expression in the glia. This outcome

again relies on the assumption that su(Hw) alone is sufficient to repress the gene expression. If there are many other factors involved, depletion of su(Hw) alone may have little effect on the gene expression, as the genes would be repressed by an accumulative effect of several proteins.

The chromatin state of active genes with no H3K36me3 may or may not play a role in helping to activate the Su(Hw)-regulated neuronal genes. However, the presence of such a chromatin state alone is in itself an intriguing observation. The primary question is how H3K36me3 is absent from these active genes. There needs to be some removal mechanism that leads to no detectable H3K36me3, and that is specifically targeted to particular genes. I can conceive of two possibilities for achieving an H3K36me3-less active state. One that these genes have H3K36me3 deposited during transcription but the methyl mark is rapidly removed by a histone de-methylase, likely KDM4a. Another possibility is that the H3K36me3 is not placed on these genes at all, meaning that the recruitment or activity of the methyltransferase is inhibited at these genes. There is some evidence against the first mechanism. Since experiments using Kdm4a mutants in *Drosophila* showed that many genes affected by a Kdm4a knock-down have no H3K36me3, even in the absence of the demethylase (162). This means that the genes modulated by Kdm4A are regulated by a mechanism other than removal of H3K36me3. One thing to note is that only 99 are affected by the Kdm4a mutant, here I found that there were over 500 genes that had the active H3K36me3-less signature. Therefore, assessing the affect of Kdm4a knockdown in neurons and glia may yield different results than seen in the study by Crona *et.al* (162).

Assessing the second possibility, that H3K36me3 is not placed at all, would be more difficult. A first experiment would be to ask whether the Ser2-phosphorylated form of the elongating polymerase was present at these genes. Since Set2 associates with this modified form of the elongating polymerase, if there is lack of this Pol II modification, then the downstream methyl-transferases would not be recruited. Another experiment would be to test the effect of a Set2 knockdown at these genes. In principal, lack of Set2 should not affect the correct expression of the H3K36me3-less genes if the methyltransferase is not active at them.

It would be very interesting to assess what effect having no H3K36me3 has on the genes. Lack of H3K36me3 certainly leads to spreading of H3K27ac, and likely H3K18ac, thus these regions would be expected to become "hyper" activated and open. The explanation that this chromatin state would help genes that are in a repressive context

to overcome the repression is certainly not the only possibility. For instance, does lack of H3K36me3, and increased histone acetylation, lead to higher cryptic transcription or affect splicing? If so, this chromatin state could be a mechanism to allow internal transcription from within the gene. Such internal transcription might be necessary to generate novel transcripts of the gene, or to allow transcription of regulatory RNAs, for example. Such questions could be addressed both computationally and experimentally. From the existing nucRNA data, which is directional, we could assess the level of internal antisense transcription at these genes. Determining internal sense transcription would be difficult since it would not be simple to define internal transcriptional start sites from normal transcription. Here, an experimental approach that captures only the very 5′ end of the transcript would be useful, such as CAGE (cap analysis of gene expression (163)). This analysis would determine if there is indeed more internal transcription initiation at the H3K36me3-less genes. If such a phenomenon as increased cryptic transcription exists at these genes, then asking whether this is an essential factor of these genes would be the next inquiry. Depending on the mechanism of generating the chromatin state, I could conceive several experiments. Knock-down of Kdm4a is one, but another is tethering Set2 to these genes, through an inactive Cas9 for example (164). Tethering Set2 to the 3′ end of specific genes could force H3K36me3 at these genes, and the consequence of this forced H3K36me3 on gene expression and the phenotype of the fly could be investigated.

Chapter 5

The role of promoter architecture in defining gene expression programs

5.1 Summary

In this chapter, I focus on a small subset of genes that represent the most extreme cases of nucleosome architecture around the promoter. These two gene sets, with either ordered or fuzzy promoter types, have vast differences in Pol II binding profiles and incorporation of H2A.Z. Interestingly, around half of the genes with extremely fuzzy nucleosome architecture have high levels of very precisely bound Pol II immediately downstream of the TSS. Conversely, the majority of genes with the most highly ordered nucleosomes show broad Pol II binding and highly variable gene expression levels. The underlying DNA sequences of these two promoter types is highly different and is likely a major contributor in establishing the differential chromatin states. I suggest a mechanism whereby the fuzzy genes have highly dynamic and competitive binding between nucleosomes and RNA Pol II, which inhibits incorporation of H2A.Z and facilitates extremely rigorous gene regulation.

5.2 Introduction

Promoter architecture plays a major role in defining how transcriptional machinery is recruited, as well as the pausing status and the elongation rate of a gene. Nucleosomes are refractory to gene expression, by hindering both the progression of Pol II through a gene and recruitment of Pol II to the promoter DNA. However, there is little evidence

that supports the notion that nucleosome positioning has an impact on gene expression. As demonstrated from the previous chapter, there are cell-type-specific genes with constitutively open promoters, whose chromatin architecture is invariant between cell types, but there are also cell-type-specific genes where the chromatin architecture is altered with changes in gene activity.

It is generally viewed that transcriptional activity and the presence of H2A.Z are linked to an open and uniform promoter architecture (41). The suggested consequence of this relationship is that when genes transition from an inactive to active state, the promoters would become cleared, incorporate H2A.Z, and adapt a more ordered architecture. Yet, this scenario is not supported when looking at specific cell types of adult tissues, as it was found that genes with cell-type-specific Pol II binding do not incorporate H2A.Z at their promoters (see (34) and chapter 4). H2A.Z was instead found to be highly similar across all cell types tested, and correlated with tissue-invariant genes. As discussed in the previous chapter (4), those cell-type-specific genes identified by Pol II binding have a different chromatin state than invariant genes: they are active, with no H3K36me3, and fuzzy nucleosome architectures. Pol II pausing could be a major regulatory mechanism governing cell-type-specific genes, as it negatively correlates with H2A.Z (18). Pol II pausing is a specialised gene regulatory mechanism that is important for both spatial and temporal gene expression (see 1.1.1). Regulation by Pol II pausing correlates with a distinct nucleosome architecture at the promoter. At highly paused genes, the nucleosome-depleted region (NDR) is occupied until the pre-initiation complex is recruited to the promoter, and the nucleosome-array is less ordered than non-paused genes (165). However the cell-type-specific genes showed little clearing of nucleosomes across the promoter, even in the cell type where Pol II is bound. Thus, the current model we have of Pol II pausing is likely to be more complicated than we thought.

The majority of studies have focused on first characterising a particular expression characteristic of the gene, such as Pol II pausing level, or (as I have done in the previous chapter 4) the cell-type-specificity of the genes expression pattern. Then the chromatin state of each gene class is assessed. In this chapter, I aimed instead to investigate chromatin function from a nucleosome-centric starting point. To aim to first identify genes with a particular nucleosome organisation across the promoter, and then look at the expression, histone modifications, and underlying sequences of those genes. This approach will ideally allow me to assess what may drive the two different types of gene

regulation– open constitutive promoters versus closed regulated promoter– without bias to expression levels or cell-type specificity.

5.2.1 Aims

- Define gene classes based on nucleosome architecture around the promoter.
- Assess activity status of ordered versus fuzzy gene classes.
- Identify underlying sequence preferences of ordered versus fuzzy gene classes.
- Compare the chromatin states of ordered versus fuzzy genes.
- Assess the pausing status of fuzzy verus ordered genes.

5.3 Pol II binding correlates with NDR formation

Once striking observation from the analysis in chapter 4 (section 4.4) was that the cell-type-specific genes based on Pol II binding had fuzzy promoter architecture compared to the invariant genes. The level of Pol II binding at the gene did not appear to have much impact on NDR formation (upper panels figure 4.3). I had expected that the cell type in which the Pol II was bound would have had a much reduced level of nucleosomes upstream of the TSS, compared to the cell type where there was no Pol II binding. Also, the +1 nucleosome was expected to be more highly ordered in the cell type where Pol II was bound. The opposite was observed for those genes identified as neuron-enriched by the nucRNA analysis (lower panels figure 4.3). Here, the nucleosomes formed highly ordered arrays across the gene body, and had a deep NDR in both neurons and glia, despite the genes being far more highly expressed in neurons.

How much does Pol II binding or RNA expression level influence the organisation of nucleosomes at the promoters? If there is more transcription machinery bound to the gene then there must be reduced nucleosome occupancy, since it is assumed that nucleosomes and Pol II would not be able to bind to the same piece of DNA at the same time. However, the view from this genomic data is only a snapshot, and does not capture the dynamics of the system. If some genes had highly dynamic transitions between a Pol II-bound state and a nucleosome-bound state, then both high Pol II and high nucleosome occupancy would be observed at the same regions. Why some genes would be regulated by a more static chromatin state, and some by a more dynamic chromatin state is an interesting question.

5. THE ROLE OF PROMOTER ARCHITECTURE IN DEFINING GENE EXPRESSION PROGRAMS

Before further analysis, I wanted to ensure that the observations made about the nucleosome architecture at cell-type-specific genes were not just the result of making average profiles of multiple genes. I wanted to look at both the influence of Pol II binding and nucRNA level on nucleosome architecture at the promoter, since they appear to have different effects. First, I examined the influence of Pol II binding. I generated a ranked list of genes by calculating the number of reads from the Pol II ChIP across a 500 bp window around the TSS, ranked from highest to lowest amount of binding (methods 6.5). I then used these ranked gene lists to generate ranked heatmaps of the MNase-seq data for neurons and glia (left side, figure 5.1). Control heatmaps of the Pol II binding data (right side, figure 5.1) show the ranking of the Pol II was performed correctly and that it is useful to compare the binding levels of Pol II with the MNase heatmap .

It can be clearly seen from these heatmaps that even at the highest level of Pol II binding there was little NDR formation at the neuron-enriched genes (neuronal genes, neuronal MNase, figure 5.1). For the glia-enriched genes, there appears to be some ordered nucleosome array clusters across a range of Pol II binding levels. However, this nucleosome arrangement is present in both the neuron and glia MNase-seq data, whereas the genes only had Pol II binding in glia. Those genes with invariant Pol II binding show highly ordered nucleosome arrays and large NDRs for all genes that show some Pol II binding. Those genes with no Pol II enrichment do not show any ordered nucleosome organisation nor do they ahve an NDR. Thus it appears that at invariant genes, the level of Pol II binding does correlate with a clearance of the nucleosomes at the promoter.

I wanted to further demonstrate that the effects of Pol II binding levels on nucleosome architecture are different between cell-type-specific and invariant genes. Therefore, I split each gene group into three categories, based on the level of Pol II binding: high, medium and low binding. For invariant genes, I generated a fourth group of very high binding as the Pol II binding is far higher in this group than any Pol II binding in the cell-type-specific groups. I then generated average profile plots of the MNase-seq data, split into the different gene catergories (figure 5.2). These results clearly demonstrate a difference in nucleosome organisation between invariant and cell-type-specific genes. Neuron-enriched genes presented a very fuzzy nucleosome architecture at the promoter, in both neuron and glia MNase-seq data. However, the neuron-enriched genes with the highest levels of Pol II binding had a decrease in nucleosome occupancy in neuronal MNase compared to glia MNase. The glia-enriched genes showed the same

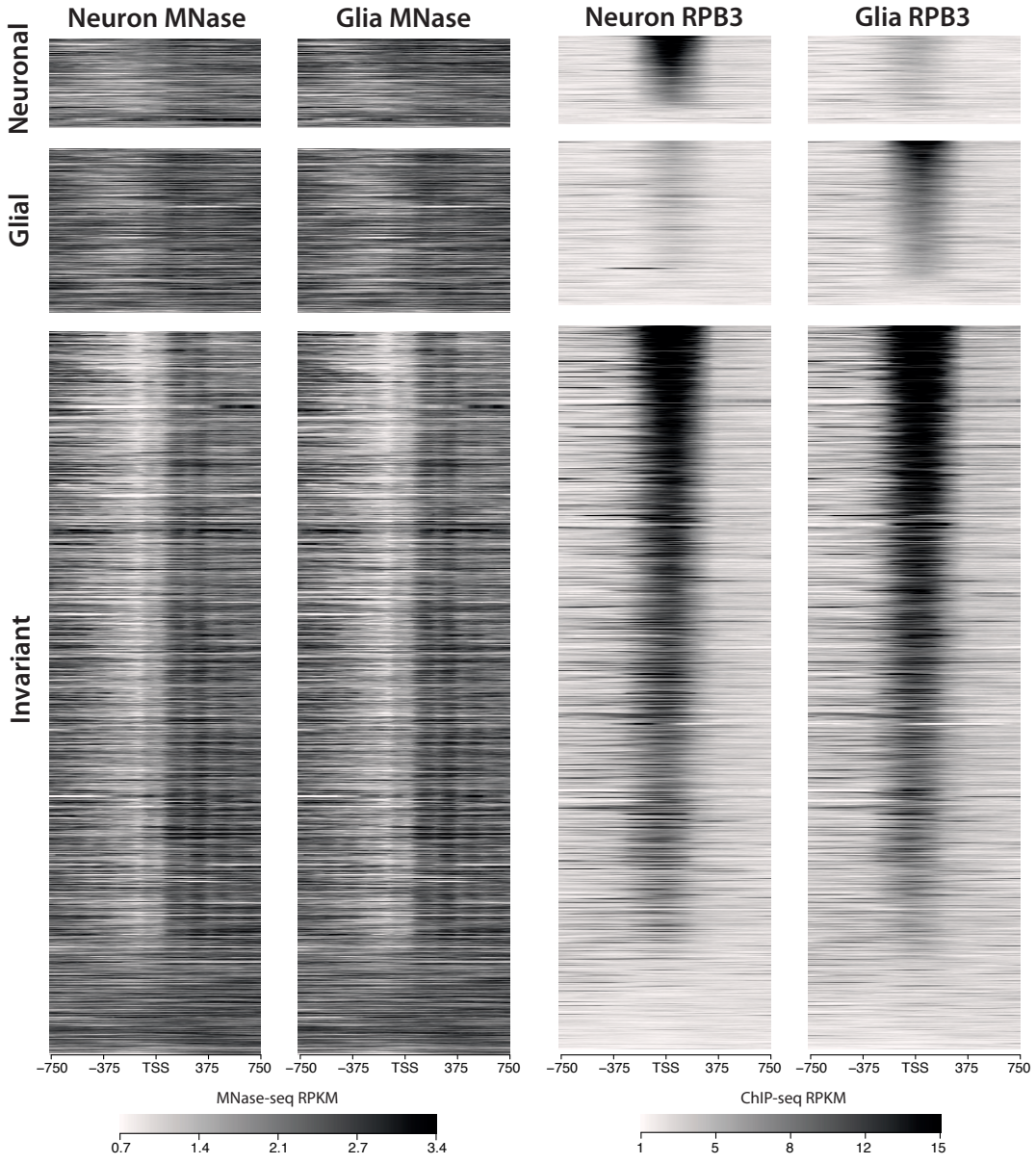


Figure 5.1: Pol II binding across TSS - Heatmap of MNase-seq and RPB3 (Pol II) ChIP-seq reads across cell-type specific and invariant genes, defined by Pol II binding. Genes ranked in descending order of read counts across a 500 bp window around the TSS. Pol II levels from neurona RPB3 CAST-ChIP were used to rank the neuron-enriched genes, glia RPB3 CAST-ChIP data was used to rank the glia-enriched genes, and whole-head RPB3-ChIP data was used to rank the invariant genes.

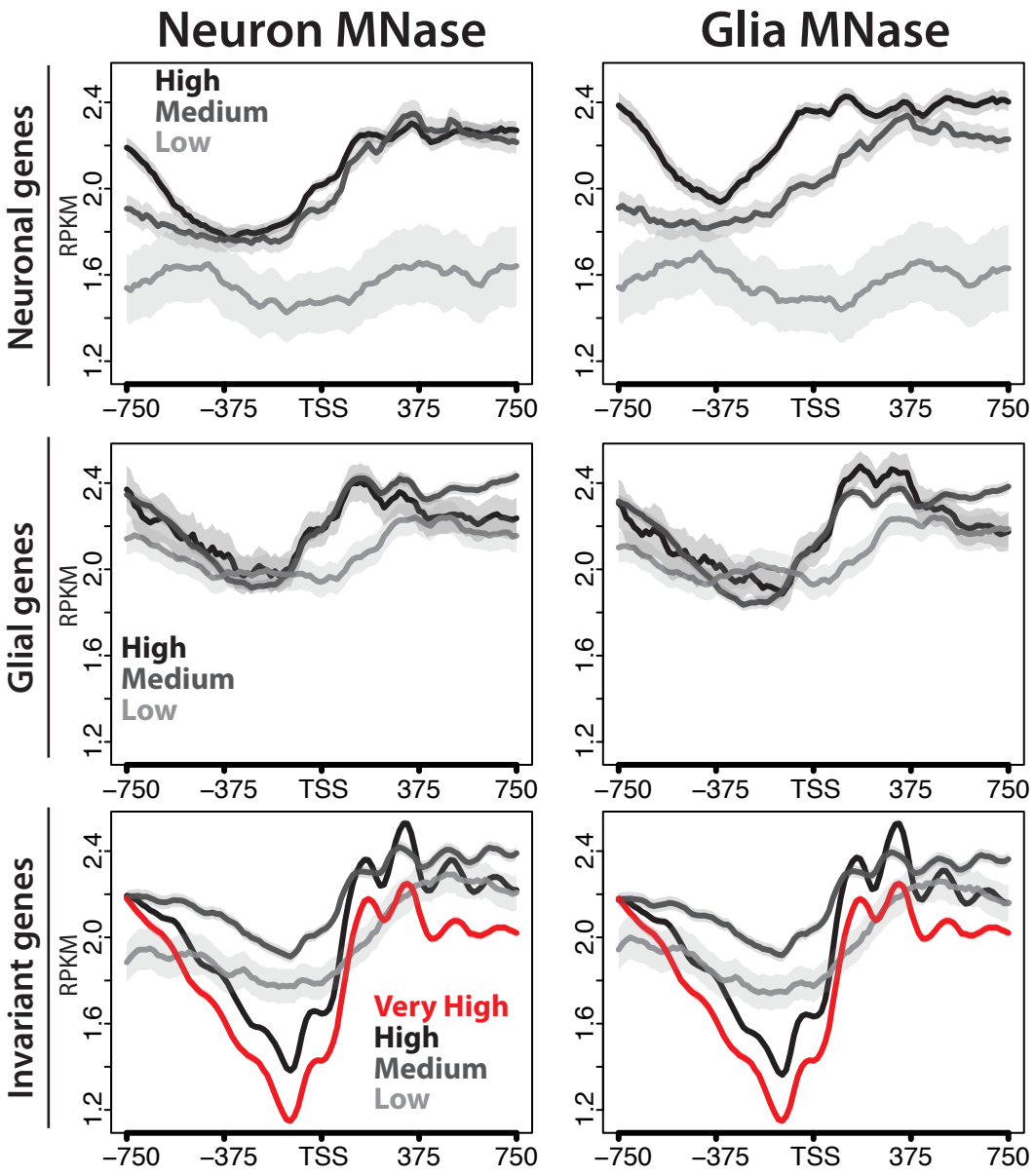


Figure 5.2: Nucleosome architecture around the promoter based on Pol II binding - Average profile plots of the neuron or glia MNase-seq data. Genes from the neuron-enriched, glia-enriched, or invariant classes were further classified as having very high (invariant only), high, med, or low Pol II counts over a 500 bp window across the TSS.

general pattern as neuron-enriched genes, although the +1 and +2 nucleosomes were more defined, and again the promoter region had a high density of nucleosomes that was slightly lower in the glia cells than the neurons. Invariant genes showed a completely different pattern than the cell-type-specific genes. The higher the Pol II binding, the more well positioned the +1 nucleosome became and the lower the density of nucleosomes over the NDR. It is clear from these analysis that there are different arrangments of nucleosome with increased Pol II binding at invariant genes compared to cell-type-specific genes.

5.4 RNA expression and nucleosome positioning are not correlated

As demonstrated in the previous chapter (chapter 4, section 4.4), there was no correlation between the level of nucRNA expression and nucleosome positioning at gene promoters. Again, to demonstrate this further, I generated heatmaps where genes were ranked from the highest level of expression to the lowest for each gene set (figure 5.3). In these heatmaps, it was very clear that the level of RNA expression had little correlationwith the level of nucleosome occupancy at the promoters of neuron-enriched genes. The nucleosome array of these genes is highly ordered in both the neuron and glia, and the level of RNA expression appears to not correlate with how ordered the nucleosomes are at these genes. The glia-enriched genes were more difficult to interpret, since the RNA levels were generally depleted from the neuron nucRNA to levels lower than the surrounding regions. Very few of the glia-enriched genes had very high expression, and most appeared to have levels comparable to background (levels outside of the scaled gene regions) in glia cells. The nucleosome architecture was completely disordered at the glia-enriched genes. Those genes with the highest levels of expression had lower nucleosome occupancy, but there was little formation of an NDR or ordered nucleosome array. Much like with the neuron-enriched genes, the invariant genes had highly ordered nucleosome arrays and deep NDRs. Also, at the invariant genes the level of RNA expression had little influence on the order of the nucleosome architecture, but does appear to influence the level of nucleosome occupancy into the gene body. The highest expressed genes have less nucleosome density into the gene, but the +1 nucleosome position and NDR remain the same at all RNA expression levels.

The observations seen in the heatmaps of figure 5.3 are more easily interpreted as average profile plots, splitting the genes into high, medium and low expression (figure

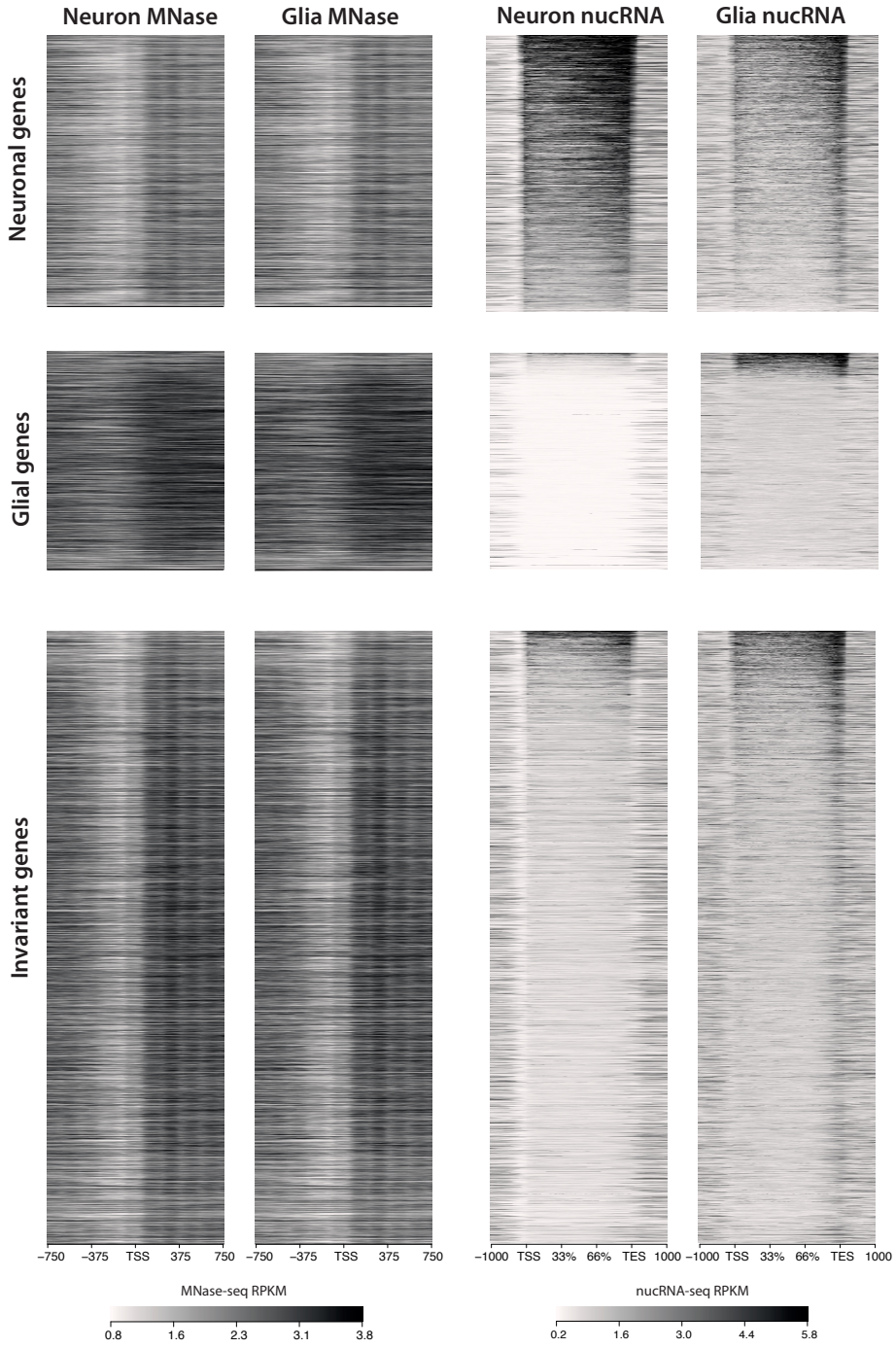


Figure 5.3: Ranking genes based on nucRNA expression - Heatmaps of MNase-seq and nucRNA-seq in neurons and glia, at cell-type-specific and invariant genes based on nucRNA expression level. Neuronal genes were ranked based on neuronal nucRNA-seq expression level, glia genes were ranked based on glial nucRNA-seq expression level, and invaraiant genes were ranked based on the average expression level of the nucRNA-seq input samples.

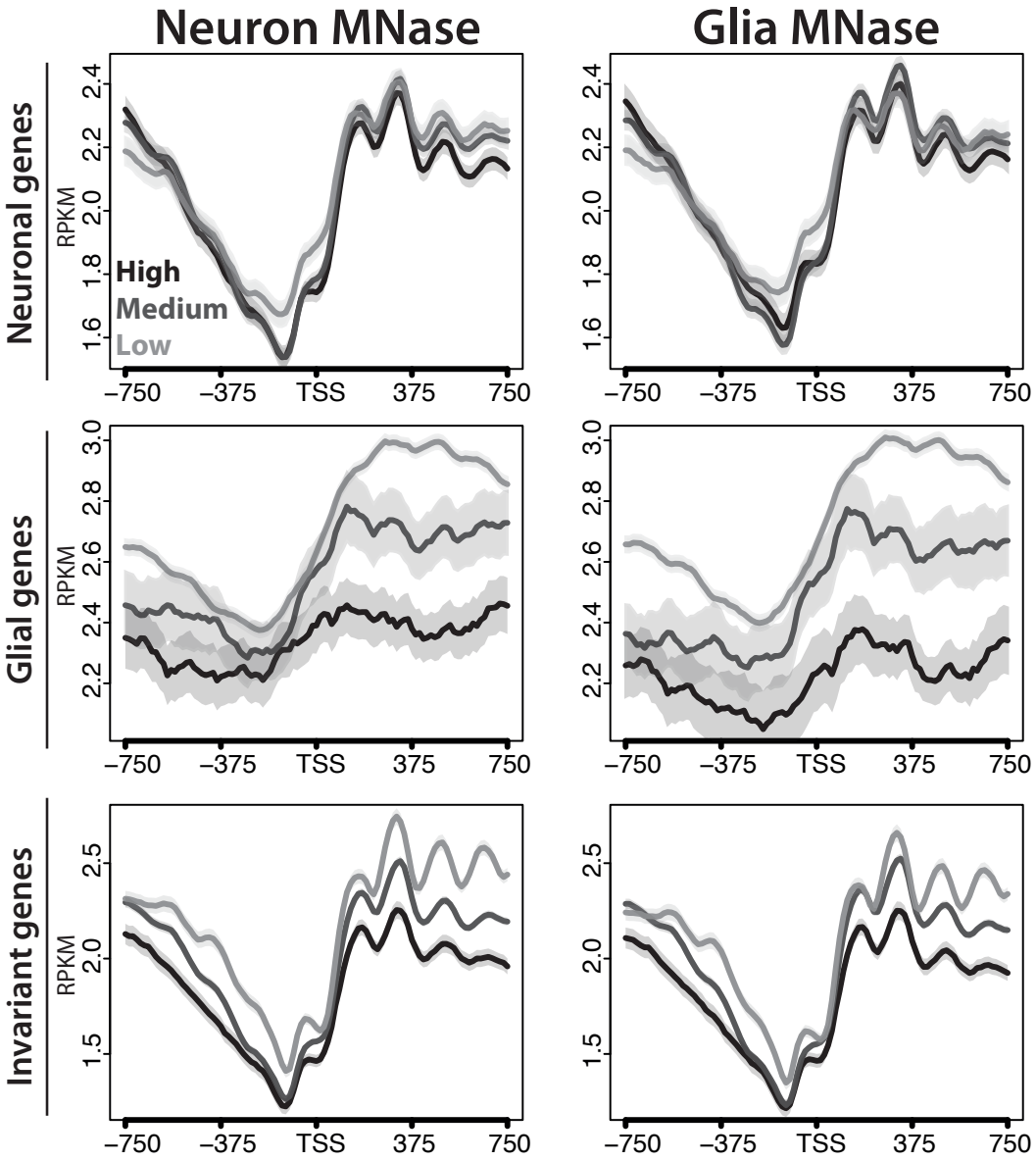


Figure 5.4: Nucleosome architecture based on expression level - Average profile plots of MNase-seq data from neurons or glia, split into high, medium, or low expression levels based on the nucRNA-seq data.

5.4). Here, the trends described from the heatmap were more clearly visible. The nucleosome architecture at neuronal genes was barely influenced by the expression level of those genes. The highest expression group showed nearly identical expression to the lowest expression group (figure 5.4, upper panels.) As seen in the heatmap, the nucleosome density into the gene body of invariant genes decreased with higher RNA levels. However, the NDR of the invariant genes remained depleted to the same degree regardless of RNA expression level (figure 5.4, lower panels). Interestingly, the glia-enriched genes have the most changes in nucleosome architecture across the different expression level groups (figure 5.4, middle panels). Here, the highest level of gene expression had a very flat nucleosome pattern, with the "NDR" region having nearly equal nucleosome density levels as the gene body. The lowest expressed genes have a larger difference between the nucleosome density across the promoter region and the gene body. Please note the difference in scale between the different gene classes. For example, for the glia-enriched genes, the lowest point on the scale bar (log2(RPKM) = 2.2) is equivalent to the second highest point on the neuronal and invariant gene graphs. Thus, the glia-enriched genes have very high nucleosome density across the entire region, compared to the neuronal and invariant genes.

5.5 Defining promoters based on nucleosome architecture

From the previous analysis (sections 5.3, and 5.4), there are clearly differences in how the nucleosomes are organised at the promoters of cell-type-specific and invariant genes. However, it is apparent that not all genes regulated cell-type-specifically have the same promoter architecture, and not all invariant genes have a constitutively open promoter architecture. It would be interesting to better understand the characteristics that underly the "fuzzy" versus "ordered" types of promoter and how they are involved in gene regulation. To achieve this, I rationalised that it would be better to define gene sets solely based on the promoter architecture, rather than a pre-defined expression pattern. With a defined set of genes whose defining characteristic was the occupancy and positioning of nucleosomes around the TSS, I could then observe what types of characteristics, i.e. expression and chromatin state, are associated with each promoter type. To define the two groups of genes, "ordered" or "fuzzy", a score of orderedness and fuzziness was calculated for every gene promoter (figure 5.5 A), using the head MNase-seq data set. This was calculated by Pawel Bednardz. To determine the "orderedness" of each gene, the ratio between the read count at the nucleosome summit and the neighbouring trough (linker) was calculated for the first four nucleosomes of the gene.

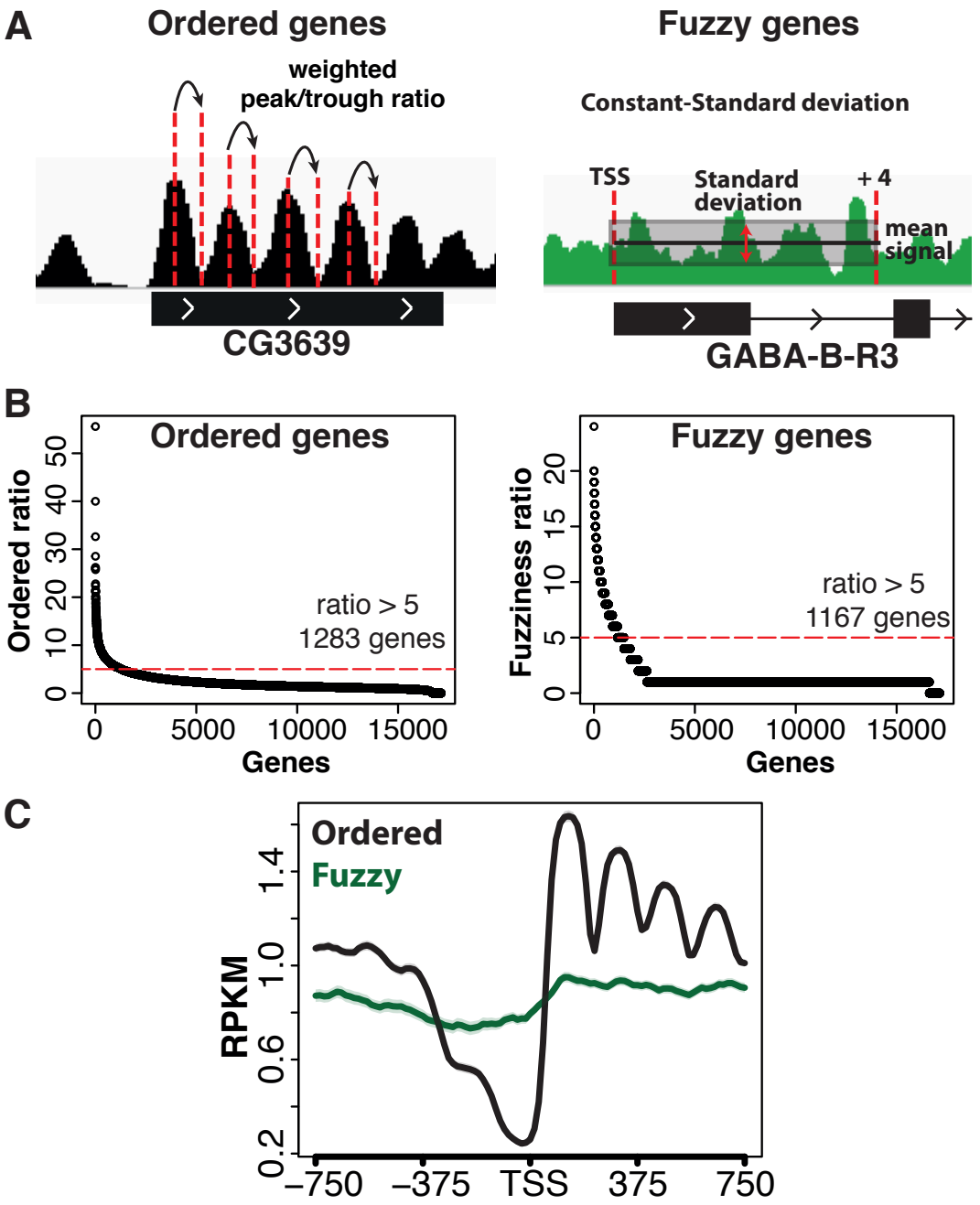


Figure 5.5: Dividing genes based on promoter nucleosome architecture - A) The orderedness and fuzziness ratios for each gene promoter was calculated. B) The most extreme genes, those with ratios over 5, were selected for further analysis. C) Metagene analysis showing that the classification into ordered and fuzzy genes based on the mathematical calculations were correct. RPKM is defined as reads per million mapped reads.

Then, the weighted average of the ratios was calculated, with higher weight given to the +1 nucleosome, then the +2 etc. The fuzziness score was calculated by taking a constant number minus standard deviation of a signal from TSS to where the +4 nucleosome was expected to be. To select genes with the most extreme ordered or fuzzy architectures for further analysis, I took all genes with a score over 5, with 1283 and 1167 genes for ordered and fuzzy categories, respectively. The majority of genes fall below the threshold score of 5 (figure 5.5 B). I generated an average profile plot, graphing the MNase-seq data across the two gene categories (figure 5.5 C). This shows that the selection was appropriate for identifying ordered and fuzzy genes. The ordered genes showed the expected features, such as a nucleosome-depleted region upstream of the TSS, a highly positioned +1 nucleosome, and an ordered nucleosome array into the body of the gene. The fuzzy genes in this selection had none of the features; there was a very slight depletion of read coverage upstream of the TSS, relative to the surround regions, and there was no defined +1 nucleosome or nucleosome array.

5.6 Removing low coverage genes from the analysis

As can be seen in Figure 5.5 (C), the average coverage of MNase-seq reads over the gene body is generally lower in the fuzzy gene class compared to the ordered gene class. It is difficult to determine how well positioned a nucleosome is over a genomic locus when the coverage of MNase-seq reads is very low. This is because if there is not enough coverage there will be little difference in read depth between where the nucleosome would be "positioned" and the adjacent linker region. Therefore those genes with very low coverage needed to be removed from further analysis. To accomplish this, I determined the number of reads per base pair for every gene (methods 6.5.1) and divided the genes into quintiles based on this read depth (figure 5.6 A). The lowest quintile (quintile 5) has many genes falling below one read per base pair, a coverage depth that is not reliable for determining nucleosome positioning. Only genes that fall into the top four quintiles were used for further analysis, this resulted in a large loss of genes within the fuzzy category (281 remaining from 1167) and also a loss of genes from the ordered category (796 remaining from 1283).

When each quintile was plotted as an average profile plot aligned to the TSS (Figure 5.6 B), the level of coverage didi not affect the depth of the NDR nor did it seem to affect the average position of the +1 nucleosome. However, the higher the coverage, the more defined the nucleosome arrays were (compare quintiles 1 (black) and 5 (red) in figure 5.6 B, for instance). Inspecting those genes that were removed from the analysis,

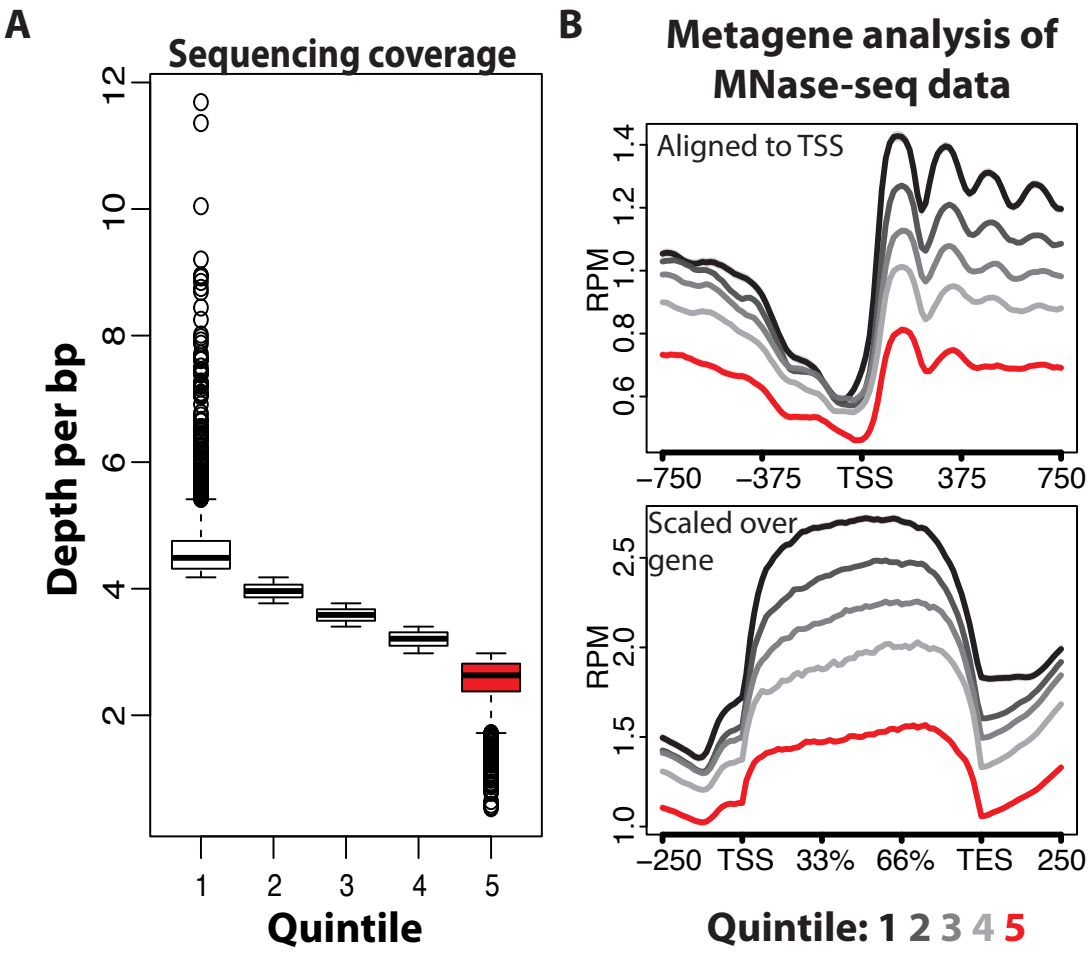


Figure 5.6: Removing low coverage genes from analysis - A) Boxplots showing the sequence coverage of the MNase-seq reads, in depth per base pair, across the read-depth quintiles. B) Average profile plots of the MNase-seq data across the coverage quintiles. In the upper panel, the data were aligned to the TSS and show the distinctive +1 nucleosome, and nucleosome-array structure. In the lower panel, the average MNase-seq coverage is scaled across the gene body of the five gene groups.

I found that they have very high expression levels (not shown). This makes sense in that these genes are likely so highly expressed that they have depletion of nucleosomes across the gene body, due to constant progression of Pol II across the gene.

5.7 Activity characteristics of ordered and fuzzy genes

To determine the overall activity characteristics of ordered and fuzzy gene classes, I plotted the read counts of RPB3-ChIP across the TSS as well as the average RPKM from whole head nucRNA-seq of each gene group (figure 5.7, B and C). I also generated heatmaps for the MNase-seq, RPB3 (Pol II) and nucRNA data, ranked using the heirarchical clustering function (158), to visualise the Pol II binding or nucRNA levels across the different gene sets (figure 5.7, D,E, and F). The boxplots for RPB3 (figure 5.7, B) show that Pol II bound to generally lower levels in the fuzzy gene group compared to ordered genes. Interestingly, the pattern of the heatmap for RPB3 was very different between the ordered and fuzzy gene classes (figure 5.7, E). The ordered gene class had a very dispersed average RPB3 binding profile, whereas the fuzzy gene class had a highly concise RPB3 profile spanning only approximately 200 base pairs, with the highest point around 50 base pairs after the TSS. The broad binding pattern observed in the ordered gene average profile appears to be produced by broad binding of Pol II across the same gene promoters, rather than precise Pol II binding at different points from the TSS for each gene.

The heatmap of the fuzzy gene class showed two clearly defined sub-populations: genes with high levels of precisely positioned Pol II binding, and genes with no Pol II present. The boxplot for nucRNA reads at the two gene classes (figure 5.7, C) showed that there was little difference in gene expression, on average, between the two gene groups. The heatmap of the nucRNA profile showed a complementary pattern to the RPB3 heatmap (figure 5.7, D). Ordered genes had a low read density across the NDR, and had relatively equal read levels between the TSS and further into the gene. Fuzzy genes, on the other hand, had a higher read density just after the TSS compared to further down stream regions. In conclusion, the Pol II binding data and nucRNA data show that those genes with ordered promoters have broad Pol II binding and highly variable levels of gene expression. Conversely, genes with fuzzy promoters are enriched for two different types of characterisitic; genes with sharp Pol II recruitment and high levels of gene expression, or genes with no Pol II recruitment and low gene expression. These observations are consistent with previous findings that invariant genes have broad Pol II binding, and cell-type-specific genes have highly specific Pol II binding sites (166).

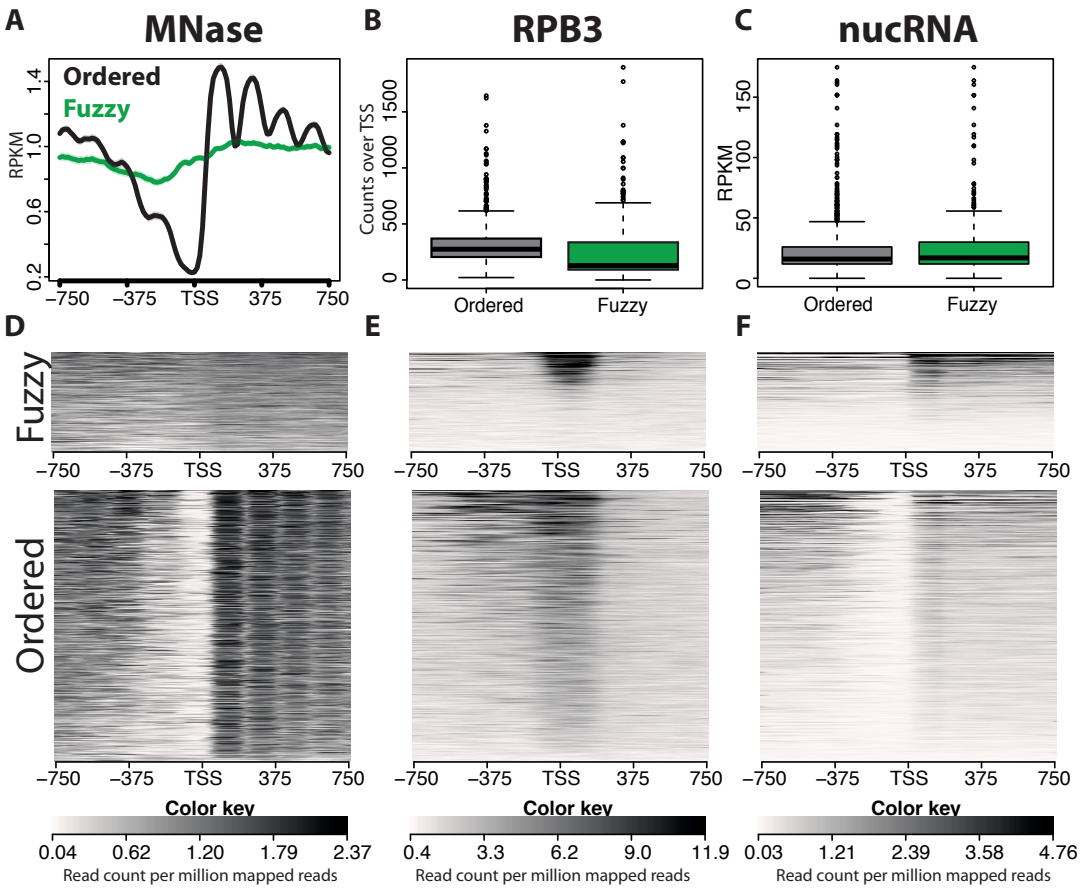


Figure 5.7: Activity characteristics of ordered and fuzzy genes - A) Average profile plots of MNase-seq data of ordered and fuzzy genes. B) Boxplots showing the levels of RPB3 binding across a the TSS ± 250 bp, for the ordered and fuzzy gene groups. C) Boxplots showing the distribution of RPKM levels from the input nucRNA analysis for the ordered and fuzzy gene classes. Heatmaps produced from NGS-plot showing the MNase-seq data (D), RPB3 ChIP-seq data (E) and input nucRNA data (F) for the ordered and fuzzy gene groups.

5.8 Defining promoter classes based on Pol II binding

There are clearly two distinct populations of genes within the fuzzy gene class, as shown by the analysis of Pol II and nucRNA (section 5.7). One group of genes had highly sharp Pol II recruitment, and were expressed on the nucRNA level. The other group of genes had neither Pol II recruitment, nor nucRNA expression. These two populations likely have different features in terms of chromatin signature and promoter architecture. As demonstrated in sections 5.3 and 5.4, the level of Pol II binding was much more correlated with changes in the surounding chromatin architecture than RNA levels. Therefore, I separated active genes from inactive genes, based on Pol II occupancy, in both the ordered and the fuzzy gene lists. This was achieved by intersecting each gene list with the the list of Pol II peak calls for the head RPB3 ChIP-seq (CCAT peak calls, (34)), resulting in four distinct gene classes: Ordered + Pol II, Ordered No Pol II, Fuzzy + Pol II, and Fuzzy No Pol II (table 5.1).

	Pol II	**No Pol II**
Ordered	402	394
Fuzzy	110	179

Table 5.1: Defining promoter classes by Pol II binding

The average Pol II profile at these four classes of genes is shown in figure 5.8 (A). The two populations previously observed for the fuzzy genes have been clearly separated, with a relatively even number of genes having either high levels of precisely localised Pol II binding or no detectable Pol II binding. There was not such a clear distinction between being Pol II bound and not Pol II bound at the ordered genes. For example, the ordered genes with no Pol II peak had an overall higher level of reads compared to the fuzzy genes with no Pol II (peak (figure 5.8 A, grey for for ordered Pol II, light green for fuzzy Pol II). The higher Pol II binding for ordered genes did not appear to result in higher transcription, as the nuclear RNA levels of the No Pol II gene classes were at the same level (figure 5.8 C).

The higher level of Pol II binding in the **ordered no Pol II** gene category compared to the **fuzzy no Pol II** category was not due to technical differences in how accessible the chromatin was to sonication. Appendix C.4 shows that the input chromatin is at the same background level in all gene categories. It is likely that the low level of read enrichment came from bound Pol II, which was just not called as Pol II-bound in the peak-calling analysis, simply because of the technical limitations of this type of

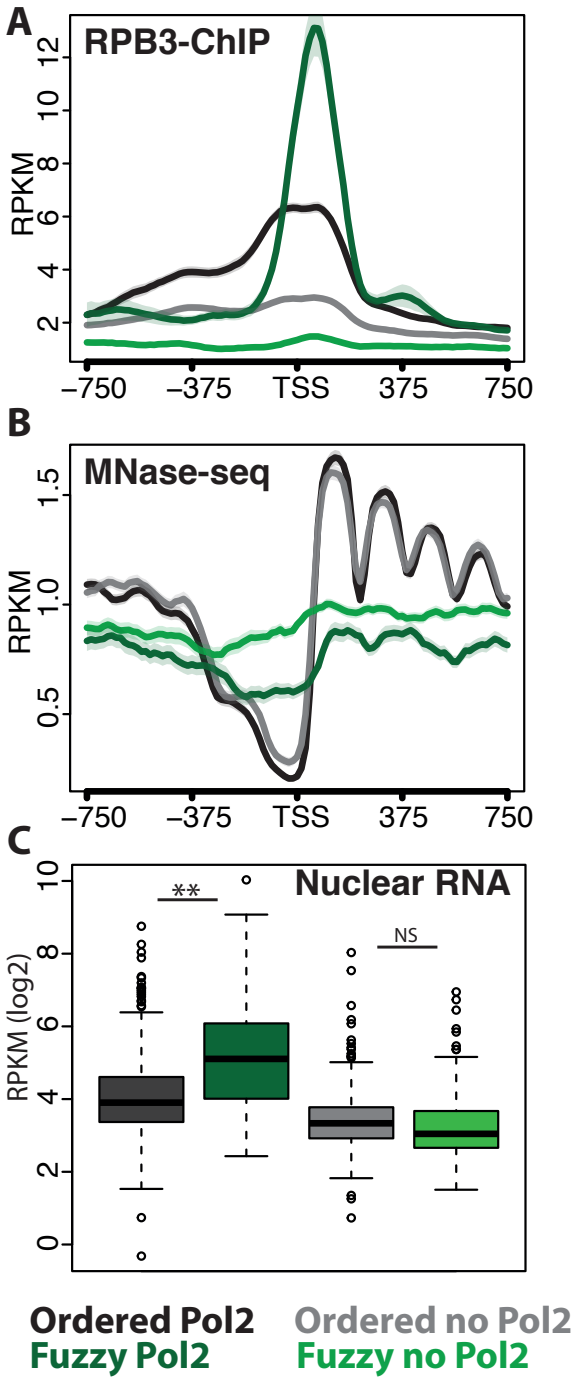


Figure 5.8: Defining different architecture groups based on Pol II binding - A) Average profile plots of RPB3 binding across the TSS of the different gene groups. B)Average profile plots of the MNase-seq data, aligned to the TSS. C) RPKM levels of the four gene classes, using the average RPKM of the head (input) nucRNA data.

analysis. Peak calling from genomics data requires a threshold to be reached above the background noise, and sharper peaks are easier to call than broadly dispersed peaks with blurry edges. Because of the nature of Pol II binding at ordered genes, which is rather lower and much broader than fuzzy genes, the Pol II peaks at ordered genes would be more difficult to call. Despite this, however, I retained the same classifications for further analysis, given that the Pol II binding in the **ordered no Pol II** class is not called as significant in the peak calling and the nucRNA levels are equally as low as the **fuzzy no Pol II** class.

Despite the differences in Pol II binding and nucRNA levels for the ordered genes, the nucleosome pattern at these promoters was nearly identical (figure 5.8, B). However, the NDR of the **ordered no Pol II** genes had a slightly higher read coverage than that of the **ordered Pol II** genes. The nucleosome pattern of fuzzy genes changed when the genes are split into the ± Pol II categories. The **fuzzy no Pol II** genes retained a completely flat nucleosome architecture, whereas the **fuzzy Pol II** genes had a reduction of reads over the promoter, and a slight positioning of the +1 and +2 nucleosome was apparent. The NDR present for the **fuzzy Pol II** genes was still not to the level of depletion seen in both classes of ordered genes. This means that despite the high level of Pol II binding at the **fuzzy Pol II** genes, the Pol II was not completely out-competing nucleosomes for binding to the promoter region.

5.9 Underlying sequence preferences of promoter classes

One explanation of why there is little difference in promoter architecture at ordered genes, regardless of activity level, is that the DNA sequence at these promoters is driving nucleosome eviction rather than chromatin remodellers or the pre-initiation complex (PIC). A long stretch of AT (or CG) nucleotides is unfavourable to nucleosome formation as it increases the rigidity of the DNA, which makes wrapping it around a nucleosome thermodynamically unfavourable. Therefore, those genes with an ordered promoter architecture and deep NDR across the promoter would be expected to be highly enriched for AT. Conversely, those genes with fuzzy promoters, where nucleosomes occupy the promoter region, would be expected to have mixed sequences that would favour nucleosome formation.

To determine if there are any differences in the composition of nucleotides between the ordered and fuzzy gene classes, I used the genomic coordinates of each annotated TSS to generate a fasta file of sequences 500 bp either side of the TSS (methods 6.5.2).

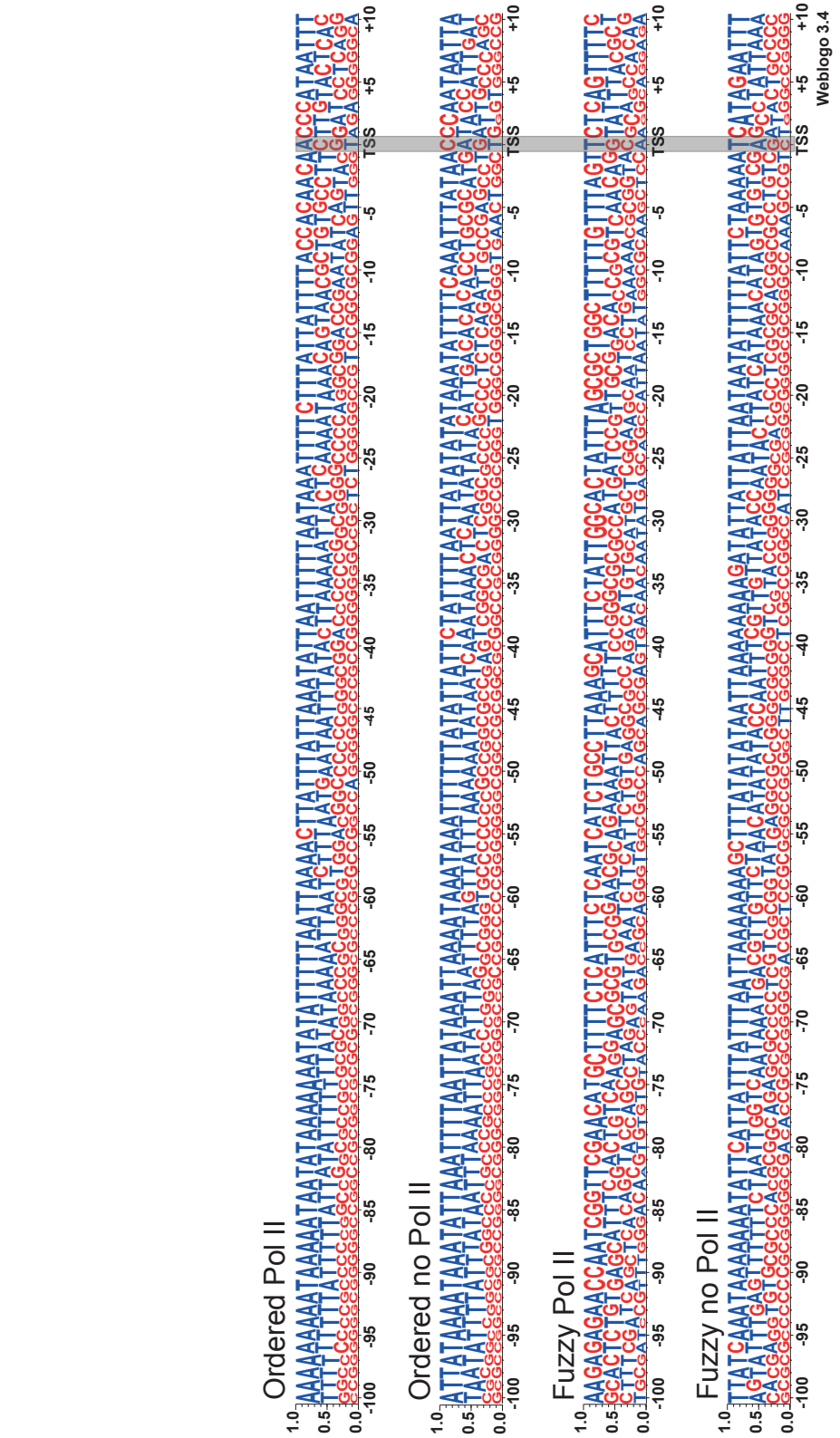


Figure 5.9: Underlying sequence preferences of different promoter classes -The regions directly upstream of the TSS were highly AT-rich for ordered genes, indicating that the DNA sequence may play a role in intrinsically excluding nucleosomes from these promoters. Those genes with fuzzy promoters had more randomly distributed nucleotide sequences upstream of the TSS; such a nucleotide composition would be favourable to nucleosome binding.

I then used the composite fasta file to make an alignment and probability-score logo of each gene class using the online web-tool weblogo 3.4 (167). From this analysis it was clear that both sets of ordered genes ± Pol II had the same level of AT enrichment (figure 5.9). When the fuzzy gene group was split based on Pol II a different picture arose (figure 5.9). In this case the genes with Pol II had no detectable sequence bias, whereas the genes with no Pol II binding showed an enrichment of A and T nucleotides. This difference in the underlying sequences of the fuzzy genes and not the ordered genes could be explained by the observations that the ordered gene class ± Pol II had more similar Pol II profiles, only one was much lower, and the fuzzy gene class had an all-or-nothing distinction in terms of Pol II binding.

5.10 Histone variant H2A.Z is enriched at ordered genes

To determine if there may be differences in the chromatin state (histone variants and histone modifications), that could contribute to the architectural differences of the gene classes, I analysed the binding of histone variant H2A.Z and several histone marks at the ordered and fuzzy genes ± Pol II. I analysed a whole-head H2A.Z ChIP-seq dataset (34), as well as the neuronal histone mark datasets H3K36me3, H3K27ac, and H3K27me3. I used the neuron-specific histone datasets as I did not perform head ChIP-seq analysis of these histone modifications, and the neuron datasets will represent a major proporation of the cells in the head. All analysis was plotted as $\log_2$ fold change over head H3-ChIP-seq to remove the influence of different nucleosome levels (34) (figure 5.10.

The results of the chromatin analysis are most intriguing for H2A.Z (figure 5.10 A), where a vast difference exists between ordered and fuzzy genes, regardless of gene activity. The fuzzy genes with Pol II had a slight peak of H2A.Z at the promoter region, which was the contribution of a small number of genes (figure 5.10 E). Ordered genes had high levels of H2A.Z around the promoter, regardless of the level of Pol II binding or transcription at the genes (figure 5.10 A). Ordered genes also had high levels of the active histone marks H3K27ac and H3K36me3 (figure 5.10 B and C), with the **Ordered no Pol II** gene class showing slightly lower levels of these marks than the **Ordered Pol II** gene class. Despite a lack of H2A.Z at the **Fuzzy Pol II** genes, these genes showed H3K27ac and H3K36me3 levels that were comparable to the ordered genes. Those genes that had fuzzy promoters, without Pol II did not have enrichment of the "active" chromatin marks, and had some increase in H3K27me3 enrichment compared to the other gene groups (figure 5.10 D).

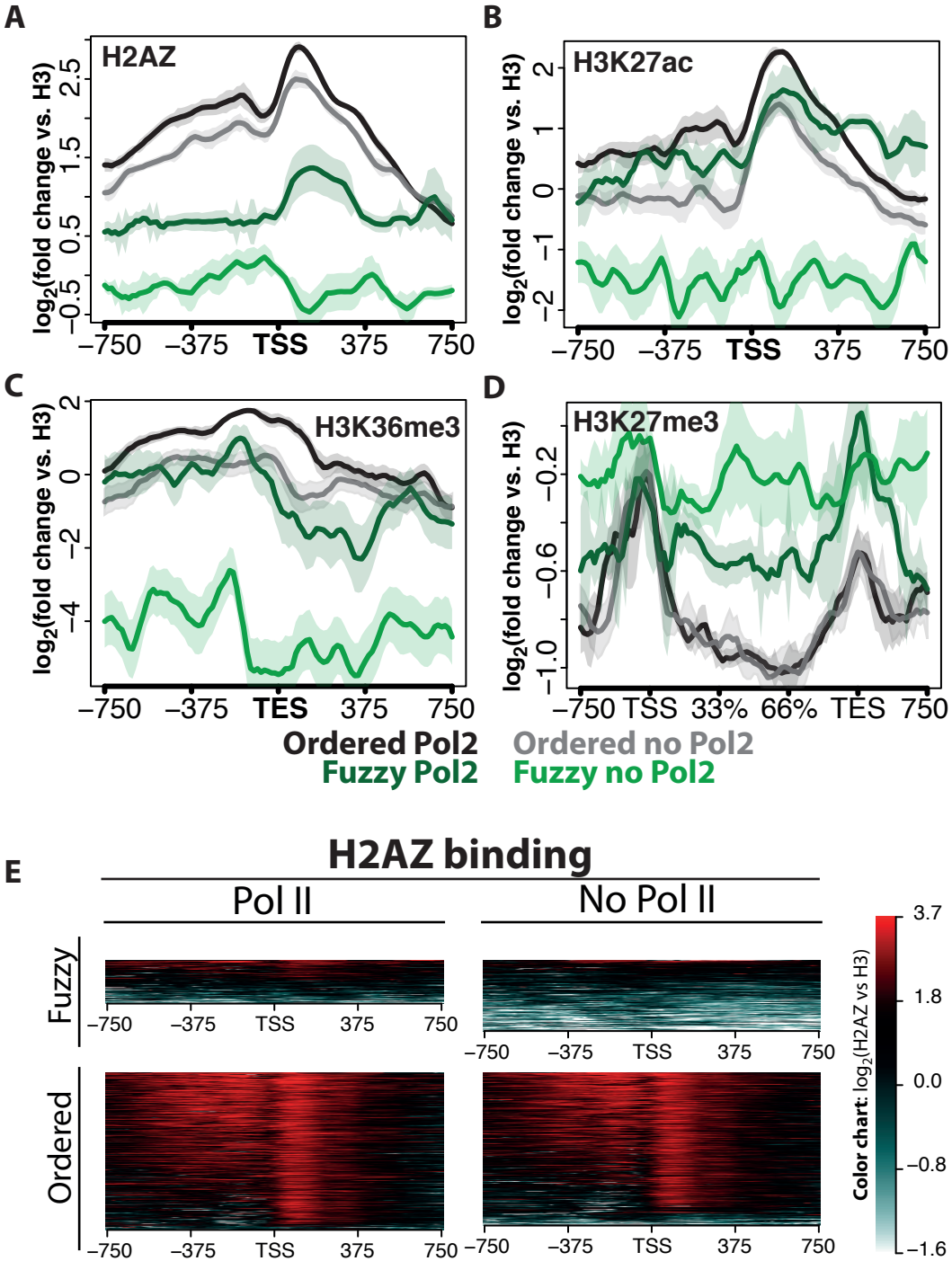


Figure 5.10: Histone marks reflect nucleosome architecture more than expression level - Average profile plots of H2A.Z (A), H3K27ac (B), H3K36me3 (C), and H3K27me3 (D) across the ordered and fuzzy ± Pol II gene groups. E) Heatmaps, ranked by heirarchical clustering, of H2A.Z binding across the tss of ordered and fuzzy genes ± Pol II genes.

These findings are in agreement with the observations discussed in chapter 3. Those genes with constitutively open promoter architectures, the invariant and neuronal-nucRNA genes, had high levels of the active histone marks regardless of the level of gene activity. Here, the ordered promoters showed the same features, high levels of active marks and H2A.Z, even when the genes had different levels of Pol II binding and gene expression. A key point of difference between these analysis and those of chapter 3 is that the fuzzy genes with Pol II had enrichment of H3K36me3. This indicates that the H3K36me3-less feature is not the predominant chromatin state for genes regulated through modulating accessibility to the promoter, by occluding the promoter with nucleosomes. Therefore a fuzzy nucleosome architecture is a feature of active-H3K36me3-less genes, but active genes with fuzzy promoters are not necessarily H3K36me3-less.

5.11 H2A.Z binding is predictive of ordered promoters

The observation that H2A.Z was highly enriched at ordered genes and depleted at fuzzy genes (figure 5.10) led me to ask what the chromatin architecture is of genes with and without H2A.Z. To do this, I generated a list of genes called to have H2A.Z (CCAT peak-calling from (34), classed as **H2A.Z**), and a list of all remaining genes that were not called to have an H2A.Z peak (**No H2A.Z**). To ensure that the genes were classified correctly, I generated an average profile plot of the H2A.Z data (figure 5.11, A), which clearly showed high levels of H2A.Z binding at the **H2A.Z** genes and no H2A.Z binding for the **No H2A.Z** genes.

Next, I used the H2A.Z gene classes to plot the head MNase-seq data (figure 5.11, B). It can be seen clearly from the analysis that splitting genes based purely on presence or absence of H2A.Z produced two types of average nucleosome profiles: an ordered nucleosome architecture for H2A.Z genes, and a fuzzy nucleosome architecture for non-H2A.Z genes. Both sets of genes were expressed, as shown by analysis of nucRNA data (figure 5.11, C), albeit to a lower level for non-H2A.Z genes (p-value = 0.0664). Interestingly, there was a difference in the level of Pol II binding across the two gene classes (figure 5.11, D). H2A.Z bound genes had a significantly higher level of Pol II binding across the promoter region than the non-H2A.Z genes (p-value = 0.004154, *Welch 2 sample t.test*). I next determined how many of the genes from each of my promoter architecture groups, ordered and fuzzy, overlapped with the H2A.Z and the No H2A.Z groups. As expected, 95 % of the genes in the ordered gene category overlapped with genes called as containing an H2A.Z peak whereas less than 40 % of the genes in the

fuzzy category overlapped with H2A.Z-bound genes. This indicates a strong link between the "orderedness" of a promoter's nucleosomes and H2A.Z, which is independent of gene expression levels.

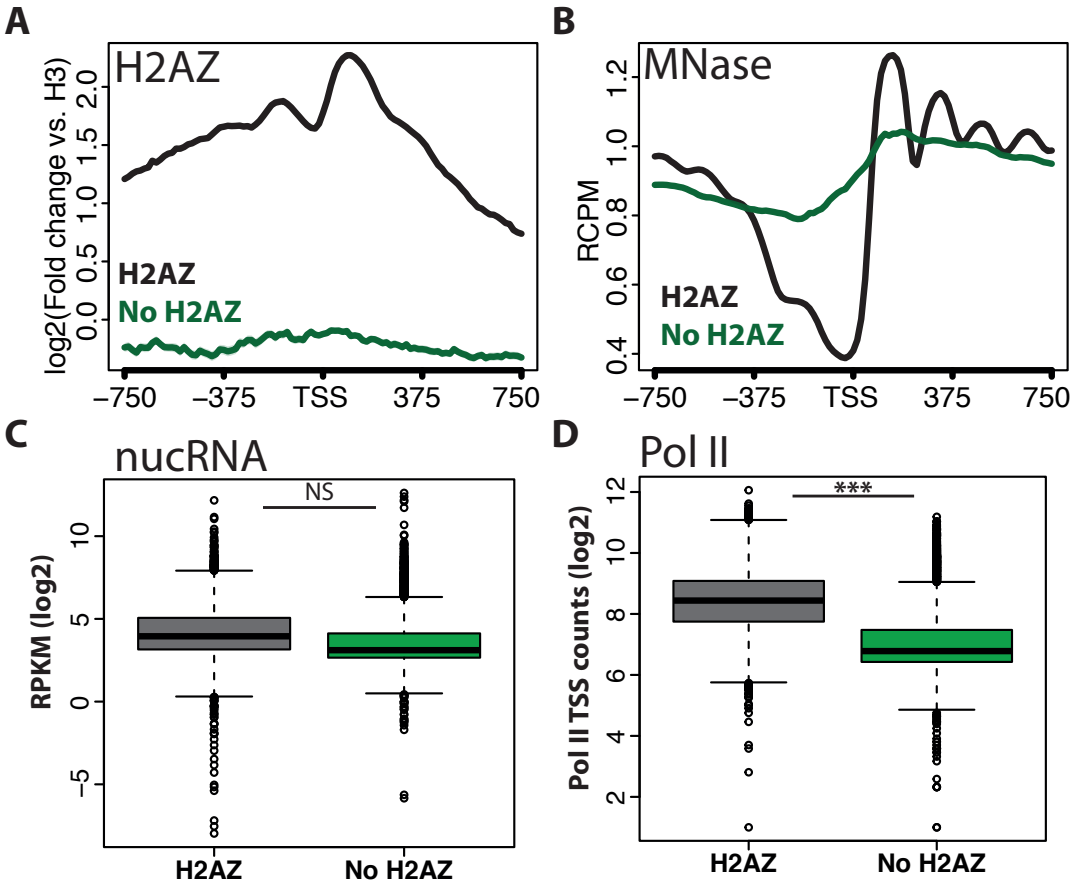


Figure 5.11: H2A.Z is present at ordered genes and not fuzzy genes - Genes were split according to whether or not there was an H2A.Z peak called at the gene or not. A) Average profile plots of H2A.Z-ChIP data showing that the gene divisions based on peaks call were correct. B) Average profile plots of head MNase-seq data aligned to the TSS at H2A.Z-positive and H2A.Z-negative genes. C) RPKMs of H2A.Z-positive and H2A.Z-negative genes. D) Difference in the level of Pol II binding across the TSS ± 250 bp between H2A.Z-positive and H2A.Z-negative genes.

5.12 Fuzzy genes are enriched for paused Pol II

The **Fuzzy Pol II** gene class shows the characteristics associated with paused Pol II (introduction 1.1.1), such as the shape and position of the Pol II binding profile and the high levels of short nucRNA reads at the 5′ end of genes. High levels of H2A.Z have been shown to be negativley correlated with Pol II pausing (18), which is in agreement with the observation made here that fuzzy genes have far lower levels of H2A.Z than the ordered genes. To test if the fuzzy genes are more frequently regulated by Pol II pausing than the ordered genes, I calculated a Pausing Index (PI), (figure 5.12 A). I then ranked all genes and split them into quintiles based on the PI (6256 genes in each group).

With the quintiles of PI, I assessed whether the RNA expression and chromatin structure follow the patterns one would expect from the different pausing quintiles. I generated average profile plots of both the nuclear RNA and the MNase-seq datasets splitting the data into the pausing quintiles (figure 5.12, C and D). Both the RNA and the MNase patterns show what would be expected from separating genes by PI. The nucRNA had a high peak of reads at the 5′ end of genes for the highest PI quintile, and also had the highest levels of transcripts overall (figure 5.12, C). The shape of the nucRNA profile was quite different across the genes in the next highest quintile (quintile 60-80). The profile at these genes was generally much flatter, with only a small peak of reads accumulating at the 5′ end of the gene. The remaining lower quintiles continued to decrease in expression levels consistent with the decreasing levels of Pol II binding.

When observing the MNase-seq data for each PI quintile, some interesting features can be noted (figure 5.12, D). First, for genes with the highest level of pausing, the nucleosome pattern was consistent with the current models for Pol II pausing. The key features are a deep NDR, a highly positioned +1 nucleosome, and the nucleosomes become quickly depleted and disordered into the gene body. Under closer inspection there were also an observable difference between the highest quintile and the 60–80 quintile in the position of the +1 relative to the TSS. The 60–80 quintile had much lower levels of Pol II than the highest quintile, however the NDR was just as deep. The middle quintile and the 60–80 quintile had highly similar RNA and nucleosome profiles, with small differences in levels that are proportional to the difference in Pol II binding. The lowest quintile and the second lowest (20–40 quintile) had very little nucleosome organisation, and there was no observable NDR. These observations show that when genes are ranked based on pausing index, the nucleosome architecture at the

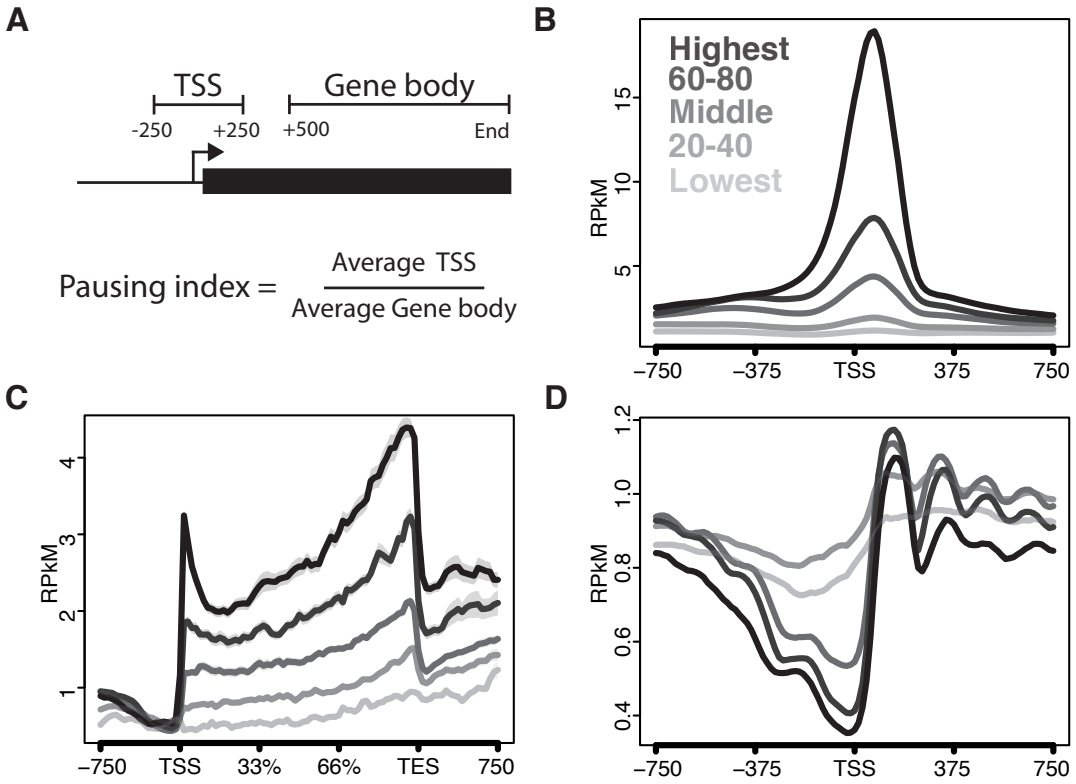


Figure 5.12: Calculating the pausing index based on RPB3 ChIP-seq - Dividing genes based on level of Pol II pausing reveals a promoter nuclear architecture fitting the current model for gene activity and promoter structure. A) Calculating the pausing index based on RPB3 ChIP-seq data. Read-counts were calculated over the TSS ± 250 bp and also over the gene body, from 500 bp into the gene to the TES. Counts were then divided by the length of the region they were calculated over to get average reads per bp for the TSS or gene body. These average numbers are then used to generate the pausing index. B) The average profile of RPB3 data across genes split into quintiles based on pausing index. C) nucRNA levels across pausing-quintiles. D) nucleosome profiles at the TSS across the pausing quintiles.

promoter falls into the expected patterns: 1) Genes with a high pausing index have a deep NDR, a highly positioned +1 nucleosome and depletion of nucleosomes into the gene; 2) Pol II-bound genes with a low pausing index show an NDR, a highly positioned +1 nucleosome, and an ordered array of nucleosomes into the gene; 3) Genes with very low or no Pol II have no NDR, the +1 nucleosome is not well positioned and there is no ordered nucleosome array into the gene.

To assess the level of pausing in the **ordered Pol II** and **fuzzy Pol II** gene sets, I intersected each gene list with the genes in each pausing index quintile, and generated pie charts to compare the distribution of the groups (figure 5.13). The ordered and fuzzy gene sets have a strikingly different distributions of the pausing quintiles. The genes with fuzzy nucleosome architecture indeed have a higher proportion of the most highly paused genes compared to the ordered genes (35 % in fuzzy versus 22 % in ordered). This was expected, as these genes had sharp Pol II binding profiles (section 5.7). The ordered genes have a larger proportion of genes within the second highest pausing quintile than the fuzzy gene class. This was also in agreement with the broad Pol II binding profile at these genes (section 5.7). There were very few genes in either the ordered or fuzzy category that had a PI in the lowest two quintiles.

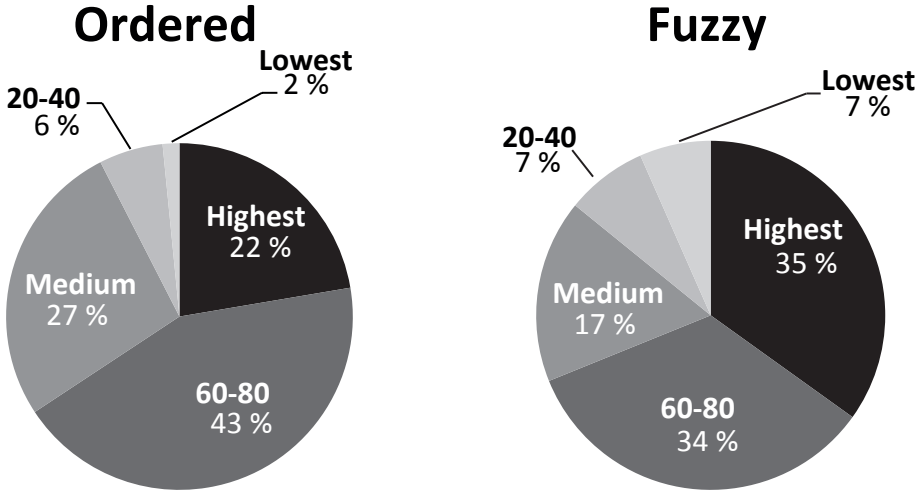


Figure 5.13: Fuzzy genes have higher pausing indices - Pausing indices were split into quintiles, and the number of genes in the ordered or fuzzy gene class that fell into each pausing index group was calculated. Ordered genes are more enriched for genes in the 60–80 and middle pausing index quintiles. Fuzzy genes are more enriched for genes that fall into the highest and 60–80 pausing quintiles.

5.13 Discussion

Specific gene expression programs require specialised regulation. Different types of genes will be subjected to more or less stringent regulatory processes, which would evolve alongside the evolution of the gene itself. Here I have described the characteristics of gene regulation through the architecture of nucleosomes across the promoter region. When observing promoter architecture from the point of view of Pol II binding, there is a clear relationship between the level of Pol II binding and clearance of nucleosomes from the promoter, i.e. NDR formation. However, when the gene expression/Pol II binding is observed from the perspective of the nucleosome architecture, the result was unexpected. I identified a subset of genes with very high Pol II binding, yet had the fuzziest nucleosome architectures of all genes in the genome. I have summarised the findings and projected model of gene regulation at constitutively open promoters of "ordered" genes compared to "fuzzy" genes (figure 5.14).

At the DNA sequence level, the ordered and fuzzy promoters are very different. The ordered promoters are highly AT-rich, which indicates that these promoters would be intrinsically unfavourable to nucleosome binding. Conversely, the fuzzy promoters have highly varied sequences across the promoter region and thus would be more favourable for nucleosome formation. These underlying DNA sequences would be the basis on which the chromatin state and regulatory mechanisms are formed. Although the cause and consequence of what drives the chromatin state at different genes is not clear, the DNA sequence differences at gene promoters is likely a fundamental factor.

Ordered gene promoters could be founded on an intrinsic NDR, guided by the underlying DNA sequence. This intrinsically open region would be a large enough platform for SWR1 to bind and deposit H2A.Z within the surrounding nucleosomes (figure 5.14, A). This mechanism would imply that H2A.Z deposition is independent of transcriptional activity, and would be in agreement with the observation that ordered genes have high levels of H2A.Z incorporation, regardless of Pol II occupancy (section 5.10). Other chromatin remodellers would be involved in defining the ordered spacing of the nucleosome array, a mechanism that is also likely to be independent of transcription. This generates an open promoter that would be available for pre-initiation complex (PIC) assembly, with little hinderance from the local chromatin architecture (figure 5.14, B). During the transcription process of the ordered genes, the active histone modifications are set in place (figure 5.14, C). Both H2A.Z and H3K36me3 have been demonstrated to negatively correlate with nucleosome turnover (51, 168). Thus at the the ordered

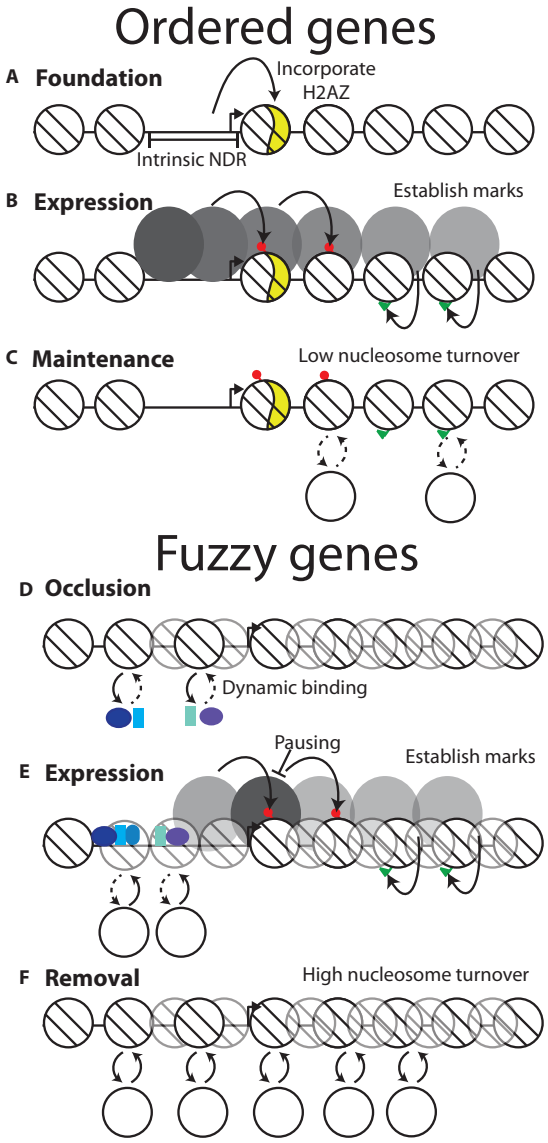


Figure 5.14: Model for establishing different nucleosome architectures - Model depicting how the ordered and fuzzy promoter arcitectures may be established. For ordered promoter architectures; A) the DNA sequence under the promoter is unfavourable for nucleosome formation. The intrinsic nucleosome-depleted region (NDR) allows SWR1 to bind and incorporate H2A.Z, even in the absence of transcription. B) Pol II is recruited to the ordered gene promoters and transcriptional elongation occurs with little regulation at teh inititation step (pausing). During transcription the active histone marks are added to the genebody nucleosomes. C) Low nucleosome turnover, likely driven by high H2A.Z and H3K36me3, results in maintainence of the active chromatin state after transcription has ceased. Fo fuzzy promoter architectures; D) the DNA sequence under the promoter is favourable for nucleosome formation, meaning transcription factors compete with nucleosomes for binding the DNA. E) When transcription factors do bind, the PIC is assembled and inititates transcription, but Pol II is paused 50-100 base pairs into the gene. Nucleosomes re-occupy the promoter region once Pol II has progressed into the paused state, so that no nucleosome-free region is available for H2A.Z incorporation. F) After transcription is complete, high nucleosome turnover ensures the chromatin is reset to a repressive state in the absence of the necessary transcription factors.

genes, nucleosome turnover is expected to be low, and therefore the active histone modifications may be maintained for long periods after transcription of the gene has stopped.

A key feature of the ordered promoters is that they are constitutively open, meaning that they have an NDR even in the cell type where the gene is not currently active. However, this does not mean that these genes are not subjected to additional modes of activation. As shown in the previous chapter (chapter 4), there are genes expressed in all cell types that have an additional increased expression level in neurons. Thus, these ordered genes likely have low levels of basal transcription that is increased, either at the level of Pol II recruitment or elongation, by additional factors. I expect that these genes, while not necessarily house-keeping genes, would not have deleterious effects if expressed when not required. This is because genes which cause problems when misexpressed would be expected to have evolved a more complex mechanism that ensures efficient and precise shutdown of the gene when it is not required, and have efficient and precise gene activation when it is required. If a gene is not deleterious to the cell when expressed even when it is not required, then there would be no evolutionary advantage to develop a highly thourough and specialised repression system for those genes.

Fuzzy gene promoters have an underlying sequence that we would expect to be preferable to nucleosome formation. Thus fuzzy promoters are intrinsically occluded by nucleosomes, and assembly of the PIC at the promoter would require a coordinated effort of many factors (figure 5.14, D and E). Dynamic competition between these factors and the nucleosomes for binding the promoter DNA ensures the precise regulation of these genes. Only when the binding is tipped in favour of the regulatory factors can the transcriptional machinery be recruited to these genes. Once the PIC is assembled, the Pol II progresses into the gene and is further regulated by pausing mechanisms (figure 5.14, E). Additionallly, the active histone marks are established across these genes during transcription (figure 5.14, E). Higher nucleosome turnover at these genes would drive dynamic removal of these marks, to ensure complete repression of the gene in the inactive state (figure 5.14, F).

High levels of Pol II pausing have been shown to alter the nucleosome architecture at the promoter, leading to a positioned +1 nucleosome, and clearing of the NDR (17). However, that high Pol II binding and pausing always leads to a cleared NDR and positioned + 1 nucleosome is not true. I have shown in this chapter that there is a subclass of genes that have the fuzziest nucleome architecture of all genes in the genome, yet have high levels of Pol II binding, a significant proportion of which is in the highest

two quartiles of paused Pol II. These genes do represent a small proportion of the total genes present in the *Drosophila* genome and are likely an exception rather than the rule. However, I that these genes will be regulated with exceptionally rigourous control, and thus exemplify the chromatin features involved in precise spatial and temporal gene expression.

The majority of the fuzzy, paused Pol II genes, lack H2A.Z. And, although a relationship between H2A.Z and ordered nucleosome architectures has been shown previously, it is not clear what the function of this relationship actually is. In yeast, depletion of H2A.Z did nothing to affect the ordered arrangement of the nucleosomes at the promoter (33). Loss of H2A.Z in *Drosophila* is lethal, presumably through a loss of global gene activity, which would make sense in that H2A.Z is found to be enriched with tissue-invariant genes (34), whose expression is necessary for basic cell functions. H2A.Z is clearly not necessary for the activity of many genes, particularly those regulated through Pol II pausing (18). However, it has yet to be shown whether ectopic incorporation of H2A.Z at non-H2A.Z genes has an effect on gene expression. I would expect that forced incorporation of H2A.Z at non-H2A.Z genes would greatly impact their ability to be temporally and specifically regulated. I propose that the rapid incorporation of nucleosomes over the promoter region is a necessary step to ensure exclusion of H2A.Z, and thus ensure the maintainence of specific regulation of that gene.

It is likely that higher binding preference of nucleosomes at the promoter is not the only driving force behind exclusion of SWR1 and H2A.Z; other chromatin remodelers will be neccessary. Interestingly, a recently published manuscript showed that loss of the ubiquitous histone chaperones FACT and Spt6 are involved in blocking H2A.Z incorporation into intragenic regions (169). The model proposed in Jeronimo et. al., (2015), describes how active transcription requires the removal of nucleosomes, and thus the exposure of DNA to the SWR1 complex. FACT and Spt6 are then required for removal of H2A.Z and incorporation of cannonical H2A, and thus the repression of internal transcription start sites. I speculate that FACT could have an equally important role in preventing H2A.Z incorporation at active fuzzy promoters, and thus ensure specific regulation. It has been shown previously that inhibition of FACT, using a temperature-sensitive FACT mutant, has highly precise phenotypic outcomes depending on the time of development FACT was inhibited (170). For example, the inability to eclose completely from the pupae, even though development to the adult form is unhindered (data not shown). This would be consistent with a loss of control of temporal and spatial specific gene regulation.

5.14 Future directions

It would be extremely interesting to determine what the consequences would be if a promoter was modified to resemble to opposite architecture, for example by editing the amount of AT. Would a transition from fuzzy to ordered promoter sequence enable the incorporation of H2A.Z? Would it cause a severe phenotype due to ectopic expression? Would the gene then exhibit an ordered nucleosome array, even though the sequence underlying the gene region was unchanged? Testing these questions would be technically challenging. One experiment could be using reporter constructs of selected gene promoters upstream of GFP, for example. The promoter regions could then be altered to be more or less nucleosome favourable. These reporter constructs would need to be incorporated into the genome at the same location, such as using phiC31 insertion sights. Such an experimental set-up would enable the nucleosome organisation, H2A.Z incorporation and expression level to be assessed as well as determining if ectopic expression at different developmental stages occurs.

Many further experiments would be useful to test the two models of gene regulation described in this chapter. Firstly, whether the underlying DNA sequence is indeed guiding an intrinsic loss of nucleosomes at the ordered gene promoters. *In vitro* nucleosome reconstitution assays would be ideal to test this hypothesis. If the system behaves in the same way that yeast promoters do, then reconstitution of nucleosomes without any cell extract or ATP would show little occupancy of nucleosomes over the ordered gene promoters, but high occupancy of nucleosomes over the fuzzy promoters. This system could also be further used to test the incorporation of H2A.Z, by adding H2A.Z, SWR1, and ATP to see if H2A.Z is indeed preferentially incorporated into the ordered genes. Testing whether there is truly more regulation by Pol II pausing at the fuzzy genes could be done by RNAi of the Negative Elongation Factor (NELF). NELFs role is to promote Pol II pausing, and removal of NELF is needed for Pol II to proceed into productive elongation. By artificially removing NELF, the pausing check point is removed and there will be alterations at the transcriptional level at pause-regulated genes.

To test whether FACT is involved in repressing incorporation of active genes with fuzzy promoters, the temperature-sensitive Dre4 mutant fly strain could be used (170). Once flies have reached adulthood, switching them to restrictive temperature would inhibit FACT function, and a non-ts mutant Dre4 can be used as a control. Comparing the H2A.Z occupancy by ChIP-seq between the $Dre4^{ts}$ mutant and control group would provide a global picture of where H2A.Z is incorporated in the absence of FACT activity.

5. THE ROLE OF PROMOTER ARCHITECTURE IN DEFINING GENE EXPRESSION PROGRAMS

Outlook

The answer to the question of how genes are expressed in the right place, and at the right time, is clearly complex. In this thesis, I have demonstrated that multiple genomics tools can be used to obtain a more refined picture of the chromatin state aids by dissecting out some of the inherent complexity. My analyses revealed levels of gene regulation that were not previously considered. Just as I have shown in chapter 3,that one method for measuring gene activity alone is not enough for grasping the entire repertoire of active genes, looking at a single type of "model" cell, or an average of a whole organism, is not sufficient for understanding the complexities of gene regulation. Indeed, precise spatial and temporal gene expression is driven by combinations of subtle mechanisms and co-operativity that is lost when looking at global averages. A key conclusion I make is that although the end result of gene regulation can be the same, i.e. that a given protein is present in a specific cell type, how that is achieved can be vastly different for each individual gene. Some genes will be regulated by Pol II recruitment, some genes will be regulated by transcriptional elongation, some genes will be regulated during mRNA translation, and perhaps at every step in-between. Including at the level of RNA processing, transport and degradation. We now have many tools to assay these steps of regulation with high spatial precision. By using combinations of these tools, we can start to build a more refined and complex understanding of gene-regulatory systems.

There are, however, still limitations to this type of analysis. For example, even though I have assessed two distinct populations of cells with defined functions, I am still averaging across multiple cells. I have used the pan-neuronal elav-Gal4 driver to label what I term "neurons", yet, there is a highly diverse mixture of neuronal sub-types that comprise this population. Each neuron in the head is likely to have some unique features involved in the functional specialisation of that neuron, which are lost when averaging the whole neuronal population. There are also limitations in that I have averaged multiple genes, defined by some unifying feature, which will average out any information about the many mechanisms governing their expression. This means that

only the mechanisms that are most prevalent in the gene population will be distinguishable. However, by refining my analyses from whole head to neurons and glia, I have uncovered that the most prevalent gene-regulatory mechanisms strongly differ between these cell types. My project has also only measured a single time point. Tested here were adult flies, which were 2–4 days old, and at a circadian time-point of 17:00–18:00. Even within a single neuron, gene expression likely changes over time, so we are still missing the context of chromatin state in the temporal regulation of genes. Thus, the analysis performed here provides a more refined view of the most abundant mechanisms involved in regulating major gene expression differences between neurons and glia at a single time point.

To further address the question about temporal gene regulation in the *Drosophila* head, we are currently performing experiments to study gene expression changes during fasting. I have obtained and mapped ChIP-seq datasets measuring the binding of Pol II in adult female heads after six and nine hours of food deprivation (not shown). These new data complement an RNA-seq dataset that measured the changes in gene expression from zero to 24 hours of starvation (T. Schauer, *unpublished*). This new avenue of research will also move forward into cell-type-specific gene regulatory mechanisms of the fasting response. These data will allow us to understand the gene regulatory changes during fasting both spatially and temporally.

Cell-type-specific tools have by no means been pushed to the limits. With further optimisation, the tools used here could be used to assay far smaller sub-populations. Particularly, the split-Gal4 system could be used to assay small, highly defined, cell populations (171). Our understanding of mechanisms driving cell-type-specific gene regulation will grow exponentially, especially when technologies for single-cell analysis become more routine. While many techniques have been developed to achieve a more refined view of general gene regulation mechanisms, such as Gro-seq for probing Pol II pausing (172), ATAC-seq for nucleosome architecture (153), ChIP-exo for looking at highly defined protein-DNA interactions (173), or ribosome foot-printing for measuring translational progression (174), these techniques have yet to be applied to cell-type-specific systems. Further technological advances are required if we aim to investigate the dynamics of the chromatin and transcriptional state in specific cell types. To my knowledge, only steady-state analyses of specific cell types have been investigated thus far, as was shown in my research here. However, methods for studying the dynamic responses of the chromatin state and expression to environmental stimuli will be needed to understand the dynamics of the gene-regulatory system.

The results presented in chapter 4 are a good example of the advantages of using multiple methods in genome-wide studies. The comparison between a Pol II-centric method for identifying gene activity and a method better suited to measuring transcriptional elongation, revealed distinct mechanisms for achieving neuron-specific gene expression. While much further analysis is needed to understand what drives the cell-type-specific elongation, this study uncovered a novel insulator-based model for achieving neuron-specific gene activity. I have shown that a particular subset of neuron-specific genes are highly enriched for the insulator Su(Hw), an insulator that is absent from neurons yet present in glia, and other tissues. Neuron-specific activation would be achieved through relief of repression by Su(Hw) in neurons, a mechanism analogous to the REST-complex function in mammals (175).

Using only a single method to identify neuron-enriched or glia-enriched genes, the complexity of the different types of gene regulation may have been lost. Had only the nucRNA analysis been used as the measure for neuron-enriched or glia-enriched gene expression, I would not then have been aware of the Su(Hw)-mediated mechanism for regulating a subset of neuron-enriched genes. Had I only used Pol II binding as the measure for gene activity, I would not have been aware of specifically regulated genes with highly invariant chromatin states. It is by comparing and contrasting the two types of gene activity measurement that we can visualise the different types of regulation.

I have generated many testable hypotheses from the data analysis presented in this research. However, the analysis performed here is only a small sample of the many types of analysis that can be performed on these data. For example, whether there is enrichment for transcription factor binding sites at the cell-type-specific nucleosome-free regions is still on going. Additionally, the nucRNA data could be explored much further, for instance by identifying non-coding RNAs from each cell type, or identifying differential 3′UTRs. I envision that incorporating many publicly available datasets into future analysis will allow us to obtain an even more detailed view of the chromatin state and gene regulatory mechanisms. Computational analysis is extremely valuable for observing the data, and generating testable hypotheses. However, taking an hypothesis formed *in silico* and testing it experimentally *in vivo* and *in vitro* is also highly important to further our understanding. My overall aim of the research presented in this thesis was to better understand how chromatin structure regulates specialised gene expression programs in neurons and glia. The system-wide approach I have taken, observing and refining the relationships between many different types of dataset, has lead

to finding multiple mechanisms governing the specialised gene regulation. As is the way with biology, more questions have arisen than definite answers. But the questions themselves have become more directed and refined than was previously possible.

Chapter 6

Materials and methods

6.1 *Drosophila* husbandry

Food preparation

All fly stocks were grown on basic food, containing: 76 g cornmeal, 60 g dried yeast, 8.4 g agar, 31.4 g saccarose, 62.7 g glucose, 8.7 g sodium potassium tartrate, 0.7 g calcium chloride, and 2.6 g nipagen per litre of solution. After cooking, the food was left to dry overnight before plugging with stoppers. Additional dried yeast was added on top of the food before use.

Stock generation and maintenance

The Elav::H2B-GFP line was maintained as a homozygous stock, containing two copies each of the Elav-Gal4 driver line. The Repo::H2B-GFP line was made homozygous, using balancer chromosomes (*CyO* for 2nd chromosome, and TM3,*Ser* for 3rd chromosomes). Repo::H2B-GFP flies were outcrossed for six generations with 2202U (wild-type) to ensure the genetic backgrounds between the Elav::H2B-GFP and Repo::H2B-GFP were as similar as possible. Stocks were maintained in small vials at 18 °C.

Line amplification and collection

Stocks for large-scale collection were expanded to bottles of standard food with additional dried yeast. Flies were flipped to new food every 2–3 days and grown at 25 °C at 70 % humidity. Flies were collected within one day of eclosion, into bottles containing standard food and sufficient yeast. Collected flies were kept at 25 °C until the following day between 17:00–18:00, where they were transferred to Falcon tubes and quickly frozen in liquid nitrogen.

6.2 Generating Repo::H2B-GFP *Drosophila* line

6.2.1 Cloning H2B-GFP behind Repo promoter

Cloning strategy

The H2B-GFP transgene was amplified using PCR 6.2.1 from k161pHS-H2B-GFP (vector map, appendix D.1). The PCR amplification used primers containing Not1 and Xba1 restriction sites (see primer list D.1). The H2B-GFP PCR amplicon was cloned into a pCasper4 vector containing the Repo regulatory region (vector map, appendix D.3). The Repo regulatory region is sufficient to drive specific expression of the H2B-GFP transgene in glial cells.

PCR amplification

PCR amplification was performed using the following components:

Table 6.1: PCR reaction setup - for 50 μL reaction

Component	Quantity
5 X Phusion buffer	10 μL
10 mM dNTPs	1 μL
10 mM primers	2.5 μL of each
Phusion polymerase	1 μL
Template	1 ng
H_20	To 50 μL

Amplification of the H2B-GFP transgene from the k161pHS-H2B-GFP plasmid was performed following the PCR cycle conditions in table 6.2.

Table 6.2: PCR conditions

	Temperature	Time
------	98 °C	1 minute
25	**98**°C	20 seconds
cycles	**55**°C	15 seconds
------	**72**°C	45 seconds
	72°C	6 seconds
	4°C	Finish

Restricition Endonuclease Digestion

PCR amplicons were purified using Wizard SV Gel and PCR clean up kit (Promega) prior to restriction digestion. Restriction digestion conditions were performed as in table 6.3.

Table 6.3: Restriction digest conditions

Reagent	H2B-GFP	pCaspeR4Repo
DNA	1520 ng	1125 ng
Buffer 3	5 μL	5 μL
BSA	0.5 μL	0.5 μL
NotI	0.5 μL	0.5 μL
XbaI	0.5 μL	0.5 μL
H2O	38.5 μL	41 μL

The samples were incubated at 37 °C for 1 hour, then the reaction was stopped by incubating the samples at 65 °C for 10 minutes. The samples were then purified using Wizard SV Gel and PCR clean up kit (Promega).

Ligation

Ligation of the H2B-GFP insert into the pCaspeR4-Repo vector was carried out using two different insert:vector ratios. Reactions were carried out in 20 μL volumes containing 1 x T4 DNA ligase buffer, 100 ng of vector, 15 or 100 ng of insert, 1 μL T4 ligase (NEB), and H_2O to 20 μL. Ligation reactions were incubated at room temperature for 20 minutes, then immediately used for bacterial transformation.

Transformation of plasmid into bacteria

Competent cells (DH5α, 50 μL) were thawed slowly on ice, 5– 10 μL of the DNA solution was mixed with the cells, and then left on ice for 30 minutes. The cells were then heat-shocked at 42° C for 30 seconds, then immediately transfered to ice and 800 μL of LB media was added to them. Cells were incubated at 37 °C with shaking for 1.5 hours. After the incubation, the cells were pelleted by centrifugation, the supernatant removed, then resuspended in 100 μL LB media and plated on LB+ampicillin media.

Identifying positive clones

After overnight incubation at 37 °C, several colonies were picked and used to inoculate first 10 μL of distilled H_2O, then LB+ampicillin broth. The inoculated H_2O sample was used as template in a PCR reaction set up as in 6.2.1. Amplicons were visualised by running the PCR reactions of a 1 % agarose gel, 80 V, for 45 minutes. Positive

colonies, already inoculated in LB+ampicillin broth, were grown at 37 °C overnight. Plasmids were purified from the overnight cultures using a mini-prep kit (Promega) and sent for sequencing. While waiting for sequencing results, larger overnight cultures of the positive clones were inoculated from the previous cultures to purify a larger quantity of plasmid (midi-prep kit, Promega) to send for *Drosophila* transformation (section 6.2.2). Positive clones were re-sequenced after midi-prep clean-up, to ensure the correct construct was maintained in the culture.

6.2.2 Generation of transgenic lines

Transgenic *Drosophila* were generated by the company TheBestGene (thebestgene.com). The Repo::H2B-GFP fly was generated using p-element insertion to incorporate the transgenes into the genome (176), thus insertion sites are random. Insertion was into the fly line 2202U, which is the same genetic background as the Elav-Gal4::H2B-GFP line. The vector contains the *white* gene as a selective marker for positive transformants. Positive transformants of the Repo::H2B-GFP were made homozygous, and then checked using western-blot analysis and immunohistochemistry (sections 6.2.3 and 6.4).

6.2.3 Western blot analysis of Repo::H2B-GFP flies

Sample preparation

Analysis of textitDrosophila head proteins was performed by isolating heads on dry ice, then transferring the heads to a microcentrifuge tube and adding an aliquot of 2 X SDS Buffer, normally to achieve one head per μL of buffer. The heads were ground using disposable dounce until only the exoskeleton remained, the samples were then boiled for two minutes and centrifuged at maximum speed for two minutes. The supernatant was collected, and stored at -20 °C until required. Analysis of FANS-isolated nuclei samples was performed by adding the appropriate amount of 2 X SDS loading buffer to the samples, and mixing the sample by pipetting. The samples were boiled for two minutes, centrifuged at maximum speed for two minutes, and the supernatant was collected and stored at -20 °C until required.

SDS-PAGE

A 15 % acrylamide gel was cast using 4.5 mL acrylamide (30 %), 2.25 ml 4x Tris-Cl/SDS buffer (pH 8.8), 2.25 mL H_2O, 40 μL APS (10 %) and 18 μL TEMED. Samples were diluted as required to volumes of 20 μL with 2 X SDS loading buffer and the SDS-PAGE was run at 200 V for 45 minutes.

Transfer to membrane

An appropriately sized piece of nitrocellulose membrane (Protran BA 83 Whatman, GE Healthcare Life Sciences) was equilibrated with transfer buffer for 5–10 minutes. The gel and membrane were arranged within the transfer apparatus between two layers of Whatmann paper, and a double layer of sponge to ensure close contact of the gel with the membrane. Assembly was done with all components saturated with transfer buffer. The samples were transferred from the gel to the membrane at 100 V for 60 minutes.

Antibody staining

The membrane was blocked for one hour at room temperature in TBST+ 5 % milk. The membrane was then incubated with α H2B-HRP antibody (ab64039, abcam), diluted 1:40 000 in TBST + 5 % Milk, for one hour at room temperature. After incubation, the membrane was washed three times, for five minutes each. The membrane was then incubated with chemiluminescent HRP-substrate (Millipore) for one minute and exposed to photographic film for detection (Kodak).

6.3 Chromatin and expression analysis

6.3.1 Quantitative-PCR

PCR primer design

The primer design for histone modifications (H3K27ac, H3K27me3, H3K36me3) were designed by selecting regions from published sequencing data (81, 177, 178). Positive control regions were defined by selecting the genomic sequence under positive peaks in the IGV browser (www.broadinstitute.org. Negative regions were selected from either nearby genomic regions with no positive enrichment of the histone modification, or other genes that showed no enrichment. Primers to amplify these regions were designed manually in Ape http://biologylabs.utah.edu with the following parameters as guides:GC content 50–60 %, length 18–24 nucleotides, melting temp 60–63 °C. Primers were analyzed against the *Drosophila* genome using BLAST http://blast.ncbi.nlm.nih.gov/Blast, to confirm specificity for the target sequence. New primer pairs were tested using a standard curve to check for non-linear amplification and off-target amplification products. Primer pairs for nucRNA qPCR analysis were designed and tested by Tamas Schauer.

Reverse-transcription for RNA analysis

Reverse transcription of RNA samples was performed using the SuperScript III (Invitrogen) reverse transcription kit. Samples were prepared as per the kit protocol using random hexamers as primers.

Reaction setup and analysis

The input samples for each ChIP were diluted 1:20 and 1:200, and the ChIP sample was diluted to 1:20 in nuclease-free water. If ChIP samples were of very low concentration, dilutions of 1:10 and 1:100 (input) and 1:10 (sample) were used. The qPCR reaction was set up at room temperature adding 6 μL DNA, 1.5 μL primer pairs (table D.1, and 7.5 μL fast SYBR green master mix (Applied Biosystems). Quantitative PCR amplification was performed using a 7500 FAST Real-Time PCR system (Applied Biosystems) using standard amplification conditions, with a melting curve. ChIP enrichment over input, as percentage input, was calculated as follows:

Calculated cycle difference between 10 F dilution of input =
Ct value (1:200) - Ct value (1:20)
Calculated 100 % = Ct value (1:200) - (2 x calculated cycle difference)
Normalised ChIP Ct = ChIP Ct - calculated 100 % / -cycle difference
% input =$10^{\,normalisedChIPCt}$ x 100

6.3.2 Isolation of *Drosophila* heads

The amount of *Drosophila* used varied depending on the required use of the nuclei. Heads were dissociated from *Drosophila* frozen at -80°C by strongly tapping the tube on the bench. Three cycles of five rapid taps on each end of the tube and placing the tube back into liquid nitrogen for one minute, were sufficient to remove most heads from the sample. The sample was sieved through a 630 μm and a 420 μm mesh. Torsos of the flies are retained in the 630 μm sieve, heads are retained in the 420 μm sieve, whilst wings and legs pass through. Heads were ground to a fine powder on dry ice using a mortar and pestle.

6.3.3 Fluorescence-activated nuclei sorting

The nuclei of the cell type of interest were labelled with the reporter gene H2B-GFP (section 1.3.3. Heads were collected from approximately 50 mL of *Drosophila*, frozen in liquid nitrogen as in section 6.3.2. The ground heads were resuspended in 50 mL of Buffer NPB (section 6.6), and incubated for 2 minutes on ice. Samples were then

cross-linked with 1 % formaldehyde for 10 minutes at room temperature, after which the formaldehyde was quenched for 5 minutes by addition of 2.5 mL of 2.5 M glycine. The samples were then transferred back to ice, and homogenised using a 100 mL electric dounce homogeniser at 4 °C, ensuring 30 up/down cycles were completed. The samples were then filtered through a 60 μm filter followed by filtration using a 10 μm filter (Millipore) that was pre-wet with NPB + 5 % BSA. Nuclei were pelleted by centrifuging at 2000 rpm, for five minutes, at 4 °C. The supernatant was carefully removed, then the pellet of nuclei were gently re-suspended with 1 mL of NPB + 5 % BSA. The nuclei were centrifuged at 1000 rpm for one minute, the supernatant removed and the nuclei again resuspended in 1 mL of NBP + 5 % BSA.

GFP-positive nuclei were collected using a FACS-Aria III cell sorter. Prior to sorting, the nuclei were diluted by adding 100 μL nuclei sample to 1 mL NPB + 5 % BSA. The nuclei were gently separated by passing 10 X through a 22G needle, then 5 X with a 25G needle, with a 1 mL syringe, and the sample was then filtered into the sorting tube through a 40 μm cell strainer (BD Biosciences). The optimal dilution was calibrated for each sort; normally the optimal was a 10-fold dilution. Events were kept to a maximum of 10 000 events/s to reduce the required pressure and increase the yield, the laser intensities were adjusted as necessary for each sorting session. For chromatin analysis, samples of one million nuclei were collected, collection tubes were pre-incubated with NBP + 5 % BSA and 1 mL of NPB was added to the collection tube for sorting. Frequent re-suspension of the samples was necessary to decrease clumping, and increase the yeild of nuclei per sample. The entire FACS-Aria system, including sample collection, is kept at 4 °C for the entire sort. After sample collection, the samples were centrifuged at 2000 rpm for 15 minutes, and the majority of the supernatant was carefully removed, leaving around 500 μL. The nuclei were then re-suspended in the remaining supernatant, and transferred to a low-binding micro-centrifuge tube (Costar). The nuclei were centrifuged for five minutes at 2000 rpm, and the remaining supernatant was removed. The nuclei pellets were stored at -80 °C until needed.

6.3.4 Chromatin Preparation

Standard Protocol

Ground heads from 15 mL of flies prepared as in section 6.3.2, were transferred to a 50 ml Falcon tube and 25 mL of NPB was added and the sample was left to incubate on ice for 5 minutes. The sample was homogenised using a 100 mL electric dounce homogeniser at 4 °C for 25 times. The sample was then cross-linked by adding 675 μL

of 37 % formaldehyde, rotating at room temperature for 10 minutes. Formaldehyde was quenched in the sample by the addition of 1.25 mL of 2.5 M glycine, rotating at room temperature for 5 minutes. The sample was filtered through a 60 μm filter (Millipore) and centrifuged at 2000 rpm for 5 minutes at 4 °C to pellet the nuclei. After removing the supernatant, the pellet was resuspended in 1 mL RIPA buffer and transferred to a 1.5 mL micro-centrifuge tube. The sample was centrifuged at 2000 rpm for 1 minute, the supernatant removed and the pellet again resuspended in 1 mL RIPA buffer. This was completed a total of 3 times. After the final wash, the pellet was resuspended in 300 μL RIPA buffer, and sonicated using a Branson Sonicator with the settings: Intensity 4; pulse 40; 15 seconds pulsing; 45 second break, repeated 7 times. The sample was then split into two 150 μL aliquots and transferred to covaris snap cap tubes. Each aliquot was sonicated using the Covaris sonicator with settings: 4 minutes, DF 200. After Covaris sonication, the samples were centrifuged at maximum speed (20, 000 RPM) at 4 °C for 10 minutes. The soluble chromatin in the supernatant was taken up and kept at -80 °C until required.

Protocol for FACS isolated nuclei

A total of one million nuclei collected by FACS (section 6.3.3), were centrifuged at 1000 rpm, 4 °C, for 20 minutes. The majority of supernatant was removed, leaving $\sim$ 0.5 mL of supernatant which was used to re-suspend the pellet. The sample was transferred to a siliconised 1.5 mL micro-centrifuge tube and centrifuged at 1000 rpm, for 10 minutes at 4 °C. The supernatant was removed and the pellet was resuspended in 150 μL RIPA buffer. The sample was transferred to a Covaris snap-cap tube, and sonicated using the Covaris sonicator using the settings: 6 minutes, DF 175. The samples were transferred to siliconised 1.5 mL micro-centrifuge tubes and were centrifuged at maximum speed (20 000 RPM) at 4 °C for 10 minutes. The soluble chromatin in the supernatant was taken up and kept at -80 °C until required.

6.3.5 Chromatin immunoprecipitation

All centrifugation steps were at 4 °C for one minute at 500rpm. Preparing the beads Protein-G Fast flow sepharose beads (GE Healthcare) were gently resuspended to a slurry (50/50 mixture of beads and water). For each ChIP sample 25 μL beads was used, however the total volume of beads required for each experiment was pooled into a single tube for the washing and blocking steps. To wash the beads, the required volume of beads was transferred to a sterile micro-centrifuge tube, 1 mL of nuclease-free water was added and the beads were rotated at 4 °C for 15 minutes. After centrifugation and removal of the supernatant, the beads were washed quickly two more times, using

1 mL nuclease-free water, centrifuging, and removing supernatant between the washes. The beads were then blocked by adding 1 mL of RIPA+BSA (section 6.6) and rotating the beads for at least one hour at 4 °C, then washed four times 5 minutes with rotation in RIPA buffer (section 6.6). Chromatin prepared as described in section 6.3.4 was thawed on ice, and 10 μg of chromatin was diluted to 650 μL in ice cold RIPA buffer.

The ChIP assay was performed by incubating 500 μL of the diluted chromatin with the beads, rotating for 3 hours at 4 °C. A 50 μL sample of the diluted chromatin was also collect as an input sample. The ChIP sample was centrifuged at 500 rpm for 1 minute, the supernatant was removed from the beads and kept in a fresh micro-centrifuge tube. This sample was used to assess the chromatin fragment size later. To the beads, 1 mL of RIPA buffer was added and the samples were washed with rotation for five minutes. The centrifugation, supernatant removal (now discarding), and addition of fresh RIPA buffer was performed a total of four times. The beads were then more stringently washed with LiCl buffer (section 6.6) for 10 minutes. The beads were then centrifuged at 500 rpm, the supernatant removed, and the beads were washed quickly twice with 1 mL of TE buffer (section 6.6). After the final wash of TE buffer was removed, the beads were resuspended in 100 μL TE buffer, and 50 μL of TE buffer was added to each input sample. All tubes, including the supernatant samples, were sealed with parafilm and incubated overnight at 65 °C with vigorous shaking (1000 rpm). The samples were then incubated with 1 μL RNase A (thermo scientific) at 37 °C for 30 minutes. The proteins within the samples were then degraded by adding 5 μL SDS, and 1 μL proteinase K (10 mg/ml, Roche) and incubating at 55 °C for 1.5 hours. The DNA was then purified using Ampure-XP beads (Agencourt) following the manufacturer's directions, with the following exceptions: three washes of 600 μL 70 % ethanol were used at step 6, a dry time of 15 minutes was used, and 15 μL was used for elution. The Ampure-XP beads were added directly to the bead/DNA slurry because material would be lost by transferring the supernatant to a fresh tube. Technical replicates of the ChIP samples were pooled together by eluting the DNA off the beads with the same 15 μL elution buffer for all samples. The concentration of DNA was measured using the Qubit dsDNA high sensitivity assay (life technologies). A minimum of 2 ng of DNA was sent to the EMBL genecore (Heidleberg) for library preparation and sequencing (Illumina hiseq 2000).

Optimizing ChIP conditions for different antibodies

Each antibody required optimisation to the one million nuclei protocol before performing ChIP-seq analysis. I first optimised performed a standard ChIP-qPCR analysis

with the recommended amount of antibody. From this analysis, I determined the optimisation procedure, normally titrating the antibody concentration, then adjusting the bead concentration before optimal conditions were achieved. The antibody and bead conditions for each antibody is shown in table 6.4.

Table 6.4: ChIP conditions for different antibodies

Antibody	Volume antibody	Volume beads
H3K36me3 (abcam ab9050)	2.5 μL	25 μL
H3K27ac (abcam ab 4729)	1 μL	25 μL
H3K27me3 (millipore 07-449)	2 μL	20 μL

6.3.6 Micrococcal nuclease titration

Nuclei were prepared as in section 6.3.3, diluted 1:10 and counted using a haemocytometer to estimate nuclei concentration. For each sample, 500 000 nuclei were centrifuged at 1 000 rpm for 1 min, and the pellet resuspended in 25 μL of MN buffer (section 6.6). MNase was added to replicate samples to 0, 0.1, 0.2, 0.3, 0.4, 0.5, and 1 units per sample. Samples were incubated at 37 °C for 10 mins with gentle shaking, immediately transferred to ice and 4 μL of stop solution (137.5 mM EDTA, 5.5 % SDS) was added. Samples were incubated overnight at 65 °C to reverse formaldehyde cross-links, RNAse-treated, and Proteinase K treated as in section 6.3.5, with the exception that SDS was not added for protienase-K treatment. Samples were mixed with 6 X orange loading dye and DNA separated using gel electrophoresis with a 3 % TBE agarose gel, 65 V, 400 mA, 6 hours. The gel was stained with ethidium bromide and visualised using a GelDoc-IT system (UVP).

6.3.7 Micrococcal Nuclease Sequencing

One million FANs-isolated nuclei were collected directly into MN buffer, centrifuged at 1500 rpm for 20 minutes at 4 °C, the majority of the supernatant removed, and the pellet was gently resuspended in the remaining buffer. The sample was transferred to a siliconised micro-centrifuge tube, and centrifuged for 10 minutes at 1500 rpm, at 4 °C. The supernatant was removed to the level of the pellet to ensure the nuclei were retained, and the pellet was resuspended in 25 μL of buffer MN. To the resuspended sample, 0.3 units of MNase was added, and the sample was incubated at 37 °C for 10 mins with gentle shaking (300 rpm). The samples were transferred to ice and 4 μL

of stop solution was added (section 6.6). Samples were incubated at 65 °C, RNase-treated, proteinase-K treated and run on an agarose gel as described in section 6.3.6. The mono-nucleosomal DNA fragments in the gel were excised under low intensity UV light, using a clean scalpel. The gel fragments were then cut into small pieces and the DNA eluted from the gel using freeze and squeeze gel extraction columns (BioRad). DNA was then purified using minelute column DNA purification kit (Qiagen) following the manufacturer's guidlines with the following exceptions: 1 mL of buffer PB was used for binding the DNA to the column, the columns were washed twice with buffer PE, the columns were spun dry for two minutes, and the columns were left open at room temperature for 2 minutes to remove excess ethanol. The DNA was eluted using 12 μL elution buffer, leaving the sample for 5 minutes before centrifugation. The eluted material was passed over the column a second time to increase the yeild of DNA eluted. The concentration of the purified DNA was measured using the Qubit dsDNA high sensitivity assay (life technologies). A total of 10 ng was sent for sequencing at the EMBL genecore facility (Heidelberg), who also prepared the libraries. samples were sequenced with 50 bp paired-end reads, initially one sample per sequencing lane, then the neuron and glia specific samples were sequenced again, with the same DNA but new library preparation, by multiplexing the four samples together on two lanes. This resulted in approximately 500 million reads for each cell type.

6.3.8 Nuclear-RNA seq

Heads were collected as previously described (section 6.3.2) from 20 mL of frozen flies, and ground to a fine powder on dry ice using a mortar and pestle. The powdered sample was then added to 50 mL of NPB (section 6.6) that contained 250 μL RNaseIn (Promega). Note that no protease inhibitor complex was added to the buffer. The samples were then homogenised 30 times using a 100 mL Dounce homogenizer, at 4 °C, then filtered sequentially through 60 μm and 10 μm filters (Millipore) that were pre-wet with NPB + 5 % BSA. Nuclei were pelleted by centrifuging at 2000 rpm, for five minutes, at 4 °C. The supernatant was carefully removed, then the pellet of nuclei were gently re-suspended with 1 mL of NPB + 5 % BSA + RNaseIN (200 units/mL). The nuclei were centrifuged at 1000 rpm for one minute, the supernatant removed and the nuclei again resuspended in 1 mL of NBP+BSA+RNasIN buffer. This washing step was repeated for a total of five times to ensure maximal removal of RNA from the outside of nuclei, whilst retaining the intergrity of the nuclei. GFP-positive nuclei were then collected using FANs (section 6.3.3). Nuclei were collected into 1 mL of NPB, containing 2 X concentrated RNaseIN. Samples of 0.5 million nuclei were collected sequentially for up to four hours. This ensured that high-population neuron isolations

were within the same experimental time-frame as glia isolations. Samples were kept on ice until sorting was finished.

Samples were then centrifuged at 2000 rpm, 4 °C, for 15 minutes. The supernatant was carefully removed until approximately 500 μL of liquid was left. The nuclei pellet was then resuspended in the remaining supernatant and transferred to a low-binding micro-centrifuge tube (Costar). The samples were then centrifuged at 2000 rpm, 4 °C, for 5 minutes and as much of the supernatant was removed as possible. During the centrifugation the lysis buffer was prepared from the Agencourt RNAdvance tissue kit (Agencourt), 400 μL of which was immediately added to the nuclei pellet after the supernatant was removed. The samples were incubated at 37 °C, for 25 minutes, then transferred to -80 °C for storage. Shortly after this, the RNA was isolated from the samples following the instructions of the Agencourt RNAdvance kit protocol, including the DNase digestion step. RNA concentration was measured using Quibit RNA detection kit, and quality of the RNA was tested using a bioanalyzer (Agilient technologies).

The Ovation human FFPE RNA-seq library preparation kit (Nugen) was used to prepare directional sequencing libraries for 5 ng of each RNA sample, as per kit directions. A custom-designed primer set was used to target the ribosomal RNA for degradation. The library was amplified using 15 PCR cycles and the libraries were purified an additional time to remove primer-dimers from the amplification. Libraries were sequenced by the EMBL genecore facility using an Illumina hiseq 2000 machine, 100 bp paired-end sequencing.

6.4 Immunohistochemistry of *Drosophila* brains

Flies were asphyxiated with CO_2, and placed in 70 % ethanol for several minutes to remove air pockets. Flies were then transferred to PBS (section 6.6) and the brains dissected. Dissected brains were incubated in 200 μL PBST (PBS, 0.3 % Triton-X100) and 4 % formaldehyde, at room temperature for 20 minutes to cross-link. The brains were then rinsed two times with 200 μL PBST, and then washed three times, for 20 minutes each, with rotation in 200 μL PBST. The dissected brains were then blocked with PBST + 5 % foetal calf serum and 5 % milk, with rotation at room temperature for 30 minutes. Primary antibody was then added at a concentration of 1:200, in PBST + 5 % foetal calf serum, and the samples incubated for two nights at 4 °C with rotation. The two quick wash steps and three longer wash steps were repeated three times, then the secondary antibody was added as 1:500 dilutions and the sample was again

incubated for 2 nights at 4 °C with rotation. The five wash steps were repeated as before, except that Hoechst was added for the first 20 minute wash step at a dilution of 1:1000. Glass slides were prepared with hole-punch reinforcement stickers to generate a shallow well. The immunostained brains were then transferred to the well and a drop of focus-clear (CelExplorer) was added to the well. Coverslips were then placed over the slides and sealed with nail polish. Images were obtained of the immunostained brains were captured with a confocal laser-scanning microscope (LSM-710, Zeiss). Image data processing was performed using Fiji (ImageJ 1.48r, http://imajej.nih.gov.ij.

6.5 Bioinformatics analysis

6.5.1 Mapping and data transformation

Mapping of raw sequencing reads to the Drosophila genome was performed by Pawel Bednarz (University of Warsaw), using the mapping tool bowtie (145). Sorting the mapped bam files was performed using the samtools sort function of samtools. Peaks were called using MACs (179). To find genes associated with peak calls, for example the H2AZ genes, defined in a bed file, the following command line was used using bedtools: intersectBed -a example.bed -b refSeq.bed -wb. To obtain the coverage of reads over certain annotated regions, as defined by a .bed file, the following command line was used: coverageBed a example.bed -b Data/genome/Drosophila_genome.gff > example_coverage.tsv

6.5.2 Analysis in R

The statistical package R was used for the majority of simple data manipulation and analysis.

Venn diagrams

The Vennerables package was used to generate the weighted venn diagrams (see http://r-forge.r-project.org/projects/vennerable). Comparisons were made between flybase gene numbers identified in the three method types.

Scatterplots and linear regression

Scatterpots were generated in R using the scatterplot function. Linear regression was calculated using the lm function.

Generating TSS and genebody bed files

The following code was used in R to define 500 bp around the TSS of all annotated genes:

```
Genes<- read.csv“gene_annotation”, sep = “\t”, header=FALSE)
Genes<- subset (Genes, select =c(1,2,3,4,6))
colnames(Genes)<- c(“chr_Name”,“start_pos”, “start_min”, “gene”, “strand”)

    NewDoc=matrix(FALSE, ncol=4, nrow=length(Genes$chr_Name))
colnames(NewDoc)=c(“chr”, “Start”, “Stp”, “gene”)
NewDoc=as.data.frame(NewDoc)

    for(i in 1:length(Genes$chr_Name)){ if(Genes$strand[i]==“+”){
middle=Genes$start_pos[i]}
else{middle=Genes$start_min[i]}

    NewDoc$Start[i]=middle-250
NewDoc$Stp[i]=middle+250
NewDoc$chr[i]=as.character(Genes$chr_Name[i])
NewDoc$gene[i]=as.character(Genes$gene[i])
```

The following code was used to define the genebody region for Pol II pausing analysis:

```
NewDoc=matrix(FALSE, ncol=5, nrow=length(Genes$chr_Name))
colnames(NewDoc)=c(“chr”, “Start”, “Stp”, “gene”,“remove”)
NewDoc=as.data.frame(NewDoc)

    for(i in 1:length(Genes$chr_Name)){
if(Genes$strand[i]==“+”){
strt=Genes$start_pos[i]+500}
else{strt=Genes$start_pos[i]}

    if(Genes$strand[i]==“+”){
stp=Genes$start_min[i]}
else{stp=Genes$start_min[i]-500}

    NewDoc$Start[i]=strt
NewDoc$Stp[i]=stp
```

```
NewDoc$chr[i]=as.character(Genes$chr_Name[i])
NewDoc$gene[i]=as.character(Genes$gene[i])
NewDoc$remove[i]=strt>stp}
write.table(NewDoc_sub, "genebody.txt", sep="\t", col.names=F, quote=F, row.names=F)
```

To remove genes where the length of the gene is less than 500 bp:

```
NewDoc_sub<- subset(NewDoc,remove==FALSE)
NewDoc_sub<-subset(NewDoc_sub, select=c(1,2,3,4))
write.table(NewDoc_sub, "genebody.txt", sep="\t", col.names=F, quote=F, row.names=F)
```

Generating fasta files for promoter regions

A file with the sequence 200 bp around the TSS of all annotated genes was generated using the following code:

```
library("seqinr")
Name_position=read.csv("ref_TSS_gene", sep = "\t", header=FALSE)
head(Name_position)
Name_position=subset(Name_position, select=c(1,2,3,4,6))
colnames(Name_position)=c("chr_Name","start_pos","start_min", "gene","strand")
```

```
fasta=read.fasta(file ="/Drosophila_melanogaster.BDGP5.69.fa", seqtype = "AA",
as.string = FALSE, forceDNAtolower = TRUE,set.attributes = TRUE, legacy.mode =
TRUE, seqonly = FALSE, strip.desc = FALSE) NewDoc3=matrix(FALSE, ncol = 4,
nrow = length(Name_position$chr_Name))
colnames(NewDoc3)=c("chr","sequence", "gene", "strand")
NewDoc3=as.data.frame(NewDoc3)
for(i in 1:length(Name_position$chr_Name)){
a=sub(">", "", Name_position$chr_Name[i])
x=which(names(fasta)==a)
```

```
if(Name_position$strand[i]=="+"){
middle=(as.numeric(as.character(Name_position$start_pos[i])))}
else{middle=(as.numeric(as.character(Name_position$start_min[i])))}
```

```
start=middle-100
stp=middle+100
```

```
    NewDoc3$sequence[i]=paste(fasta[[x]][start:stp],collapse=")
NewDoc3$chr[i]=as.character(Name_position$chr_Name[i])
NewDoc3$gene[i]=as.character(Name_position$gene[i])
NewDoc3$strand[i]=as.character(Name_position$strand[i])}
```

An example of generating the fasta file of the neuronal gene promoter sequence is shown in the following code:

```
library("Hmisc") indexes=c() for(i in 1:length(NewDoc$chr)){
if (NewDoc$gene[i]%nin% invariant$V1){indexes=append(indexes,i)}}

    invariant_seq <- NewDoc[-indexes,]
indexes<-c()
for(i in 1:length(NewDoc$chr)){
if (NewDoc$gene[i]%nin% neuronal$V1){indexes=append(indexes,i)}}

    neuronal_seq <- NewDoc[-indexes,]
```

Calculating Pol II pausing index

The number of counts from head RPB3 ChIP-seq replicates were calculated using bedtools (section 6.5.1) over the TSS.bed and the genebody.bed files generated above. The pausing index was then calculated for each replicate separately, using the following code in R:

First, the data were normalised for each replicate to generate coverage per base pair over each region. An example of this calculation is shown below:

```
TSS_RPB3A<- read.table("RPB3_Fed_6h_A_TSS.txt")
head(TSS_RPB3A) colnames(TSS_RPB3A)<- c("chr","start","stop","gene","depth",
"bases_at_depth", "length","fraction")

    depth_per_bp<-TSS_RPB3A$depth/TSS_RPB3A$length
head(depth_per_bp)

    NewDoc<-matrix(FALSE, ncol=7,nrow=length(TSS_RPB3A$gene)) colnames(NewDoc)<-
c("chr","start","stop","gene","depth", "length", "depth_per_bp")
NewDoc<- as.data.frame(NewDoc)
```

```
    NewDoc$chr<-TSS_RPB3A$chr
NewDoc$start<-TSS_RPB3A$start
NewDoc$stop<-TSS_RPB3A$stop
NewDoc$gene<-TSS_RPB3A$gene
NewDoc$depth<-TSS_RPB3A$depth
NewDoc$length<-TSS_RPB3A$length
NewDoc$depth_per_bp<-depth_per_bp
TSS_RPB3A_normalised<-(NewDoc)
write.table(TSS_RPB3A_normalised, "TSS_RPB3_normalised.txt", sep="\t", col.names=T,
row.names=F, quote=F)
```

Then, the ratio between the Pol II binding at the TSS and the genebody were calculated using the following code:

```
NewDoc<-matrix(FALSE, ncol=5, nrow=length(TSS_RPB3A_normalised$chr))
colnames(NewDoc)<-c("chr", "gene","depth_TSS","depth_GB", "PI")
NewDoc<-as.data.frame(NewDoc)

    NewDoc$chr<-TSS_RPB3A_normalised$chr
NewDoc$gene<-TSS_RPB3A_normalised$gene
NewDoc$depth_TSS<-TSS_RPB3A_normalised$depth_per_bp
NewDoc$depth_GB<-GB_RPB3A_normalised$depth_per_bp
NewDoc$PI<-NewDoc$depth_TSS/NewDoc$depth_GB

    RPB3_A_PI<-NewDoc write.table(RPB3_A_PI, "RPB3_A_PI.txt", sep="\t", col.names=T,
row.names=F, quote=F)
```

The average ratio between the replicates was calculated and used for ranking genes by pausing index.

6.5.3 Average profile plots and heatmaps

Average profile plots and heatmaps were generated using the tool NGS-plot (158). Average profile plots were made as comparable as possible by using the configuration file option. This file allows mapping of the specified data into the same plot, for example the neuron and glia nucRNA samples split by neuron, glia, and invariant gene classes were within the same plot.

6.5.4 Sequencing enrichment analysis

The analysis of nucleotide enrichment across promoter regions was performed with the online tool Weblogo3 (167). Fasta files generated from each gene class (section 6.5.2) were used to generate the graphs.

6.6 Buffers

LiCl buffer
250 mM LiCl, 10 mM Tris-Hcl (pH 8), 1 mM EDTA, 0.5 % NP-40, 0.5 % DOC. Protease inhibitor complex and PMSF were added fresh on the day of use.
MN buffer
60 mM KCl, 15 mM NaCl, 15 mM Tris (pH7.4), 0.5 mM DTT, 0.25 mM sucrose, 1 mM CaCl.
Nuclei-purification buffer (NPB), pH 7
20 mM MOPS, 40 mM NaCl, 90 mM KaCl,2 mM EDTA, 0.5 mM EGTA, 0.1 % NP-40. Stored at 4 °C for up to three months. 0.5 mM spermidine, 0.2 mM spermine, ad protease-inhibitor complex were freshly added on the day of use.
PBS
137 mM NaCl, 2.7 mM KCl, 10 mM Na2HPO4, 1.8 mM KH2PO4.
RIPA buffer
150 mM NaCl, 25 mM HEPES (pH 7.5), 1 mM EDTA, 1 % Triton-X 100, 0.1 % SDS, 0.1% DOC. Protease inhibitor complex and PMSF were added fresh on the day of use.
10 X TBS/ TBST
0.2 M Tris (pH 7.5), 1.2 M NaCl. TBST was made from 1 X TBS + 5 mL 10 % Tween-20.
TE buffer
10 mM Tris-HCl (pH 8), 1 mM EDTA.

References

[1] R G Roeder and W J Rutter. **Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms.** *Nature*, **224**(5216):234–237, October 1969. 2

[2] M Gniazdowski, J L Mandel, Jr, F Gissinger, C Kedinger, and P Chambon. **Calf thymus RNA polymerases exhibit template specificity**. *Biochem Biophys Res Commun*, **38**(6):1033–40, Mar 1970. 2

[3] C Kedinger, M Gniazdowski, J L Mandel, Jr, F Gissinger, and P Chambon. **Alpha-amanitin: a specific inhibitor of one of two DNA-pendent RNA polymerase activities from calf thymus**. *Biochem Biophys Res Commun*, **38**(1):165–71, Jan 1970. 2

[4] R Weinmann, H J Raskas, and R G Roeder. **Role of DNA-dependent RNA polymerases II and III in transcription of the adenovirus genome late in productive infection**. *Proc Natl Acad Sci U S A*, **71**(9):3426–39, Sep 1974. 2

[5] R Weinmann and R G Roeder. **Role of DNA-dependent RNA polymerase 3 in the transcription of the tRNA and 5S RNA genes**. *Proc Natl Acad Sci U S A*, **71**(5):1790–4, May 1974. 2

[6] Tatsuo Kanno, Bruno Huettel, M Florian Mette, Werner Aufsatz, Estelle Jaligot, Lucia Daxinger, David P Kreil, Marjori Matzke, and Antonius J M Matzke. **Atypical RNA polymerase subunits required for RNA-directed DNA methylation.** *Nature genetics*, **37**(7):761–765, July 2005. 2

[7] A J Herr, M B Jensen, T Dalmay, and D C Baulcombe. **RNA polymerase IV directs silencing of endogenous DNA.** *Science (New York, NY)*, **308**(5718):118–120, April 2005. 2

[8] Yasuyuki Onodera, Jeremy R Haag, Thomas Ream, Pedro Costa Nunes, Olga Pontes, and Craig S Pikaard. **Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation.** *Cell*, **120**(5):613–622, March 2005. 2

[9] Julia E Kravchenko, Igor B Rogozin, Eugene V Koonin, and Peter M Chumakov. **Transcription of mammalian messenger RNAs by a nuclear RNA polymerase of mitochondrial origin.** *Nature*, **436**(7051):735–739, August 2005. 2

[10] P A Weil, D S Luse, J Segall, and R G Roeder. **Selective and accurate initiation of transcription at the Ad2 major late promotor in a soluble system dependent on purified RNA polymerase II and DNA.** *Cell*, **18**(2):469–484, October 1979. 2

[11] P A Weil, J Segall, B Harris, S Y Ng, and R G Roeder. **Faithful transcription of eukaryotic genes by RNA polymerase III in systems reconstituted with purified DNA templates.** *The Journal of biological chemistry*, **254**(13):6163–6173, July 1979. 2

[12] Mary C Thomas and Cheng-Ming Chiang. **The general transcription machinery and general cofactors.** *Critical reviews in biochemistry and molecular biology*, **41**(3):105–178, May 2006. 2, 3

[13] T Matsui, J Segall, P A Weil, and R G Roeder. **Multiple factors required for accurate initiation of transcription by purified RNA polymerase II.** *The Journal of biological chemistry*, **255**(24):11992–11996, December 1980. 2

[14] Benjamin L Allen and Dylan J Taatjes. **The Mediator complex: a central integrator of transcription.** *Nature reviews Molecular cell biology*, **16**(3):155–166, March 2015. 3

[15] David Ries and Michael Meisterernst. **Control of gene transcription by Mediator in chromatin**. *Seminars in cell & developmental biology*, August 2011. 3

[16] I Jonkers, H Kwak, and J T Lis. **Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons**. *eLife*, **3**(0):e02407–e02407, January 2014. 3

[17] Daniel A Gilchrist, Gilberto Dos Santos, David C Fargo, Bin Xie, Yuan Gao, Leping Li, and Karen Adelman. **Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation**. *Cell*, **143**(4):540–551, November 2010. 3, 129

[18] Christopher M Weber, Srinivas Ramachandran, and Steven Henikoff. **Nucleosomes Are Context-Specific, H2A.Z-Modulated Barriers to RNA Polymerase**. *Molecular cell*, **53**(5):819–830, March 2014. 3, 9, 102, 124, 130

[19] D S Gilmour and J T Lis. **RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells.** *Molecular and Cellular Biology*, **6**(11):3984–3989, November 1986. 3

[20] A E Rougvie and J T Lis. **Postinitiation transcriptional control in Drosophila melanogaster.** *Molecular and Cellular Biology*, 1990. 3

[21] L J Core, J J Waterfall, and J T Lis. **Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters**. *Science (New York, NY)*, **322**(5909):1845–1848, December 2008. 3

[22] Karen Adelman and John T Lis. **Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans.** *Nature reviews Genetics*, **13**(10):720–731, October 2012. 3

[23] T Fujita, S Ryser, I Piuz, and W Schlegel. **Up-Regulation of P-TEFb by the MEK1-Extracellular Signal-Regulated Kinase Signaling Pathway Contributes to Stimulated Transcription Elongation of Immediate Early Genes in Neuroendocrine Cells**. *Molecular and Cellular Biology*, **28**(5):1630–1643, March 2008. 3

[24] Ramendra N Saha, Erin M Wissink, Emma R Bailey, Meilan Zhao, David C Fargo, Ji-Yeon Hwang, Kelly R Daigle, J Daniel Fenn, Karen Adelman, and Serena M Dudek. **Rapid activity-induced transcription of Arc and other IEGs relies on poised RNA polymerase II**. *Nature Publishing Group*, **14**(7):848–856, May 2011. 3

[25] Steven J Petesch and John T Lis. **Overcoming the nucleosome barrier during transcript elongation.** *Trends in Genetics*, **28**(6):285–294, June 2012. 4

[26] Curt A Davey, David F Sargent, Karolin Luger, Armin W Maeder, and Timothy J Richmond. **Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution.** *Journal of molecular biology*, **319**(5):1097–1113, June 2002. 5

REFERENCES

[27] D Rhodes. **Nucleosome cores reconstituted from poly (dA-dT) and the octamer of histones.** *Nucleic Acids Research*, **6**(5):1805–1816, 1979. 6

[28] R T Simpson and P Künzler. **Cromatin and core particles formed from the inner histones and synthetic polydeoxyribonucleotides of defined sequence.** *Nucleic Acids Research*, **6**(4):1387–1415, April 1979. 6

[29] W Linxweiler and W Hörz. **Reconstitution of mononucleosomes: characterization of distinct particles that differ in the position of the histone core.** *Nucleic Acids Research*, **12**(24):9395–9413, December 1984. 6

[30] P C FitzGerald and R T Simpson. **Effects of sequence alterations in a DNA segment containing the 5 S RNA gene from Lytechinus variegatus on positioning of a nucleosome core particle in vitro.** *The Journal of biological chemistry*, **260**(28):15318–15324, December 1985. 6

[31] W Linxweller and W Hörz. **Reconstitution experiments show that sequence-specific histone-DNA interactions are the basis for nucleosome phasing on mouse satellite DNA.** *Cell*, **42**(1):281–290, August 1985. 6

[32] Z Zhang, C J Wippo, M Wal, E Ward, P Korber, and B F Pugh. **A Packing Mechanism for Nucleosome Organization Reconstituted Across a Eukaryotic Genome**. *Science (New York, NY)*, **332**(6032):977–980, May 2011. 6, 7

[33] Julia Pointner, Jenna Persson, Punit Prasad, Ulrika Norman-Axelsson, Annelie Str aring lfors, Olga Khorosjutina, Nils Krietenstein, J Peter Svensson, Karl Ekwall, and Philipp Korber. **CHD1 remodelers regulate nucleosome spacing in vitro and align nucleosomal arrays over gene coding regions in S. pombe**. *The EMBO journal*, **31**(23):4388–4403, October 2012. 6, 130

[34] Tamás Schauer, Petra C Schwalie, Ava Handley, Carla E Margulies, Paul Flicek, and Andreas G Ladurner. **CAST-ChIP Maps Cell-Type-Specific Chromatin States in the Drosophila Central Nervous System.** *CellReports*, October 2013. 8, 16, 25, 28, 30, 32, 36, 46, 77, 102, 116, 120, 122, 130

[35] T H Thatcher and M A Gorovsky. **Phylogenetic analysis of the core histones H2A, H2B, H3, and H4.** *Nucleic Acids Research*, **22**(2):174–179, January 1994. 8

[36] R K Suto, M J Clarkson, D J Tremethick, and K Luger. **Crystal structure of a nucleosome core particle containing the variant histone H2A.Z.** *Nature structural biology*, **7**(12):1121–1124, December 2000. 8, 9

[37] A van Daal and S C Elgin. **A histone variant, H2AvD, is essential in Drosophila melanogaster.** *Molecular biology of the cell*, **3**(6):593–602, June 1992. 8

[38] Christophe Redon, Duane Pilch, Emmy Rogakou, Olga Sedelnikova, Kenneth Newrock, and William Bonner. **Histone H2A variants H2AX and H2AZ.** *Current opinion in genetics & development*, **12**(2):162–169, April 2002. 8

[39] M J Clarkson, J R Wells, F Gibson, R Saint, and D J Tremethick. **Regions of variant histone His2AvD required for Drosophila development.** *Nature*, **399**(6737):694–697, June 1999. 8, 9

[40] T J Leach, M Mazzeo, H L Chotkowski, J P Madigan, M G Wotring, and R L Glaser. **Histone H2A.Z is widely but nonrandomly distributed in chromosomes of Drosophila melanogaster.** *The Journal of biological chemistry*, **275**(30):23267–23272, July 2000. 9

[41] Travis N Mavrich, Cizhong Jiang, Ilya P Ioshikhes, Xiaoyong Li, Bryan J Venters, Sara J Zanton, Lynn P Tomsho, Ji Qi, Robert L Glaser, Stephan C Schuster, David S Gilmour, Istvan Albert, and B Franklin Pugh. **Nucleosome organization in the Drosophila genome.** *Nature*, **453**(7193):358–362, May 2008. 9, 102

[42] Zhenhai Zhang and B Franklin Pugh. **Genomic organization of H2Av containing nucleosomes in Drosophila heterochromatin.** *PloS one*, **6**(6):e20511, 2011. 9

[43] Gaku Mizuguchi, Xuetong Shen, Joe Landry, Wei-Hua Wu, Subhojit Sen, and Carl Wu. **ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex.** *Science (New York, NY)*, **303**(5656):343–348, January 2004. 9

[44] Anand Ranjan, Gaku Mizuguchi, Peter C Fitzgerald, Debbie Wei, Feng Wang, Yingzi Huang, Ed Luk, Christopher L Woodcock, and Carl Wu. **Nucleosome-free region dominates histone acetylation in targeting SWR1 to promoters for H2A.Z replacement.** *Cell*, **154**(6):1232–1245, September 2013. 9

[45] Paul D Hartley and Hiten D Madhani. **Mechanisms that specify promoter nucleosome location and identity.** *Cell*, **137**(3):445–458, May 2009. 9

[46] Chunyuan Jin and Gary Felsenfeld. **Nucleosome stability mediated by histone variants H3.3 and H2A.Z.** *Genes & Development*, **21**(12):1519–1529, June 2007. 9

[47] S Henikoff, J G Henikoff, A Sakai, G B Loeb, and K Ahmad. **Genome-wide profiling of salt fractions maps physical properties of chromatin**. *Genome Research*, **19**(3):460–469, December 2008. 9

[48] Young-Jun Park, Pamela N Dyer, David J Tremethick, and Karolin Luger. **A new fluorescence resonance energy transfer approach demonstrates that the histone variant H2AZ stabilizes the histone octamer within the nucleosome.** *The Journal of biological chemistry*, **279**(23):24274–24282, June 2004. 9

[49] Jiaxu Li, Danesh Moazed, and Steven P Gygi. **Association of the histone methyltransferase Set2 with RNA polymerase II plays a role in transcription elongation.** *The Journal of biological chemistry*, **277**(51):49383–49388, December 2002. 10

[50] Bing Li, LeAnn Howe, Scott Anderson, John R Yates, and Jerry L Workman. **The Set2 histone methyltransferase functions through the phosphorylated carboxyl-terminal domain of RNA polymerase II.** *The Journal of biological chemistry*, **278**(11):8897–8903, March 2003. 10

[51] Michaela Smolle, Swaminathan Venkatesh, Madelaine M Gogol, Hua Li, Ying Zhang, Laurence Florens, Michael P Washburn, and Jerry L Workman. **Chromatin remodelers Isw1 and Chd1 maintain chromatin structure during transcription by preventing histone exchange.** *Nature Structural & Molecular Biology*, **19**(9):884–892, September 2012. 10, 127

[52] Feng Tie, Rakhee Banerjee, Carl A Stratton, Jayashree Prasad-Sinha, Vincent Stepanik, Andrei Zlobin, Manuel O Diaz, Peter C Scacheri, and Peter J Harte. **CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing**. *Development (Cambridge, England)*, **136**(18):3131–3141, September 2009. 10

[53] Michael J Carrozza, Bing Li, Laurence Florens, Tamaki Suganuma, Selene K Swanson, Kenneth K Lee, Wei-Jong Shia, Scott Anderson, John Yates, Michael P Washburn, and Jerry L Workman. **Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription.** *Cell*, **123**(4):581–592, November 2005. 10

[54] John W Edmunds, Louis C Mahadevan, and Alison L Clayton. **Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation.** *The EMBO journal*, **27**(2):406–420, January 2008. 10

[55] Sílvia Carvalho, Ana Cláudia Raposo, Filipa Batalha Martins, Ana Rita Grosso, Sreerama Chaitanya Sridhara, José Rino, Maria Carmo-Fonseca, and Sérgio Fernandes de Almeida. **Histone methyltransferase SETD2 coordinates FACT recruitment with nucleosome dynamics during transcription.** *Nucleic Acids Research*, **41**(5):2881–2893, March 2013. 10

[56] Jeremy M Simon, Kathryn E Hacker, Darshan Singh, A Rose Brannon, Joel S Parker, Matthew Weiser, Thai H Ho, Pei-Fen Kuan, Eric Jonasch, Terrence S Furey, Jan F Prins, Jason D Lieb, W Kimryn Rathmell, and Ian J Davis. **Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects.** *Genome Research*, **24**(2):241–250, February 2014. 10

[57] Sérgio Fernandes de Almeida, Ana Rita Grosso, Frederic Koch, Romain Fenouil, Sílvia Carvalho, Jorge Andrade, Helena Levezinho, Marta Gut, Dirk Eick, Ivo Gut, Jean-Christophe Andrau, Pierre Ferrier, and Maria Carmo-Fonseca. **Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36.** *Nature Structural & Molecular Biology*, **18**(9):977–983, September 2011. 10

[58] Loredana Verdone, Micaela Caserta, and Ernesto Di Mauro. **Role of histone acetylation in the control of gene expression.** *Biochemistry and Cell Biology*, **83**(3):344–353, June 2005. 10

[59] Peter Tessarz and Tony Kouzarides. **Histone core modifications regulating nucleosome structure and dynamics.** *Nature reviews Molecular cell biology*, **15**(11):703–708, November 2014. 10

[60] V V Ogryzko, R L Schiltz, V Russanova, B H Howard, and Y Nakatani. **The transcriptional coactivators p300 and CBP are histone acetyltransferases.** *Cell*, **87**(5):953–959, November 1996. 10

[61] A J Bannister and T Kouzarides. **The CBP coactivator is a histone acetyltransferase.** *Nature*, **384**(6610):641–643, December 1996. 10

[62] A Imhof, X J Yang, V V Ogryzko, Y Nakatani, A P Wolffe, and H Ge. **Acetylation of general transcription factors by histone acetyltransferases.** *Current biology : CB*, **7**(9):689–692, September 1997. 10

[63] V Perissi, J S Dasen, R Kurokawa, Z Wang, E Korzus, D W Rose, C K Glass, and M G Rosenfeld. **Factor-specific modulation of CREB-binding protein acetyltransferase activity.** *Proceedings of the National Academy of Sciences of the United States of America*, **96**(7):3652–3657, March 1999. 11

[64] B L Kee, J Arias, and M R Montminy. **Adaptor-mediated recruitment of RNA polymerase II to a signal-dependent activator.** *The Journal of biological chemistry*, **271**(5):2373–2375, February 1996. 11

[65] Qihuang Jin, Li Rong Yu, Lifeng Wang, Zhijing Zhang, Lawryn H Kasper, Ji Eun Lee, Chaochen Wang, Paul K Brindle, Sharon YR Dent, and Kai Ge. **Distinct roles of GCN5/PCAFmediated H3K9ac and CBP/p300mediated H3K18/27ac in nuclear receptor transactivation**. *The EMBO journal*, **30**(2):249–262, 2011. 11

[66] Timothy R O'Connor and Timothy L Bailey. **Creating and validating cis-regulatory maps of tissue-specific gene expression regulation.** *Nucleic Acids Research*, **42**(17):11000–11010, 2014. 11

[67] M P Creyghton, A W Cheng, G G Welstead, T Kooistra, B W Carey, E J Steine, J Hanna, M A Lodato, G M Frampton, P A Sharp, L A Boyer, R A Young, and R Jaenisch. **From the Cover: Histone H3K27ac separates active from poised enhancers and predicts developmental state**. *Proceedings of the National Academy of Sciences*, **107**(50):21931–21936, December 2010. 11

[68] Justin Cotney, Jing Leng, Sunghee Oh, Laura E DeMare, Steven K Reilly, Mark B Gerstein, and James P Noonan. **Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb.** *Genome Research*, **22**(6):1069–1080, June 2012. 11

[69] Raphaël Margueron and Danny Reinberg. **The Polycomb complex PRC2 and its mark in life.** *Nature*, **469**(7330):343–349, January 2011. 11

[70] Jeffrey A Simon and Robert E Kingston. **Occupying chromatin: Polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put.** *Molecular cell*, **49**(5):808–824, March 2013. 11

[71] Klaus H Hansen and Kristian Helin. **Epigenetic inheritance through self-recruitment of the polycomb repressive complex 2.** *Epigenetics*, **4**(3):133–138, April 2009. 11

[72] Frank W Schmitges, Archana B Prusty, Mahamadou Faty, Alexandra Stützer, Gondichatnahalli M Lingaraju, Jonathan Aiwazian, Ragna Sack, Daniel Hess, Ling Li, Shaolian Zhou, Richard D Bunker, Urs Wirth, Tewis Bouwmeester, Andreas Bauer, Nga Ly-Hartig, Kehao Zhao, Homan Chan, Justin Gu, Heinz Gut, Wolfgang Fischle, Jürg Müller, and Nicolas H Thomä. **Histone methylation by PRC2 is inhibited by active chromatin marks.** *Molecular cell*, **42**(3):330–341, May 2011. 12

[73] Ling Cai, Scott B Rothbart, Rui Lu, Bowen Xu, Wei-Yi Chen, Ashutosh Tripathy, Shira Rockowitz, Deyou Zheng, Dinshaw J Patel, C David Allis, Brian D Strahl, Jikui Song, and Gang Greg Wang. **An H3K36 methylation-engaging Tudor motif of polycomb-like proteins mediates PRC2 complex targeting.** *Molecular cell*, **49**(3):571–582, February 2013. 12

[74] Ina Kycia, Srikanth Kudithipudi, Raluca Tamas, Goran Kungulovski, Arunkumar Dhayalan, and Albert Jeltsch. **The Tudor domain of the PHD finger protein 1 is a dual reader of lysine trimethylation at lysine 36 of histone H3 and lysine 27 of histone variant H3t.** *Journal of molecular biology*, **426**(8):1651–1660, April 2014. 12

[75] Peter V Kharchenko, Artyom A Alekseyenko, Yuri B Schwartz, Aki Minoda, Nicole C Riddle, Jason Ernst, Peter J Sabo, Erica Larschan, Andrey A Gorchakov, Tingting Gu, Daniela Linder-Basso, Annette Plachetka, Gregory Shanower, Michael Y Tolstorukov, Lovelace J Luquette, Ruibin Xi, Youngsook L Jung, Richard W Park, Eric P Bishop, Theresa P Canfield, Richard Sandstrom, Robert E Thurman, David M Macalpine, John A Stamatoyannopoulos, Manolis Kellis, Sarah C R Elgin, Mitzi I Kuroda, Vincenzo Pirrotta, Gary H Karpen, and Peter J Park. **Comprehensive analysis of the chromatin landscape in Drosophila melanogaster**. *Nature*, pages 1–7, December 2010. 12

[76] Guillaume J Filion, Joke G van Bemmel, Ulrich Braunschweig, Wendy Talhout, Jop Kind, Lucas D Ward, Wim Brugman, Inês J de Castro, Ron M Kerkhoven, Harmen J Bussemaker, and Bas Van Steensel. **Systematic protein location mapping reveals five principal chromatin types in Drosophila cells**. *Cell*, **143**(2):212–224, October 2010. 12, 13, 16, 33, 69, 87

REFERENCES

[77] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, Manching Ku, Timothy Durham, Manolis Kellis, and Bradley E Bernstein. **Mapping and analysis of chromatin state dynamics in nine human cell types**. *Nature*, pages 1–9, March 2011. 12

[78] R David Hawkins, Gary C Hon, Chuhu Yang, Jessica E Antosiewicz-Bourget, Leonard K Lee, Que-Minh Ngo, Sarit Klugman, Keith A Ching, Lee E Edsall, Zhen Ye, Samantha Kuan, Pengzhi Yu, Hui Liu, Xinmin Zhang, Roland D Green, Victor V Lobanenkov, Ron Stewart, James A Thomson, and Bing Ren. **Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency**. *Cell research*, August 2011. 12

[79] Emily J Brown and Doris Bachtrog. **The chromatin landscape of Drosophila: comparisons between species, sexes, and chromosomes.** *Genome Research*, **24**(7):1125–1137, July 2014. 13

[80] Weisheng Wu, Yong Cheng, Cheryl A Keller, Jason Ernst, Swathi Ashok Kumar, Tejaswini Mishra, Christapher Morrissey, Christine M Dorman, Kuan-Bei Chen, Daniela Drautz, Belinda Giardine, Yoichiro Shibata, Lingyun Song, Max Pimkin, Gregory E Crawford, Terrence S Furey, Manolis Kellis, Webb Miller, James Taylor, Stephan C Schuster, Yu Zhang, Francesca Chiaromonte, Gerd A Blobel, Mitchell J Weiss, and Ross C Hardison. **Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration**. *Genome Research*, August 2011. 13

[81] Stefan Bonn, Robert P Zinzen, Charles Girardot, E Hilary Gustafson, Alexis Perez-Gonzalez, Nicolas Delhomme, Yad Ghavi-Helm, Bartek Wilczyński, Andrew Riddell, and Eileen E M Furlong. **Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development**. *Nature genetics*, pages 1–55, January 2012. 13, 28, 141

[82] A Rasim Barutcu, Andrew J Fritz, Sayyed K Zaidi, André J vanWijnen, Jane B Lian, Janet L Stein, Jeffrey A Nickerson, Anthony N Imbalzano, and Gary S Stein. **C-ing the Genome: A Compendium of Chromosome Conformation Capture Methods to Study Higher-Order Chromatin Organization**. *Journal of cellular physiology*, pages n/a–n/a, June 2015. 13

[83] Xiaobin Zheng, Youngjo Kim, and Yixian Zheng. **Identification of lamin-B-regulated chromatin regions based on chromatin landscapes**. *molbiolcell.org*, 2015. 14

[84] Caterina Strambio-De-Castillia, Mario Niepel, and Michael P Rout. **The nuclear pore complex: bridging nuclear transport and gene regulation**. *Nature reviews Molecular cell biology*, **11**(7):490–501, July 2010. 14

[85] Tom Sexton, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. **Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome**. *Cell*, pages 1–15, January 2012. 14, 16

[86] Yad Ghavi-Helm, Felix A Klein, Tibor Pakozdi, Lucia Ciglar, Daan Noordermeer, Wolfgang Huber, and Eileen E M Furlong. **Enhancer loops appear stable during development and are associated with paused polymerase.** *Nature*, July 2014. 14

[87] Jesse R Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, Yarui Diao, Jing Liang, Huimin Zhao, Victor V Lobanenkov, Joseph R Ecker, James A Thomson, and Bing Ren. **Chromatin architecture reorganization during stem cell differentiation.** *Nature*, **518**(7539):331–336, February 2015. 14

[88] Li Li, Xiaowen Lyu, Chunhui Hou, Naomi Takenaka, Huy Q Nguyen, Chin-Tong Ong, Caelin Cubeñas-Potts, Ming Hu, Elissa P Lei, Giovanni Bosco, Zhaohui S Qin, and Victor G Corces. **Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing.** *Molecular cell*, **58**(2):216–231, April 2015. 14

[89] Helen Pickersgill, Bernike Kalverda, Elzo De Wit, Wendy Talhout, Maarten Fornerod, and Bas Van Steensel. **Characterization of the Drosophila melanogaster genome at the nuclear lamina**. *Nature genetics*, **38**(9):1005–1014, September 2006. 14, 95

[90] Jason Blanton, Miklos Gaszner, and Paul Schedl. **Protein:protein interactions and the pairing of boundary elements in vivo.** *Genes & Development*, **17**(5):664–675, March 2003. 15

[91] Todd Schoborg and Mariano Labrador. **Expanding the roles of chromatin insulators in nuclear architecture, chromatin organization and genome function.** *Cellular and molecular life sciences : CMLS*, **71**(21):4089–4113, November 2014. 15

[92] T I Gerasimova, K Byrd, and V G Corces. **A chromatin insulator determines the nuclear localization of DNA.** *Molecular cell*, **6**(5):1025–1035, November 2000. 15, 16

[93] Alexey A Soshnev, Ryan M Baxley, J Robert Manak, Kai Tan, and Pamela K Geyer. **The insulator protein Suppressor of Hairy-wing is an essential transcriptional repressor in the Drosophila ovary.** *Development (Cambridge, England)*, **140**(17):3613–3623, September 2013. 16, 91

[94] Ashley M Wood, Kevin Van Bortle, Edward Ramos, Naomi Takenaka, Margaret Rohrbaugh, Brian C Jones, Keith C Jones, and Victor G Corces. **Regulation of Chromatin Organization and Inducible Gene Expression by a Drosophila Insulator**. *Molecular cell*, **44**(1):29–38, October 2011. 16, 92

[95] Leah H Matzat, Ryan K Dale, Nellie Moshkovich, and Elissa P Lei. **Tissue-specific regulation of chromatin insulator function.** *PLoS Genetics*, **8**(11):e1003069, 2012. 16

[96] Catarina C F Homem and Juergen A Knoblich. **Drosophila neuroblasts: a model for stem cell biology**. *dev.biologists.org*, 2012. 16

[97] A Prokop and G M Technau. **The origin of postembryonic neuroblasts in the ventral nerve cord of Drosophila melanogaster.** *Development (Cambridge, England)*, **111**(1):79–88, January 1991. 16

[98] T Lee, A Lee, and L Luo. **Development of the Drosophila mushroom bodies: sequential generation of three distinct types of neurons from a neuroblast.** *Development (Cambridge, England)*, **126**(18):4065–4076, September 1999. 17

[99] Tara N Edwards and Ian A Meinertzhagen. **The functional organisation of glia in the adult brain of Drosophila and other insects.** *Progress in neurobiology*, **90**(4):471–497, April 2010. 17

[100] Ava Handley, Tamás Schauer, Andreas G Ladurner, and Carla E Margulies. **Designing Cell-Type-Specific Genome-wide Experiments.** *Molecular cell*, **58**(4):621–631, May 2015. 18

[101] Chris M Hempel, Ken Sugino, and Sacha B Nelson. **A manual method for the purification of fluorescently labeled neurons from the mammalian brain.** *Nature Protocols*, **2**(11):2924–2929, 2007. 20

[102] Emi Nagoshi, Ken Sugino, Ela Kula, Etsuko Okazaki, Taro Tachibana, Sacha Nelson, and Michael Rosbash. **Dissecting differential gene expression within the circadian neuronal circuit of Drosophila.** *Nature neuroscience*, **13**(1):60–68, January 2010. 20

[103] Rebecca M Fox, Joseph D Watson, Stephen E Von Stetina, Joan Mcdermott, Thomas M Brodigan, Tetsunari Fukushige, Michael Krause, and David M Miller. **The embryonic muscle transcriptome of Caenorhabditis elegans**. *Genome Biology*, **8**(9):R188, January 2007. 20

[104] Christian Berger; Heike Harzer; ThomasnbspR Burkard; Jonas Steinmann; Suzanne vannbspdernbspHorst; Anne-Sophie Laurenson; Maria Novatchkova; Heinrich Reichert; JuergennbspA Knoblich, Heike Harzer, Thomas R Burkard, Jonas Steinmann, Suzanne Van Der Horst, Anne-Sophie Laurenson, Maria Novatchkova, Heinrich Reichert, and Juergen A Knoblich. **FACS Purification and Transcriptome Analysis of Drosophila Neural Stem Cells Reveals a Role for Klumpfuss in Self-Renewal**. *CellReports*, **2**(2):407–418, Aug 2012. 20, 95

[105] Virginia Espina, Julia D Wulfkuhle, Valerie S Calvert, Amy VanMeter, Weidong Zhou, George Coukos, David H Geho, Emanuel F Petricoin, and Lance A Liotta. **Laser-capture microdissection.** *Nature Protocols*, **1**(2):586–603, 2006. 21

[106] Liang Cheng, Shaobo Zhang, Gregory T MacLennan, Sean R Williamson, Darrell D Davidson, Mingsheng Wang, Timothy D Jones, Antonio Lopez-Beltran, and Rodolfo Montironi. **Laser-assisted microdissection in translational research: theory, technical considerations, and future applications.** *Applied immunohistochemistry & molecular morphology : AIMM / official publication of the Society for Applied Immunohistochemistry*, **21**(1):31–47, January 2013. 21

[107] B W Okaty, K Sugino, and S B Nelson. **Cell Type-Specific Transcriptomics in the Brain**. *Journal of Neuroscience*, **31**(19):6939–6943, May 2011. 22

[108] Simon Haenni, Zhe Ji, Mainul Hoque, Nigel Rust, Helen Sharpe, Ralf Eberhard, Cathy Browne, Michael O Hengartner, Jane Mellor, Bin Tian, and André Furger. **Analysis of C. elegans intestinal gene expression and polyadenylation by fluorescence-activated nuclei sorting and 3'-end-seq.** *Nucleic Acids Research*, **40**(13):6304–6318, July 2012. 22

[109] Yan Jiang, Anouch Matevossian, Hsien-Sung Huang, Juerg Straubhaar, and Schahram Akbarian. **Isolation of neuronal chromatin from brain tissue**. *BMC neuroscience*, **9**:42, January 2008. 22

[110] Stefan Bonn, Robert P Zinzen, Charles Girardot, E Hilary Gustafson, Alexis Perez-Gonzalez, Nicolas Delhomme, Yad Ghavi-Helm, Bartek Wilczyński, Andrew Riddell, and Eileen E M Furlong. **Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development**. *Nat Genet*, pages 1–11, Jan 2012. 22, 24

[111] Moritz J Rossner, Johannes Hirrlinger, Sven P Wichert, Christine Boehm, Dieter Newrzella, Holger Hiemisch, Gisela Eisenhardt, Carolin Stuenkel, Oliver von Ahsen, and Klaus-Armin Nave. **Global transcriptome analysis of genetically identified neurons in the adult cortex.** *Journal of Neuroscience*, **26**(39):9956–9966, September 2006. 22

[112] Roger B Deal and Steven Henikoff. **The INTACT method for cell type–specific gene expression and chromatin profiling in Arabidopsis thaliana**. *Nature Protocols*, **6**(1):56–68, January 2011. 23

[113] F. A Steiner, P. B Talbert, S Kasinathan, R. B Deal, and S Henikoff. **Cell type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling**. *Genome Research*, pages 1–30, Jan 2012. 23

[114] G. L Henry, F. P Davis, S Picard, and S. R Eddy. **Cell type-specific genomics of Drosophila neurons**. *Nucleic Acids Research*, pages 1–14, Aug 2012. 23, 28

[115] F A Steiner, P B Talbert, S Kasinathan, R B Deal, and S Henikoff. **Cell type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling**. *Genome Research*, pages 1–30, January 2012. 24

[116] Y Jiang, B Langley, F D Lubin, W Renthal, M A Wood, D H Yasui, A Kumar, E J Nestler, S Akbarian, and A C Beckel-Mitchener. **Epigenetics in the Nervous System**. *Journal of Neuroscience*, **28**(46):11753–11759, November 2008. 24

[117] Peter J Roy, Joshua M Stuart, Jim Lund, and Stuart K Kim. **Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans.** *Nature*, **418**(6901):975–979, August 2002. 24

[118] Myriam Heiman, Anne Schaefer, Shiaoching Gong, Jayms D Peterson, Michelle Day, Keri E Ramsey, Mayte Suárez-Fariñas, Cordelia Schwarz, Dietrich A Stephan, D James Surmeier, Paul Greengard, and Nathaniel Heintz. **A translational profiling approach for the molecular characterization of CNS cell types**. *Cell*, **135**(4):738–748, November 2008. 24, 25

[119] Elisenda Sanz, Linghai Yang, Thomas Su, David R Morris, G Stanley Mcknight, and Paul S Amieux. **Cell-type-specific isolation of ribosome-associated mRNA from complex tissues.** *Proceedings of the National Academy of Sciences*, **106**(33):13939–13944, August 2009. 24

[120] Z Yang. **Isolation of mRNA from specific tissues of Drosophila by mRNA tagging**. *Nucleic Acids Research*, **33**(17):e148–e148, September 2005. 24

[121] Florencia Pauli, Yueyi Liu, Yoona A Kim, Pei-Jiun Chen, and Stuart K Kim. **Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in C. elegans.** *Development (Cambridge, England)*, **133**(2):287–295, January 2006. 24

[122] Hirofumi Kunitomo, Hiroko Uesugi, Yuji Kohara, and Yuichi Iino. **Identification of ciliated sensory neuron-expressed genes in Caenorhabditis elegans using targeted pull-down of poly(A) tails.** *Genome Biology*, **6**(2):R17, 2005. 24

[123] Stephen E Von Stetina, Joseph D Watson, Rebecca M Fox, Kellen L Olszewski, W Clay Spencer, Peter J Roy, and David M Miller. **Cell-specific microarray profiling experiments reveal a comprehensive picture of gene expression in the C. elegans nervous system**. *Genome Biology*, **8**(7):R135, January 2007. 24

[124] W Clay Spencer, Georg Zeller, Joseph D Watson, Stefan R Henz, Kathie L Watkins, Rebecca D McWhirter, Sarah Petersen, Vipin T Sreedharan, Christian Widmer, Jeanyoung Jo, Valerie Reinke, Lisa Petrella, Susan Strome, Stephen E Von Stetina, Menachem Katz, Shai Shaham, Gunnar Rätsch, and David M Miller. **A spatial and temporal map of C. elegans gene expression.** *Genome Research*, **21**(2):325–341, February 2011. 24

[125] Joseph P Doyle, Joseph D Dougherty, Myriam Heiman, Eric F Schmidt, Tanya R Stevens, Guojun Ma, Sujata Bupp, Prerana Shrestha, Rajiv D Shah, Martin L Doughty, Shiaoching Gong, Paul Greengard, and Nathaniel Heintz. **Application of a translational profiling approach for the comparative analysis of CNS cell types**. *Cell*, **135**(4):749–762, November 2008. 25

REFERENCES

[126] J D Dougherty, E F Schmidt, M Nakajima, and N Heintz. **Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells**. *Nucleic Acids Research*, **38**(13):4218–4230, July 2010. 25

[127] M Hupe, M X Li, K Gertow Gillner, R H Adams, and J M Stenman. **Evaluation of TRAP-sequencing technology with a versatile conditional mouse model**. *Nucleic Acids Research*, October 2013. 25

[128] Michael R Miller, Kristin J Robinson, Michael D Cleary, and Chris Q Doe. **TU-tagging: cell type-specific RNA isolation from intact complex tissues.** *Nature Methods*, **6**(6):439–441, June 2009. 25

[129] Leslie Gay, Kate V Karfilis, Michael R Miller, Chris Q Doe, and Kryn Stankunas. **Applying thiouracil tagging to mouse transcriptome analysis.** *Nature Protocols*, **9**(2):410–420, February 2014. 25

[130] Michael D Cleary, Christopher D Meiering, Eric Jan, Rebecca Guymon, and John C Boothroyd. **Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay.** *Nature Biotechnology*, **23**(2):232–237, February 2005. 25

[131] Tony D Southall, Katrina S Gold, Boris Egger, Catherine M Davidson, Elizabeth E Caygill, Owen J Marshall, and Andrea H Brand. **Cell-Type-Specific Profiling of Gene Expression and Chromatin Binding without Cell Isolation: Assaying RNA Pol II Occupancy in Neural Stem Cells**. *Developmental cell*, **26**(1):101–112, July 2013. 25

[132] B van Steensel and S Henikoff. **Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase.** *Nature Biotechnology*, **18**(4):424–428, April 2000. 25

[133] Amanda Thomas, Pei-Jung Lee, Justin E Dalton, Krystle J Nomie, Loredana Stoica, Mauro Costa-Mattioli, Peter Chang, Sergey Nuzhdin, Michelle N Arbeitman, and Herman A Dierick. **A versatile method for cell-specific profiling of translated mRNAs in Drosophila.** *PloS one*, **7**(7):e40276, 2012. 27

[134] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nature Biotechnology*, **28**(5):511–515, May 2010. 27

[135] Thomas J Hardcastle and Krystyna A Kelly. **baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data**. *BMC Bioinformatics*, **11**(1):422, 2010. 27

[136] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics*, **26**(1):139–140, January 2010. 27

[137] Simon Anders and Wolfgang Huber. **Differential expression analysis for sequence count data.** *Genome Biology*, **11**(10):R106, 2010. 27

[138] Li Shen, Ning-Yi Shao, Xiaochuan Liu, Ian Maze, Jian Feng, and Eric J Nestler. **diffReps: Detecting Differential Chromatin Modification Sites from ChIP-seq Data with Biological Replicates**. *PloS one*, **8**(6):e65598, June 2013. 27

[139] Peter Glaus, Antti Honkela, and Magnus Rattray. **Identifying differentially expressed transcripts from RNA-seq data with biological variation.** *Bioinformatics*, **28**(13):1721–1728, July 2012. 27

[140] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. **Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.** *Genome Biology*, **14**(9):R95, 2013. 27

[141] Aaron T L Lun and Gordon K Smyth. **De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly.** *Nucleic Acids Research*, **42**(11):e95, 2014. 27

[142] Venkateswara R Chintapalli, Jing Wang, and Julian A T Dow. **Using FlyAtlas to identify better Drosophila melanogaster models of human disease**. *Nature genetics*, **39**(6):715–720, June 2007. 28

[143] S Robinow and K White. **Characterization and spatial distribution of the ELAV protein during Drosophila melanogaster development.** *Journal of neurobiology*, **22**(5):443–461, July 1991. 35

[144] Y Yuasa. **Drosophila homeodomain protein REPO controls glial differentiation by cooperating with ETS and BTB transcription factors**. *Development (Cambridge, England)*, **130**(11):2419–2428, June 2003. 35

[145] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biology*, **10**(3):R25, 2009. 40, 149

[146] Kaifu Chen, Yuanxin Xi, Xuewen Pan, Zhaoyu Li, Klaus Kaestner, Jessica Tyler, Sharon Dent, Xiangwei He, and Wei Li. **DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing.** *Genome Research*, **23**(2):341–351, February 2013. 40, 43

[147] A Klaes, T Menne, A Stollewerk, H Scholz, and C Klämbt. **The Ets transcription factors encoded by the Drosophila gene pointed direct glial cell differentiation in the embryonic CNS**. *Cell*, **78**(1):149–160, July 1994. 47

[148] Venkateswara R Chintapalli, Jing Wang, and Julian A T Dow. **Using FlyAtlas to identify better Drosophila melanogaster models of human disease**. *Nat Genet*, **39**(6):715–720, Jun 2007. 51

[149] I Komuro, M Schalling, L Jahn, R Bodmer, N A Jenkins, N G Copeland, and S Izumo. **Gtx: a novel murine homeobox-containing gene, expressed specifically in glial cells of the brain and germ cells of testis, has a transcriptional repressor activity in vitro for a serum-inducible promoter.** *The EMBO journal*, **12**(4):1387–1401, April 1993. 53

[150] D Fambrough and C S Goodman. **The Drosophila beaten path gene encodes a novel secreted protein that regulates defasciculation at motor axon choice points.** *Cell*, **87**(6):1049–1058, December 1996. 62

[151] P G Giresi, J Kim, R M Mcdaniell, V R Iyer, and J D Lieb. **FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin**. *Genome Research*, **17**(6):877–885, June 2007. 63

[152] N Ramsay, G Felsenfeld, B M Rushton, and J D McGhee. **A 145-base pair DNA sequence that positions itself precisely and asymmetrically on the nucleosome core.** *The EMBO journal*, **3**(11):2605–2611, November 1984. 63

[153] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. **Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position**. *Nature Methods*, **10**(12):1213–1218, October 2013. 63, 134

[154] L Stirling Churchman and Jonathan S Weissman. **Nascent transcript sequencing visualizes transcription at nucleotide resolution**. *Nature*, **469**(7330):368–373, April 2104. 65, 68

[155] Justin E Dalton, Tanvi S Kacheria, Simon RV Knott, Matthew S Lebo, Allison Nishitani, Laura E Sanders, Emma J Stirling, Ari Winbush, and Michelle N Arbeitman. **Dynamic, mating-induced gene expression changes in female head and brain tissues of Drosophila melanogaster.** *BMC Genomics*, **11**:541, 2010. 68

[156] Marijana Radonjic, Jean-Christophe Andrau, Philip Lijnzaad, Patrick Kemmeren, Thessa T J P Kockelkorn, Dik van Leenen, Nynke L van Berkum, and Frank C P Holstege. **Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon S. cerevisiae stationary phase exit**. *Molecular cell*, **18**(2):171–183, April 2005. 68

[157] B D Strahl and C D Allis. **The language of covalent histone modifications.** *Nature*, **403**(6765):41–45, January 2000. 68

[158] Li Shen, Ningyi Shao, Xiaochuan Liu, and Eric Nestler. **ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases.** *BMC Genomics*, **15**:284, 2014. 69, 114, 153

[159] Bing Li, Madelaine Gogol, Mike Carey, Samantha G Pattenden, Chris Seidel, and Jerry L Workman. **Infrequently transcribed long genes depend on the Set2/Rpd3S pathway for accurate transcription.** *Genes & Development*, **21**(11):1422–1430, June 2007. 86

[160] Yad Ghavi-Helm, Felix A Klein, Tibor Pakozdi, Lucia Ciglar, Daan Noordermeer, Wolfgang Huber, and Eileen E M Furlong. **Enhancer loops appear stable during development and are associated with paused polymerase.** *Nature*, July 2014. 95

[161] Iris Jonkers and John T Lis. **Getting up to speed with transcription elongation by RNA polymerase II.** *Nature reviews Molecular cell biology*, **16**(3):167–177, March 2015. 97

[162] Filip Crona, Olle Dahlberg, Lina E Lundberg, Jan Larsson, and Mattias Mannervik. **Gene regulation by the lysine demethylase KDM4A in Drosophila.** *Developmental biology*, **373**(2):453–463, January 2013. 99

[163] Toshiyuki Shiraki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, Akira Watahiki, Mari Nakamura, Takahiro Arakawa, Shiro Fukuda, Daisuke Sasaki, Anna Podhajska, Matthias Harbers, Jun Kawai, Piero Carninci, and Yoshihide Hayashizaki. **Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.** *Proceedings of the National Academy of Sciences of the United States of America*, **100**(26):15776–15781, December 2003. 100

[164] Nicola A Kearns, Hannah Pham, Barbara Tabak, Ryan M Genga, Noah J Silverstein, Manuel Garber, and René Maehr. **Functional annotation of native enhancers with a Cas9-histone demethylase fusion.** *Nature Methods*, **12**(5):401–403, May 2015. 100

[165] Daniel A Gilchrist, Gilberto Dos Santos, David C Fargo, Bin Xie, Yuan Gao, Leping Li, and Karen Adelman. **Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation**. *Cell*, **143**(4):540–551, Nov 2010. 102

[166] Roger A Hoskins, Jane M Landolin, James B Brown, Jeremy E Sandler, Hazuki Takahashi, Timo Lassmann, Charles Yu, Benjamin W Booth, Dayu Zhang, Kenneth H Wan, Li Yang, Nathan Boley, Justen Andrews, Thomas C Kaufman, Brenton R Graveley, Peter J Bickel, Piero Carninci, Joseph W Carlson, and Susan E Celniker. **Genome-wide analysis of promoter architecture in Drosophila melanogaster.** *Genome Research*, **21**(2):182–192, February 2011. 114

[167] Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner. **WebLogo: a sequence logo generator.** *Genome Research*, **14**(6):1188–1190, June 2004. 120, 154

[168] Christopher M Weber, Jorja G Henikoff, and Steven Henikoff. **H2A.Z nucleosomes enriched over active genes are homotypic.** *Nature Structural & Molecular Biology*, **17**(12):1500–1507, December 2010. 127

[169] Célia Jeronimo, Shinya Watanabe, Craig D Kaplan, Craig L Peterson, and François Robert. **The Histone Chaperones FACT and Spt6 Restrict H2A.Z from Intragenic Locations**. *Molecular cell*, pages 1–12, May 2015. 130

[170] T J Sliter and L I Gilbert. **Developmental arrest and ecdysteroid deficiency resulting from mutations at the dre4 locus of Drosophila.** *Genetics*, **130**(3):555–568, March 1992. 130, 131

[171] Barret D Pfeiffer, Teri-T B Ngo, Karen L Hibbard, Christine Murphy, Arnim Jenett, James W Truman, and Gerald M Rubin. **Refinement of tools for targeted gene expression in Drosophila.** *Genetics*, **186**(2):735–755, October 2010. 134

[172] Leighton J Core, André L Martins, Charles G Danko, Colin T Waters, Adam Siepel, and John T Lis. **Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers**. *Nature Publishing Group*, **46**(12):1311–1320, November 2014. 134

[173] Ho Sung Rhee and B Franklin Pugh. **Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution**. *Cell*, **147**(6):1408–19, Dec 2011. 134

[174] Nicholas T Ingolia, Sina Ghaemmaghami, John R S Newman, and Jonathan S Weissman. **Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling.** *Science (New York, NY)*, **324**(5924):218–223, April 2009. 134

[175] Irfan A Qureshi, Solen Gokhan, and Mark F Mehler. **REST and CoREST are transcriptional and epigenetic regulators of seminal neural fate decisions.** *Cell cycle (Georgetown, Tex.)*, **9**(22):4477–4486, November 2010. 135

[176] G M Rubin and A C Spradling. **Genetic transformation of Drosophila with transposable element vectors.** *Science (New York, NY)*, **218**(4570):348–353, October 1982. 140

[177] G L Henry, F P Davis, S Picard, and S R Eddy. **Cell type-specific genomics of Drosophila neurons**. *Nucleic Acids Research*, pages 1–14, August 2012. 141

[178] modENCODE Consortium, Sushmita Roy, Jason Ernst, Peter V Kharchenko, Pouya Kheradpour, Nicolas Negre, Matthew L Eaton, Jane M Landolin, Christopher A Bristow, Lijia Ma, Michael F Lin, Stefan Washietl, Bradley I Arshinoff, Ferhat Ay, Patrick E Meyer, Nicolas Robine, Nicole L Washington, Luisa Di Stefano, Eugene Berezikov, Christopher D Brown, Rogerio Candeias, Joseph W Carlson, Adrian Carr, Irwin Jungreis, Daniel Marbach, Rachel Sealfon, Michael Y Tolstorukov, Sebastian Will, Artyom A Alekseyenko, Carlo Artieri, Benjamin W Booth, Angela N Brooks, Qi Dai,

REFERENCES

Carrie A Davis, Michael O Duff, Xin Feng, Andrey A Gorchakov, Tingting Gu, Jorja G Henikoff, Philipp Kapranov, Renhua Li, Heather K MacAlpine, John Malone, Aki Minoda, Jared Nordman, Katsutomo Okamura, Marc Perry, Sara K Powell, Nicole C Riddle, Akiko Sakai, Anastasia Samsonova, Jeremy E Sandler, Yuri B Schwartz, Noa Sher, Rebecca Spokony, David Sturgill, Marijke van Baren, Kenneth H Wan, Li Yang, Charles Yu, Elise Feingold, Peter Good, Mark Guyer, Rebecca Lowdon, Kami Ahmad, Justen Andrews, Bonnie Berger, Steven E Brenner, Michael R Brent, Lucy Cherbas, Sarah C R Elgin, Thomas R Gingeras, Robert Grossman, Roger A Hoskins, Thomas C Kaufman, William Kent, Mitzi I Kuroda, Terry Orr-Weaver, Norbert Perrimon, Vincenzo Pirrotta, James W Posakony, Bing Ren, Steven Russell, Peter Cherbas, Brenton R Graveley, Suzanna Lewis, Gos Micklem, Brian Oliver, Peter J Park, Susan E Celniker, Steven Henikoff, Gary H Karpen, Eric C Lai, David M Macalpine, Lincoln D Stein, Kevin P White, and Manolis Kellis. **Identification of functional elements and regulatory circuits by Drosophila modENCODE**. *Science*, **330**(6012):1787–97, Dec 2010. 141

[179] Marcin Kruczyk, Husen M Umer, Stefan Enroth, and Jan Komorowski. **Peak Finder Metaserver - a novel application for finding peaks in ChIP-seq data.** *BMC Bioinformatics*, **14**(1):280, September 2013. 149

Appendix A

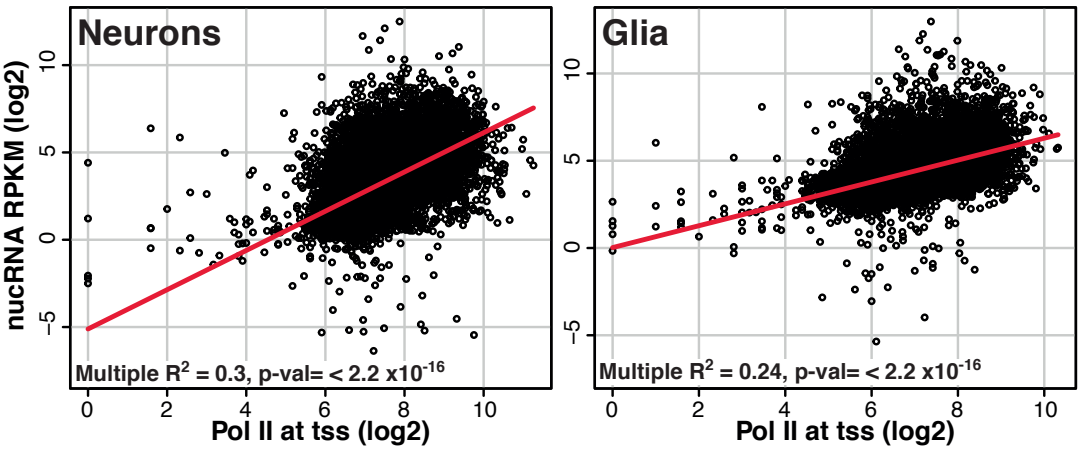


Figure A.1: Pol II binding correlates with RNA levels - Scatterplots comparing the Pol II binding level over the TSS (log2) with the nucRNA RPKM (log2). All genes were assessed, if there was zero coverage at any gene, either in the Pol II binding or the nucRNA, then the gene was removed from analysis. Neurons means the Pol II binding level was calculated from the neurons CAST-ChIP RPB3 data and the nucRNA RPKM was of the neuron-specific nucRNA data. Glia compares the glia RPB3 CAST-ChIP versus the glia-specific nucRNA data. The regression was computed using the *lm* function in R.

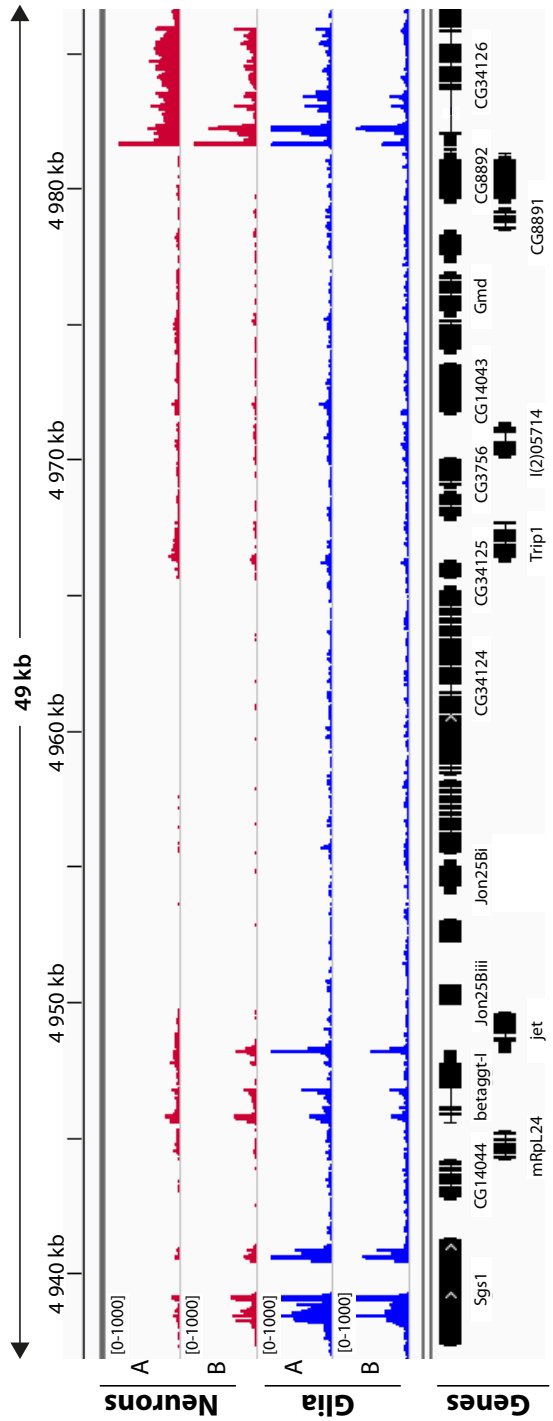


Figure A.2: Genome-browser snapshots of nucRNA data

The nucRNA data mapped to the genome without any annotation guidance for each biological replicate. The low level of genomic DNA contamination can be seen in the glia-specific samples in that there are reads over genomic regions without any annotation.

Appendix B

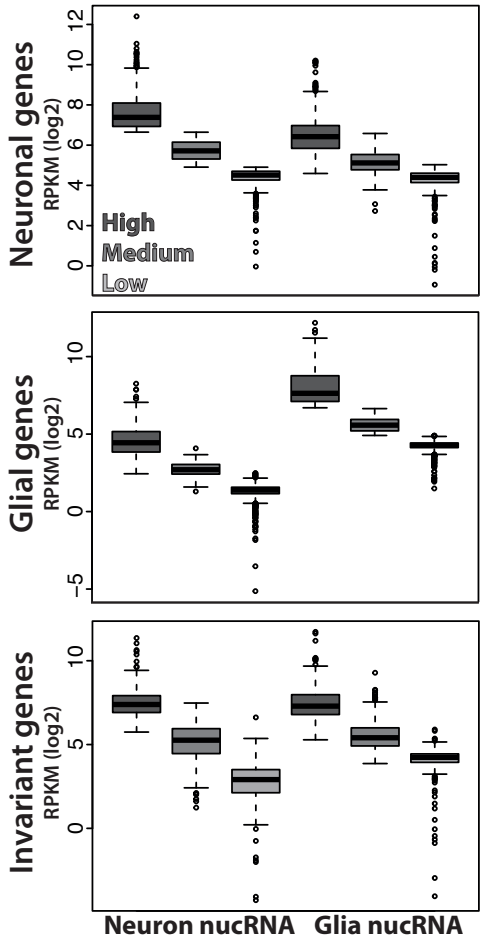


Figure B.1: Splitting genes into high, medium and low expression - Boxplots of RPKM for the neuronal nucRNA and glial nucRNA datasets. Genes were split according to the DESeq gene calls (neuronal genes, glial genes, and invariant genes).Then further split into high, medium, and low expression based on the RPKM of the cell type where the gene was called specific, and the inputs for invariant genes.

B.

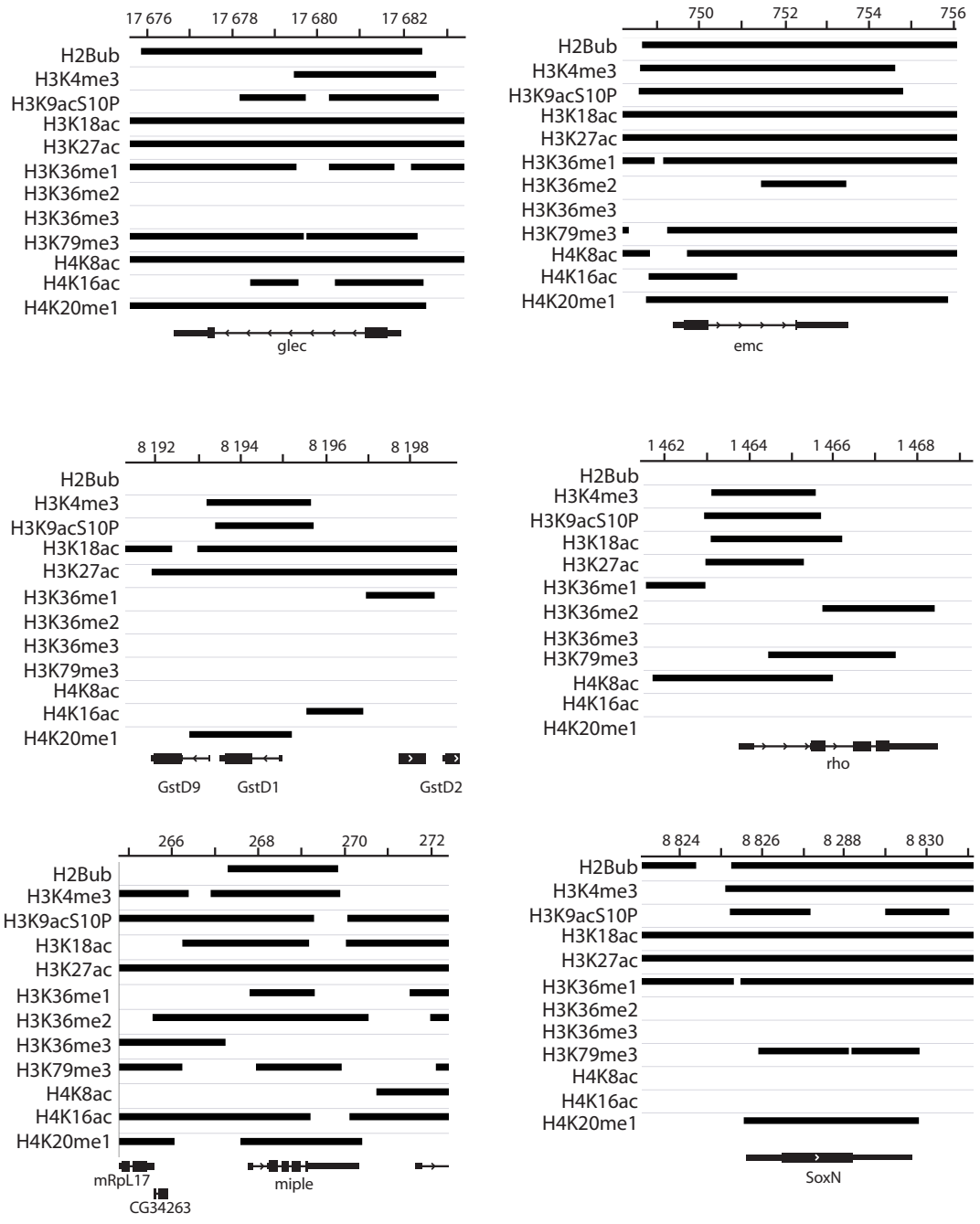


Figure B.2: Active-H3K36me3 genes with broad H3K27ac also enriched for other active modifications - Genome-browser snapshots showing the same gene set examples shown in 4.9, using available data from the modencode consortium. The data are for Adult head, mixed population and are shown as the regions where peaks for each histone modification was called. These data are in agreement with the data obtained in this work, in that these genes have no H3K36me3 and have broad H3K37ac peaks. Enrichment of H3K18ac and H3K4me3 is also seen at these genes, indicating that multiple active marks may be enriched, and perhaps even broadly spread, across these genes. Data from the modencode consortium, http://www.modencode.org

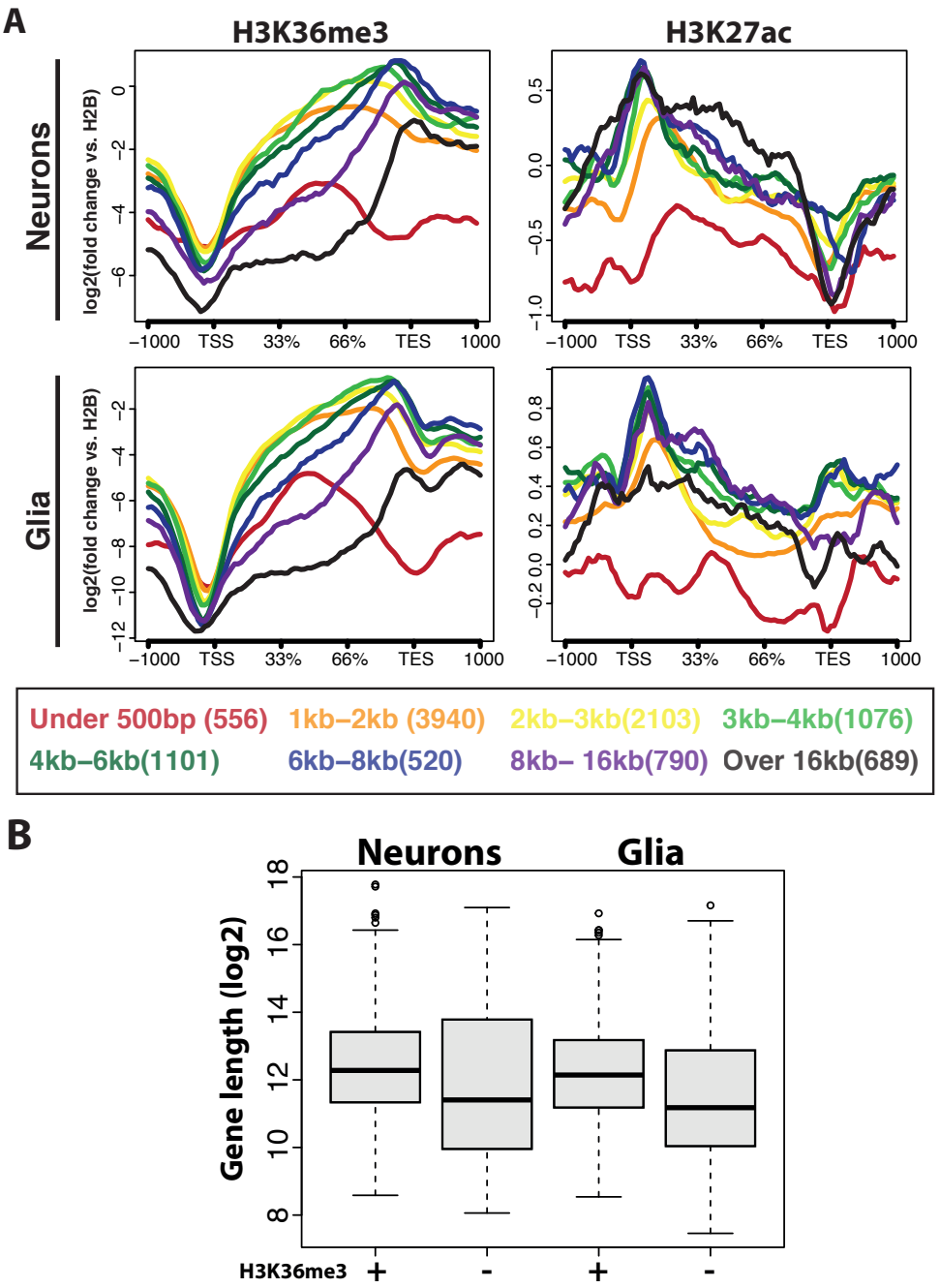


Figure B.3: Histone marks and gene length - A) *Drosophila* genes were split into eight categories based on the entire genomic region of the gene from the TSS to the TES. The shortest gene groups are 500 bp or shorter. These genes are theoretically too short for incorporation of H3K36me3, as the Pol II finishes transcribing the genes before the Set2 is recruited. Further groups were made with increased by 1kb intervals: 1kb-2kb, 2kb-3kb, 3kb-4kb, three broader groups: 4kb-6kb, 6kb-8kb and 8kb-16kb, and a final group of genes that are 16kb and longer. The average profiles of H3K36me3 and H3K27ac data are shown for each group in each cell type. B) Average gene lengths of genes with H3K36me3 and without H3K36me3

Appendix C

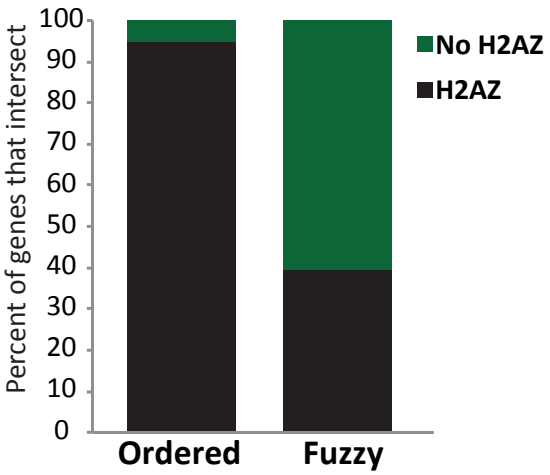
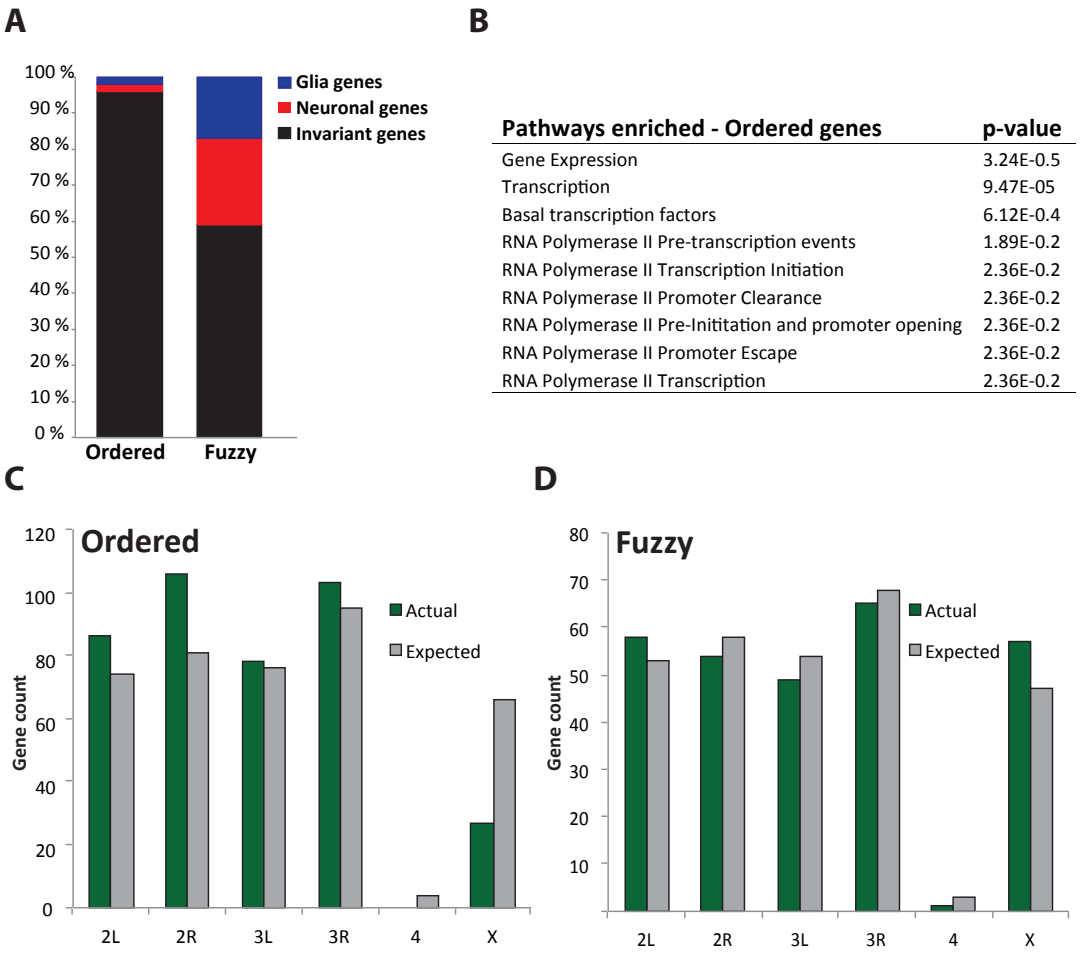


Figure C.1: Fuzzy genes have little H2AZ - The proportion of ordered genes with H2AZ is much higher than ordered genes with no H2AZ. The Fuzzy genes have a higher proportion of genes with no H2AZ compared to fuzzy genes with H2AZ



Pathways enriched - Ordered genes	p-value
Gene Expression	3.24E-0.5
Transcription	9.47E-05
Basal transcription factors	6.12E-0.4
RNA Polymerase II Pre-transcription events	1.89E-0.2
RNA Polymerase II Transcription Initiation	2.36E-0.2
RNA Polymerase II Promoter Clearance	2.36E-0.2
RNA Polymerase II Pre-Inititation and promoter opening	2.36E-0.2
RNA Polymerase II Promoter Escape	2.36E-0.2
RNA Polymerase II Transcription	2.36E-0.2

Figure C.2: Ordered and Fuzzy gene groups have different features - A) Cell-type-specificity of different gene groups. The ordered genes were highly enriched in the invariant gene class and the fuzzy genes contained more cell-type-specific genes. B) GO-enrichment. Ordered genes were enriched for pathways related to gene expression and transcriptional events, whereas fuzzy genes had no pathway enrichment. C) Distribution of genes across the genome. Genes in the ordered gene class are depleted on the X-chromosome, but the fuzzy gene class are enriched on the X-chromosome. The fact that X-chromosome genes are specifically to achieve dosage compensation is consistent with the enrichment of cell-type-specific genes in the fuzzy gene class, these genes need to be regulated in a specialised way and thus have a different chromatin architecture than the invariant or housekeeping genes.

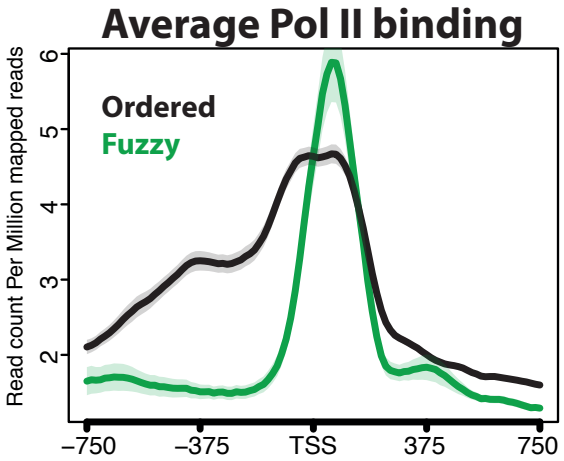


Figure C.3: Broad Pol II peak at ordered promoters, sharp Pol II at fuzzy promoters - Metagene analysis of the Pol II binding across the TSS of fuzzy or ordered genes. This shows the broad binding peak of the ordered genes, indeed there appear to be two peaks. The fuzzy genes have a very sharp binding peak that is highest around 50 bp into the gene body.

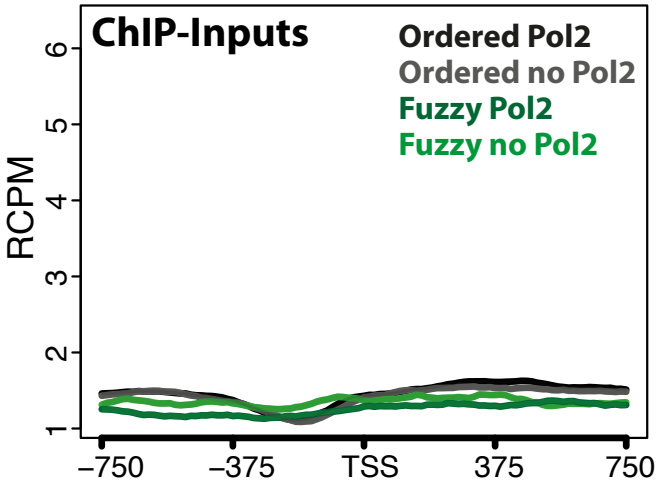


Figure C.4: ChIP inputs show no background binding - ChIP-seq inputs from the head RPB3-ChIP were plotted, they show no enrichment reads due to the chromatin background.

Appendix D

Vectors and sequences

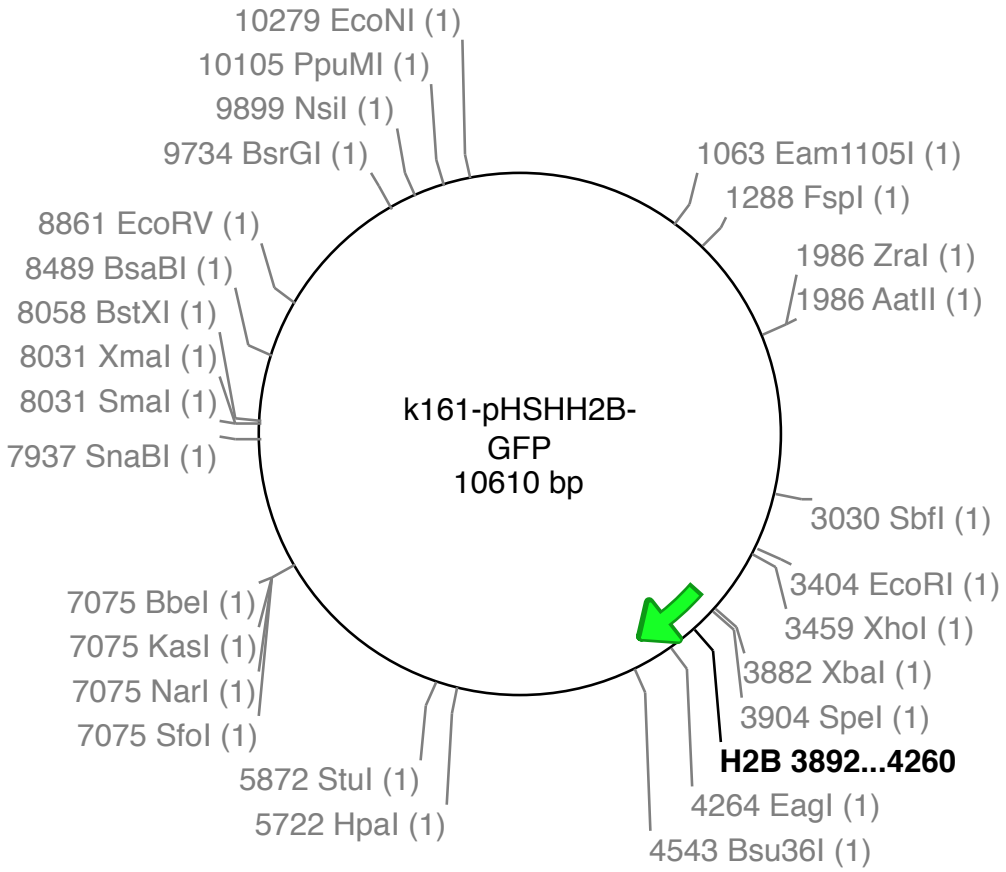


Figure D.1: k161_PHS_H2B-GFP - Plasmid from which H2B-GFP transgene was PCR-amplified. Vector graphics made with Ape (A plamid Editor v2.0.26)

D. VECTORS AND SEQUENCES

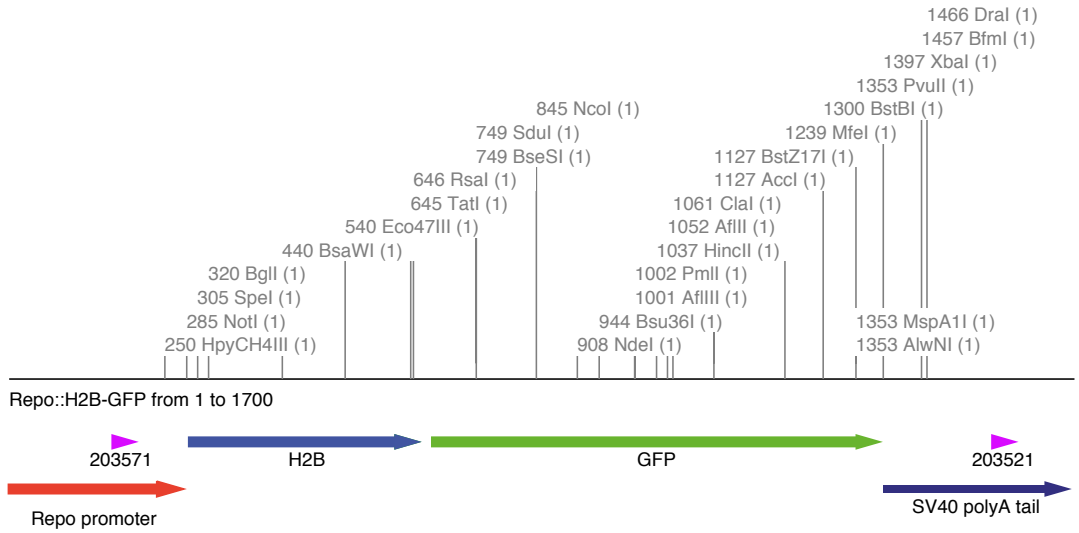


Figure D.2: Cloning construct Repo::H2B-GFP - The H2B-GFP gene construct was cloned into the pCasPer. Graphics made with Ape (A plamid Editor v2.0.26)

Repo::H2B-GFP translated sequence:

MPPKTSGKAAKKAGKAQKNITKTDKKKKRKRKESYAIYIY

KVLKQVHPDTGISSKAMSIMNSFVNDIFERIAAEASRLAH

YNKRSTITSREIQTAVRLLLPGELAKHAVSEGTKAVTKYT

SSKHRPVATMSKGEELFTGVVPILVELDGDVNGHKFSVSG

EGEGDATYGKLTLKFICTTGKLPVPWPTLVTTFTYGVQCF

SRYPDHMKRHDFFKSAMPEGYVQERTIFFKDDGNYKTRAE

VKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVY

IMADKQKNGIKANFKTRHNIEDGGVQLADHYQQNTPIGDG

PVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITH

GMDELYK

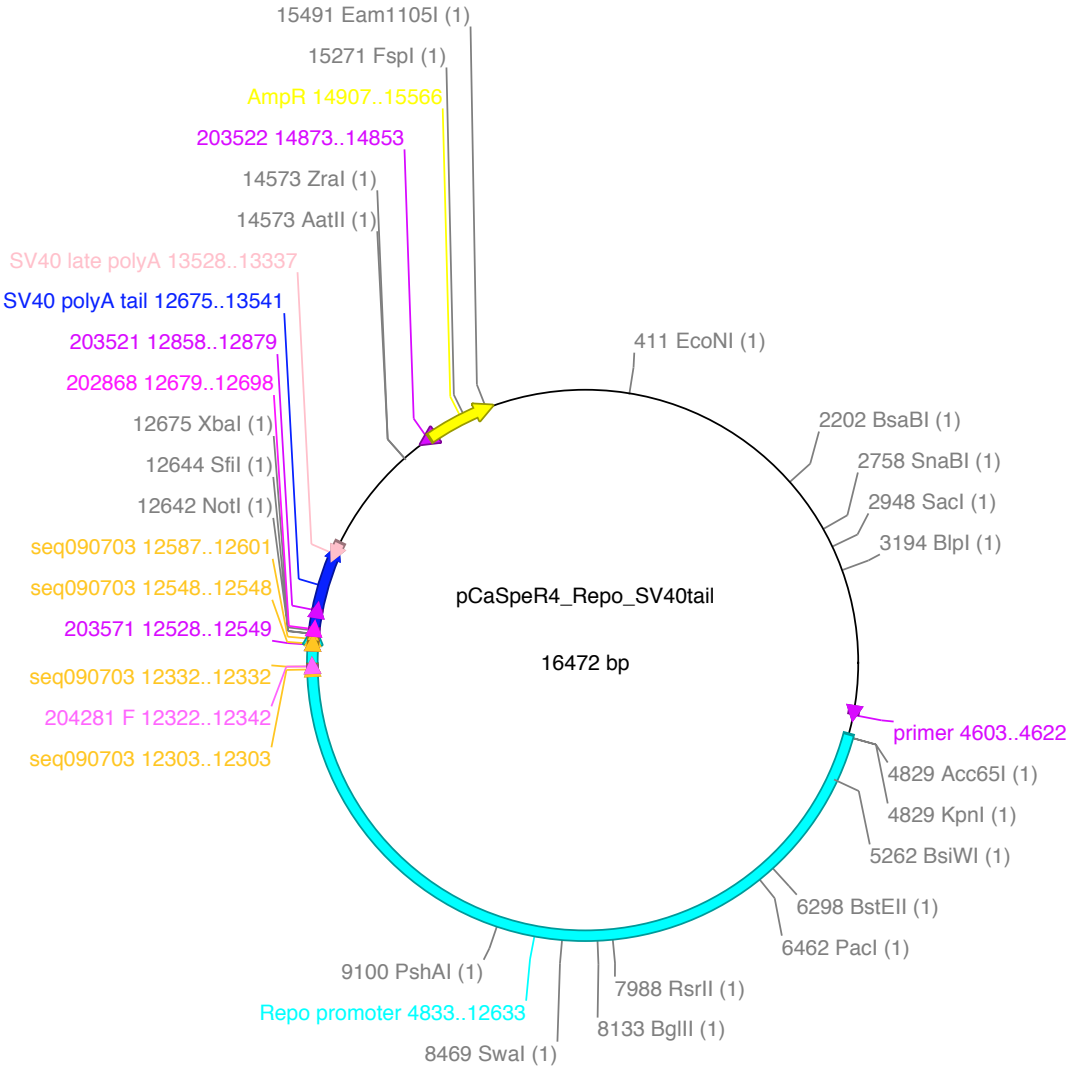


Figure D.3: pCaSpeR4 vector with Repo promoter region - This vector was used to clone the H2B-GFP transgene and iset into the *Drosophila* genome using p-elemtn insertion. The vector contains a large regulatory region that controls glia-specific expression of the *repo* gene. Vector graphics made with Ape (A plamid Editor v2.0.26).

Table D.1: Primers used in this research

Target gene	Primer ID	sequence	Use
H2B-GFP	Not1_H2B_F	AAATAT GCGGCCGC ATGCCTCCGAAAACTAGT	Repo::H2B-GFP fly strain.
	Xba1_GFP_R	GCTCTAGA TTATTTGTATAGTTCATCC	
Zen	Zen_k27me3_F	CTGGAGCACATGCCGTTGTT	Positive for H3K27me3
	Zen_k27me3_R	AAGCGCCTCATGTCCTACTCC	
RPS3 3' end	RPS3_3'_F	GCCCGAGACCGAGTACAAGA	Positive for H3K36me3
	RPS3_3'_R	TGAGGGCAACTCTTTCAGCT	
RPL32 Promoter	RPL32P_F	TTCACGATCTTGGGCCTGTATG	Positive for H3K27ac
	RPL32P_R	TTGTTGTGTCCTTCCAGCTTCA	
CG7943	FZ_CG7943_F	ACGAGTGTGTTGTTTGTGTTCGC	Positive for H3K4me1
	FZ_CG7943_R	AGATGTACGCTTAAGCCTCGTTCG	
Fasculin promoter	fas +1 F	GTTACGTTTCGACGGCCAAC	Positive for RPB3
	fas +1 R	GCGAGTATGAAGAAGAAGAACCA	
ocnus	ocn_F	TGCCCTGAAGAAGGCTCCGAT	Negative for H3K4me1
	ocn_R	TCTGCTTTGTTGGTTAGCAGCC	
Ultra bithorax	Ubx_F4 pre F	TAGTCTTATCTGTATCTCGCTCTTA	Negative for H3K36me3,
	Ubx_F4 pre R	CAGAACCAAAGTGCCGATAACTC	H3K27ac, and H3K27me3
Elav	elav RT1 F	CAACCGAAGTAACCATAACTGGA	Positive for elav mRNA
	elav RT1 R	TCCTTGCTCTCTGCTTCGAT	
Repo	repo RT1 F	ATCCCAATGGCATCAAGAAG	Positive for repo mRNA
	repo RT1 R	ACACGGGATTCGCTCAGAT	
RP49	Rp49 F	GGTATCGACAACAGAGTGCG	Normalisation of RT-qPCR
	Rp49 R	GAACTTCTTGAATCCGGTGGG	