FUNCTIONS OF HISTONE H2A.Z IN REGULATING TRANSCRIPT LEVELS IN BUDDING YEAST

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Biology, Medicine and Health

2016

By
Muxin Gu
School of Biological Sciences
Contents

Abstract ................................................................. 13

Declaration ............................................................. 15

Copyright ............................................................... 16

1 Introduction ......................................................... 19
  1.1 Histone Modifications ........................................... 20
  1.2 Histone Variant H2A.Z ........................................ 22
  1.3 Chromatin Organisation in Gene Vicinity ...................... 22
    1.3.1 Patterns of histone marks at promoters .................. 24
    1.3.2 Remodelling of promoter during gene activation .......... 26
    1.3.3 Chromatin organisation inside coding regions .......... 27
    1.3.4 Resetting chromatin after elongation ................... 28
  1.4 Non-coding RNA Transcripts ................................... 29
    1.4.1 Intergenic lncRNA ........................................ 30
    1.4.2 Promoter-associated lncRNA in higher eukaryotes .......... 31
    1.4.3 Promoter-associated lncRNA in *S.cerevisiae* ............ 31
    1.4.4 Promoter bidirectionality ................................ 34
  1.5 Project Aims .................................................... 35

2 Methods .................................................................. 37
  2.1 Strains ............................................................ 37
  2.2 Cell culture and DNA library ................................... 37
  2.3 ChIP-seq Data Analysis .......................................... 38
    2.3.1 ChIP-seq data processing .................................... 38
    2.3.2 Peak identification ........................................... 39
  2.4 Strand-specific RNA-seq Analysis .............................. 40
2.4.1 RNA-seq data processing ........................................ 40
2.4.2 Differential expression ........................................... 41
2.4.3 Transcriptome assembly ........................................... 42
2.4.4 Hierarchical clustering ............................................ 42
2.5 Acquisition of Published Data ...................................... 42
2.5.1 ChIP data .......................................................... 42
2.5.2 Data processing .................................................... 42
2.5.3 Gene and transcript annotations ................................. 43

3 H2A.Z Positively Regulates Antisense Transcripts ................. 45
3.1 Introduction .......................................................... 45
3.1.1 H2A.Z stops spread of telomeric heterochromatin .......... 45
3.1.2 H2A.Z is required for gene activation ......................... 46
3.1.3 H2A.Z occupancy negatively correlates with transcription rates in \textit{S. cerevisiae} .................................................. 46
3.1.4 Data resolution is important for quantification ............... 47
3.1.5 Accurate gene annotation is essential for studying H2A.Z-transcript relationship ........................................... 47
3.1.6 Dissecting direct and indirect effects of H2A.Z on transcription .................................................. 48
3.1.7 Objectives .......................................................... 48
3.2 H2A.Z is enriched at 3' ends of genes ............................. 50
3.3 H2A.Z outside TSS marks active antisense promoters ........... 52
3.3.1 Cryptic transcripts are revealed in \textit{rrp6Δ} strain ............... 52
3.3.2 H2A.Z at 3' ends of genes marks start sites of AS transcripts 54
3.3.3 H2A.Z-marked 3'-end regions are enriched for active chromatin marks .................................................. 56
3.3.4 Abundance of 3' H2A.Z correlates with AS transcript levels 58
3.3.5 H2A.Z has stronger association with cryptic than stable transcripts .................................................. 59
3.4 H2A.Z at 3' promoters affects expression by regulating AS transcripts 61
3.4.1 Effects of H2A.Z on transcriptome ............................. 62
3.4.2 H2A.Z is required for expression of AS transcripts .......... 64
3.4.3 AS has negative effects on sense transcripts .................. 66
3.5 H2A.Z-related AS are coregulated with adjacent sense transcripts 70
3.5.1 H2A.Z-related AS transcripts predominantly occur at tandem genes ........................................... 70
3.5.2 AS and adjacent sense transcripts are coregulated by H2A.Z .................................................. 71
3.6 H2A.Z peaks inside coding regions negatively correlate with AS transcripts .............................................. 75
  3.6.1 H2A.Z peaks in CDS are associated with AS transcripts .......................................................... 75
  3.6.2 CDS H2A.Z has positive effects on AS and negative effects on sense transcripts .......................... 78
3.7 Discussion ........................................................................................................................................... 79
  3.7.1 ChIP-seq improves quantification of H2A.Z .................................................................................. 79
  3.7.2 H2A.Z at promoter is required for both sense and AS transcripts ................................................. 79
  3.7.3 Overall effect of H2A.Z on sense transcripts ................................................................................. 80
  3.7.4 H2A.Z is likely marking bidirectional promoters ......................................................................... 81
  3.7.5 H2A.Z’s correlation with transcript levels .................................................................................... 81
  3.7.6 Subtelomeric H2A.Z does not mark AS transcripts .................................................................... 82
  3.7.7 Repressive effect of H2A.Z on AS transcripts in *S. pombe* ......................................................... 82
  3.7.8 Steady state versus stress response ............................................................................................. 83
3.8 Summary ............................................................................................................................................. 84
3.9 Supplementary Figure ...................................................................................................................... 85

4 Non-coding RNAs Between Tandem Genes Are Induced During Stress Response 87
4.1 Introduction ........................................................................................................................................ 87
  4.1.1 Transcriptomic changes in stress response .................................................................................. 87
  4.1.2 Upstream *cis*-lncRNAs as repressor ...................................................................................... 88
  4.1.3 lncRNA in *trans* as repressor .................................................................................................... 90
  4.1.4 lncRNA as gene activator ........................................................................................................ 90
  4.1.5 Objective .................................................................................................................................... 91
4.2 Caffeine Causes Significant Transcriptomic Changes ......................................................................... 92
  4.2.1 Induced and repressed genes .................................................................................................... 92
  4.2.2 Caffeine response is similar to general ESRs ............................................................................ 94
  4.2.3 Gene expression recovers after 30 minutes ............................................................................... 94
4.3 Long Non-coding RNAs Are Significantly Induced by Caffeine .................................................... 98
  4.3.1 lncRNA levels are increased in during caffeine response ......................................................... 98
  4.3.2 Induced lncRNAs are related to stress-response genes .............................................................. 100
4.3.3 Induced lncRNA is accompanied by gene activation . . . . 102
4.4 LncRNA Links Co-activated Tandemly Arranged Gene Pairs . . 104
  4.4.1 Classification of lncRNA by relative location to gene . . . 104
  4.4.2 Promoter-associated AS lncRNAs are coregulated with genes106
  4.4.3 AS lncRNAs occur between tandem gene pairs . . . . . . . 108
  4.4.4 Tandem gene pairs with intergenic lncRNA tend to be
                  coregulated . . . . . . . . . . . . . . . . . . . . . . . 109
4.5 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 111
  4.5.1 Organisation of lncRNAs in relation to genes . . . . . . . 111
  4.5.2 AS:sense coregulation is likely due to promoter sharing . . 111
  4.5.3 Regulation between convergent transcripts . . . . . . . . . 112
  4.5.4 Possible mechanisms of sense:AS coregulation . . . . . . . 113
  4.5.5 Comparison to other organisms . . . . . . . . . . . . . . . 114
  4.5.6 Outstanding questions . . . . . . . . . . . . . . . . . . . 114
4.6 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 115
4.7 Supplementary Figure . . . . . . . . . . . . . . . . . . . . . . . 117

5 Delayed caffeine-induced H2A.Z Drop in Mutants Causes Slower
Gene Recovery . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 119
  5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . 119
    5.1.1 H2A.Z becomes essential under stress . . . . . . . . . . . 119
    5.1.2 Cellular H2A.Z abundance decreases during caffeine response120
    5.1.3 Mutant strains to mimic effect of htz1Δ . . . . . . . . . . . 120
    5.1.4 Objective . . . . . . . . . . . . . . . . . . . . . . . . . 121
    5.1.5 Limitation of conventional normalisation methods . . . . 121
5.2 Decreased Chromatin-bound H2A.Z is Revealed by Normalised Data124
    5.2.1 Signal-to-background ratios decrease during caffeine response124
    5.2.2 Normalisation for WT and 4KR agrees with experimental
            data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 126
    5.2.3 OE samples are normalised to sequencing depth . . . . . . 126
5.3 H2A.Z Drop Occurs at Promoters . . . . . . . . . . . . . . . . 130
    5.3.1 Promoter H2A.Z decreases during caffeine response . . . 130
    5.3.2 H2A.Z in coding regions increases during caffeine response 134
5.4 Promoter H2A.Z Drop is Linked with Gene Activation . . . . . 136
    5.4.1 Changes in 5’ H2A.Z levels anti-correlate changes in gene
            expression . . . . . . . . . . . . . . . . . . . . . . . . . . 136
5.4.2 H2A.Z drop at 30 minutes is primarily contributed by gene activation

5.5 Delayed H2A.Z Drop Accompanies Reduced Gene Recovery

5.5.1 4KR and OE have negligible impact on steady-state transcriptome

5.5.2 Dynamic range of H2A.Z is reduced in 4KR and OE strains

5.5.3 4KR and OE share the same set of genes with reduced H2A.Z drop

5.5.4 Reduced H2A.Z drop is accompanied by slower H2A.Z recovery

5.5.5 Majority of reduced H2A.Z drop is independent of gene activation

5.5.6 Reduced H2A.Z drop is accompanied by slower gene expression recovery

5.5.7 Reduced 5’ H2A.Z drop affects highly responsive genes

5.6 Discussion

5.6.1 Limitations and improvements for normalisation

5.6.2 Majority of 5’ H2A.Z drop can be explained by gene activation

5.6.3 Evidence for nucleosome-independent H2A.Z eviction

5.6.4 Reduced 5’ H2A.Z drop is accompanied by delayed expression recovery

5.6.5 Similarity to ino80Δ mutant

5.7 Summary

5.8 Supplementary Figure

6 Discussion and Outlook

6.1 Promoter H2A.Z as a Positive Regulator of Transcription

6.2 Asymmetry of Bidirectional Promoters

6.3 Non-coding Transcript as a Regulatory Signal for Distal Targets

6.4 H2A.Z May Affect Chromatin Reset after Induction

6.5 Concluding Remarks

Appendix: Published Paper

Word Count: 51,402
List of Tables

1.1 Histone acetylation and methylation in *S.cerevisiae*. . . . . . . . . . . . 21

2.1 Yeast strains used in this study. . . . . . . . . . . . . . . . . . . . . . . . . . 37

2.2 Summary of ChIP-seq mapping. . . . . . . . . . . . . . . . . . . . . . . . . . 39

2.3 Summary of strand-specific RNA-seq mapping. . . . . . . . . . . . . . . . . . 41

2.4 ChIP-chip and ChIP-seq data imported from published studies. . . . . . . 42
## List of Figures

1.1 Histone modifications and variants at promoters of genes. ........ 25  
1.2 Chromatin remodelling at promoters during transcription initiation. 27  
1.3 Restoring chromatin structure after transcription. ................. 29  
1.4 Types of long non-coding RNAs in *S. cerevisiae*. ................. 33  
1.5 Pre-initiation complexes at bidirectional promoters. ............... 34  

3.1 Distribution of H2A.Z at genic and intergenic regions. ............ 51  
3.2 Cryptic transcripts are revealed in *rrp6Δ*. .......................... 53  
3.3 Co-localisation between 3’ H2A.Z peaks and AS transcripts. ....... 55  
3.4 H2A.Z-occupied 3’ ends are enriched for active histone marks. .... 57  
3.5 Comparison between 3’ H2A.Z and histone marks that are depleted from active promoters. ............................................ 58  
3.6 H2A.Z at 3’ ends of genes correlates with gene activity .......... 59  
3.7 Classification of cryptic and stable transcripts. ....................... 60  
3.8 Three-prime H2A.Z has stronger association with CUTs than SUTs. 61  
3.9 Effect of *HTZ1* deletion on transcriptome in Rrp6+ and Rrp6− backgrounds. .......................................................... 63  
3.10 Positive effect of 3’ H2A.Z on AS transcript levels. .................. 65  
3.11 Positive effect of 5’ H2A.Z on sense transcripts of genes .......... 66  
3.12 AS transcripts have negative effects on sense transcripts. ........ 67  
3.13 Combinatorial effects of 5’ and 3’ H2A.Z on sense transcripts. ... 68  
3.14 Antisense-associated H2A.Z are predominantly found in tandemly transcribed genes. ................................................. 71  
3.15 Examples of H2A.Z-related AS transcripts coregulated with adjacent sense transcript. ............................................... 72  
3.16 Coregulation between H2A.Z-related AS and adjacent sense transcripts. ............................................................... 73
3.17 H2A.Z peaks inside coding regions of genes are associated with AS transcripts. .................................................. 76
3.18 Effects of CDS H2A.Z on AS and sense transcripts. ............... 78
3.19 Summary of the effect of H2A.Z on transcripts. ....................... 83
S3.1 Correlation between 5’ H2A.Z and expression. ..................... 85
S3.2 CDS H2A.Z in subtelomeric regions. ................................. 85

4.1 Locations of lncRNAs relative to genes and mechanisms of regulatory functions. ........................................... 89
4.2 Activated and repressed genes at 30 minutes of caffeine response. 93
4.3 Caffeine response is similar to other environmental stress responses. 95
4.4 Recovery of expression from 30 to 120 minutes in caffeine response. 96
4.5 ncRNAs are increased during caffeine response. ................... 99
4.6 ncRNAs are related to stress-response genes. ........................ 101
4.7 Increased lncRNAs are accompanied by increased sense transcripts. 103
4.8 Classification of AS transcripts by relative location to genes. ... 105
4.9 AS transcripts upstream of genes are coregulated with the gene. . 107
4.10 AS transcripts downstream of TSS and inside CDSs are less coregulated with sense transcripts. ......................... 108
4.11 AS transcripts starting upstream of TSS link tandem gene pairs. . 109
4.12 Sense:AS:sense cluster shows coregulation during caffeine response. 110
4.13 Coregulated sense:AS:sense cluster. ................................. 115
S4.1 Convergent gene pairs are not coregulated in caffeine response. . 117

5.1 Decreasing cellular abundance of H2A.Z during caffeine response. 121
5.2 Decreased signal-to-background ratios of ChIP-seq during caffeine response. ............................................. 125
5.3 Normalised ChIP-seq data show decreased chromatin-bound H2A.Z during caffeine response. ............................ 127
5.4 Normalised data in OE strain show decreased chromatin levels of H2A.Z. .................................................... 128
5.5 Decrease of H2A.Z levels at the 5’ end of genes during caffeine response. .................................................. 131
5.6 Recovery of 5’ H2A.Z occupancy from 30 minutes to 120 minutes. 133
5.7 Change of H2A.Z occupancy inside coding regions of genes during caffeine response. ............................... 135
5.8 Negative correlation between changes of expression and 5’ H2A.Z occupancy from 0 to 30 minutes. ................. 137
5.9 Promoter H2A.Z drop is contributed by gene activation. ....... 139
5.10 4KR and OE mutations have minor effects on 0-minute gene expression. ................................................. 140
5.11 Reduced dynamic ranges of promoter H2A.Z in 4KR and OE strains. 141
5.12 Reduced drop of 5’ H2A.Z at 30 minutes in mutant strains compared to WT. ........................................... 143
5.13 Reduced H2A.Z drop reflects delayed H2A.Z drop during caffeine response. ........................................... 144
5.14 Majority of the reduced 5’ H2A.Z drop in 4KR and OE is independent of differences in gene activation. .......... 145
5.15 Reduced drop of 5’ H2A.Z is accompanied by delayed recovery of gene expression. ................................ 147
5.16 Highly induced genes tend to have reduced drop of 5’ H2A.Z. .. 149
S5.1 Gene expression of Swr1, Ino80 and Htz1 in caffeine time course. 154
S5.2 Changes of H2A.Z in CDS are uncorrelated with changes in sense or AS transcripts. ................................. 155
Abstract

**Functions of Histone H2A.Z in Regulating Transcript Levels in Budding Yeast**
Muxin Gu
A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy, 2016

The histone variant H2A.Z is an important regulator of transcription. One unsolved mystery is that why H2A.Z can have both activating and repressive effects on gene expression. By examining both coding and non-coding RNA transcripts in *S. cerevisiae*, we established that H2A.Z is present at the promoters of both coding and non-coding transcripts and have positive effects on transcript levels. The repressive effect of H2A.Z can be partially explained by the sense transcripts being antagonised by H2A.Z-activated antisense transcripts. We also established that H2A.Z-associated non-coding transcripts are predominantly located at bidirectional promoters. The sense and antisense pairs produced from bidirectional promoters show high degrees of coregulation (especially co-activation) during stress response. Surprisingly, we found that the non-coding RNAs that are co-activated with stress-response genes tend to spread the activation signal to the neighbouring gene further upstream, indicating their potential functions in gene regulation. In addition, we also observed that accumulation of H2A.Z at gene promoters is associated with slower recovery from gene induction, which could be related to the Ino80 pathway. In general, our results confirmed the interleaved nature of regulatory system in eukaryotes and highlighted the importance of taking both coding and non-coding transcripts into account while studying transcriptional regulation in eukaryotic genomes.
Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on presentation of Theses.
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4KR</td>
<td>Four lysine-to-arginine mutant</td>
</tr>
<tr>
<td>4KQ</td>
<td>Four lysine-to-glutamine mutant</td>
</tr>
<tr>
<td>AS</td>
<td>Antisense</td>
</tr>
<tr>
<td>CDS</td>
<td>Coding domain sequence</td>
</tr>
<tr>
<td>ChIP</td>
<td>Chromatin immunoprecipitation</td>
</tr>
<tr>
<td>CUT</td>
<td>Cryptic unstable transcript</td>
</tr>
<tr>
<td>ESR</td>
<td>Environmental stress response</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>HAT</td>
<td>Histone acetyltransferase</td>
</tr>
<tr>
<td>HDAC</td>
<td>Histone deacetylase</td>
</tr>
<tr>
<td>IGR</td>
<td>Intergenic region</td>
</tr>
<tr>
<td>IncRNA</td>
<td>Long non-coding RNA</td>
</tr>
<tr>
<td>ncRNA</td>
<td>Non-coding RNA</td>
</tr>
<tr>
<td>NDR</td>
<td>Nucleosome depleted region</td>
</tr>
<tr>
<td>NNS</td>
<td>Nrd1-Nab3-Sen1 complex</td>
</tr>
<tr>
<td>NUT</td>
<td>Nrd1-unterminated transcript</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>OE</td>
<td>Overexpression</td>
</tr>
<tr>
<td>PIC</td>
<td>Pre-initiation complex</td>
</tr>
<tr>
<td>Pol II</td>
<td>RNA polymerase II</td>
</tr>
<tr>
<td>RNAi</td>
<td>RNA interference</td>
</tr>
<tr>
<td>RPKM</td>
<td>Reads per kilobase per million mapped reads</td>
</tr>
<tr>
<td>snoRNA</td>
<td>Small nucleolar RNA</td>
</tr>
<tr>
<td>SRT</td>
<td>Ssu72-repressed transcript</td>
</tr>
<tr>
<td>SUT</td>
<td>Stable unannotated transcript</td>
</tr>
<tr>
<td>TM</td>
<td>Transmembrane</td>
</tr>
<tr>
<td>TES</td>
<td>Transcription end site</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcription start site</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
<tr>
<td>WT</td>
<td>Wild type</td>
</tr>
<tr>
<td>XUT</td>
<td>Xrn1-unstable transcript</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The ability to orchestrate gene activity is an essential feature for all living organisms and is closely related to the organism’s versatility. Regulation of gene expression takes place at all stages from transcription to the post-translational modification of the protein product. Understanding how transcription is regulated has been the interest of many researches for a long time. Pioneering studies established that the prokaryotic transcription is primarily regulated by transcription factors (TFs) that bind to specific DNA sequences at *cis*-regulatory elements to activate or repress the target gene. In the post-genomic era, researchers had been able to build gene regulatory networks based on TF binding and reliably simulate gene expression in prokaryotic organisms (Babu, 2008). However, such TF-based networks failed to accurately predict eukaryotic gene behaviours, indicating the existence of other factors. Indeed, eukaryotic genomes are much larger and far more complex and require more advanced mechanisms to modulate their transcription. Besides the larger pool of available TFs and co-factors, it has become clear that chromatin organisation also plays significant roles. Insights into chromatin organisation would be of vital importance in understanding how transcription is regulated in eukaryotic organisms, which would be useful in many areas such as discovering cures for gene-misregulation related human diseases.

Eukaryotic DNA is packaged into a highly organised structure called chromatin. Every 145-147 bp of DNA are wrapped by a histone octamer core that contains two copies of H2A, H2B, H3 and H4, forming the repeating unit of chromatin called nucleosome. Nucleosomes are connected by short DNA segments forming a beads-on-a-string array with a diameter of 11-nm (Luger et al.,
The beads-on-a-string structure affects how genes are transcribed in different ways. Firstly, the nucleosomes create a physical barrier between the DNA and transcriptional machinery that needs to be displaced in a controlled manner during transcriptional initiation and elongation. Also, each histone within the nucleosome can be post-transcriptionally modified at many residues. The post-translational modifications not only modulate the chemical property of the nucleosome but also create interfaces for the binding of chromatin remodelling complexes. In addition, canonical histones can be substituted by histone variants at specific genomic locations such as gene promoters, which would affect the transcription of specific genes. All these factors add new levels of complexity to the eukaryotic genome, which enables high degrees of control over their gene activities. In this chapter, we will cover the current understanding of chromatin organisation in eukaryotic cells and its close relationship with transcription. Since our study was conducted on the budding yeast \textit{S. cerevisiae}, we will emphasise \textit{S. cerevisiae} more than other eukaryotic organisms.

### 1.1 Histone Modifications

Histones undergo extensive post-translational modifications (PTMs) in all eukaryotic organisms. PTMs on over 60 different amino acid residues have been detected by mass spectrometry or specific antibodies (Kouzarides, 2007). How these PTMs contribute towards chromatin organisation has been an interesting topic. Up to now, two different mechanisms have been characterised. The first mechanism is that PTMs directly exert structural perturbation on their containing nucleosome. For example, lysine acetylations effectively neutralise the positive charge on lysine and presumably weaken the charge-charge interactions between DNA and nucleosome, thereby leading to easier access of DNA-binding proteins (Venkatesh and Workman, 2015). The second mechanism of PTM is to create interfaces for the binding of other chromatin factors. Many histone modifications can be recognised by specific protein domains such as Bromo and PHD (Table 1.1) which are usually associated with multi-subunit protein complexes that contain other chromatin remodelers.

The enzymes that catalyse or reverse the PTMs can also affect chromatin organisation. Acetylation and methylation are amongst the most well-characterised histone modifications and nearly all the enzymes that catalyse and reverse the
modifications have been identified (Table 1.1). Acetylations are added or removed by histone acetyltransferases (HATs) and deacetylases (HDACs). Each HAT or HDAC is capable of modifying multiple lysine residues. In contrast, methylation and demethylation are usually catalysed by residue-specific histone methyltransferases and demethylases (Kouzarides, 2007). Histone-modifying enzymes are often associated with multiprotein complexes (e.g. Gcn5 in the SAGA complex) which can work in conjunction with other chromatin remodellers or transcriptional machineries.

<table>
<thead>
<tr>
<th>PTM</th>
<th>Position</th>
<th>Enzyme</th>
<th>Rev. By</th>
<th>Recog. By</th>
</tr>
</thead>
<tbody>
<tr>
<td>H2A</td>
<td>K4 K7</td>
<td>Esa1</td>
<td>Rpd3</td>
<td></td>
</tr>
<tr>
<td>H2B</td>
<td>K11 K16</td>
<td>Gcn5, Esa1</td>
<td>Rpd3, Hda1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>K9</td>
<td>Gcn5, Rtt109</td>
<td></td>
<td>Bromo, PHD</td>
</tr>
<tr>
<td></td>
<td>K14</td>
<td>Gcn5, Hpa2, Esa1, Elp3, Sas2, Sas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3</td>
<td>K18</td>
<td>Gcn5</td>
<td>Rpd3, Set3, Hda1, Hst1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>K23</td>
<td>Gcn5, Sas3</td>
<td>Hos2, Hst1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>K27</td>
<td>Gcn5, Rtt109</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>K36</td>
<td>Gcn5, Rtt109</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>K56</td>
<td>Rtt109</td>
<td>Hst3/4</td>
<td>PH, (Snf5)*</td>
</tr>
<tr>
<td>H4</td>
<td>K5</td>
<td>Esa1</td>
<td>Rpd3, Set3, Hos2, Hst1</td>
<td>Bromo</td>
</tr>
<tr>
<td></td>
<td>K8</td>
<td>Esa1, Elp3, Gcn5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>K12</td>
<td>Hat1, Esa1, Hpa2</td>
<td>Sir2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>K16</td>
<td>Sas2, Esa1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-me</td>
<td>H3</td>
<td>Set1</td>
<td>Jhd2</td>
<td>Tudor, Chromo, ADD, PHD MBT, Zf-CW, WD40, PHD, PWWP, Chromo</td>
</tr>
<tr>
<td></td>
<td>K36</td>
<td>Set2</td>
<td>Rph1</td>
<td></td>
</tr>
<tr>
<td>H4</td>
<td>K79</td>
<td>Dot1</td>
<td>**</td>
<td>Tudor</td>
</tr>
</tbody>
</table>

Table 1.1: Histone acetylation and methylation in *S. cerevisiae*. Table is adapted from (Smolle and Workman, 2013). Enzymes known to catalyse and remove, and the protein domains that recognise the histone modification are shown for *S. cerevisiae*. H3K56ac is recognised by Snf5 but no domain for recognition has been identified (*). Enzymes that demethylate H3K79 has not been identified in *S. cerevisiae*(**).
1.2 Histone Variant H2A.Z

Eukaryotic organisms also incorporate histone variants at specific genomic locations to further increase the diversity of chromatin organisation. Similar to histone modifications, histone variants affect chromatin organisation by either altering the stability of nucleosomes or facilitating the binding of other chromatin factors. Amongst all the characterised histone variants, H2A.Z is particularly interesting since it is highly conserved across all eukaryotic organisms with the sequence identity of approximately 90% (Iouzalen et al., 1996). It has $\sim$60% sequence similarity to the canonical H2A and H2A.Z-containing nucleosomes have largely similar structures to the H2A-containing ones (Suto et al., 2000). Its universal presence and high degree of conservation across all eukaryotic species implies its biological importance but its exact functions have not been fully understood.

Previous studies generally supported a destabilising role of H2A.Z on its containing nucleosome. Crystallography reported subtle destabilisation of the interaction between the H2A.Z/H2B dimer and the H3/H4 tetramer (Suto et al., 2000). Nucleosome core particles reconstituted with recombinant H2A.Z/H2B dimers also exhibited reduced stability (Abbott et al., 2001). H2A.Z can be acetylated at 4 lysine sites in its N-terminal tail redundantly by the histone acetyltransferase Esa1 and deacetylated by Hda1 (Millar et al., 2006; Mehta et al., 2010). Acetylation of H2A.Z is associated with promoters of actively transcribed genes (Millar et al., 2006). These findings suggested that H2A.Z is likely a destabilising factor for nucleosomes, which presumably leads to easier access of transcription factors and gene activation.

1.3 Chromatin Organisation in Gene Vicinity

With the advancement in high-throughput genome mapping technologies such as ChIP-chip and ChIP-seq, information on the genomic distribution of histone modifications and variants has increased dramatically in the past decade. Specific patterns of histone marks (i.e. modifications and variants) have been identified at various genomic features such as the promoters, enhancers, replication origins, heterochromatin regions, centromeric and telomeric regions. The chromatin
organisation at genic regions such as the promoter and coding region, is of particular interest since it is closely related to transcription. Transcription occurs in three stages - initiation, elongation and termination. In each stage, chromatin undergoes significant changes with the involvement of many chromatin remodelling complexes that also interact with transcriptional machineries. Regulation of chromatin organisation is of vital importance in producing normal levels of transcripts, which has profound impact on the characteristics of the cell. In this section, we will discuss the patterns of histone marks in genic regions and their functions in transcriptional regulation.
1.3.1 Patterns of histone marks at promoters

Promoters are regulatory regions that are located upstream of the transcription start sites (TSSs) of genes. High-resolution nucleosome maps have shown that RNA polymerase II (Pol II)-dependent promoters tend to have a nucleosome depleted region (NDR) flanked by two well positioned nucleosomes (Fig. 1.1 A). NDR is established and maintained at the AT-rich region with promoters by ATP-dependent chromatin remodelers. In *S. cerevisiae*, the +1 nucleosomes are generally well positioned at TATA-less promoters and are "fuzzy" at TATA-containing promoters (Ioshikhes et al., 2006). Histone modifications that are located at promoters include H3K4 methylations, H3K27me3 (only in higher eukaryotes) and all H3/H4 acetylations. H3K4me3 and H3/H4 acetylations have highest occupancy at the -1 and +1 nucleosomes, both of which strongly correlate with gene expression in *S. cerevisiae* and higher eukaryotes (Millar and Grunstein, 2006; Rando and Chang, 2009; Smolle and Workman, 2013). Individual H3/H4 acetylations tend to have similar patterns to each other, which is consistent with the fact that many lysine sites are acetylated by the same enzyme, and all H3/H4 acetylations correlate with transcription. H3K27me3 is only present in high eukaryotic genomes and marks the promoters and coding regions of inactive genes. Notably, H3K4me1 and H3K4me2 are both slightly enriched at inactive promoters in *S. cerevisiae*, which is different from higher eukaryotes (Fig 1.1 B). The genomic distribution of histone H2A.Z has been characterised in many species. H2A.Z is highly enriched at the -1 and +1 nucleosomes around the NDR in both *S. cerevisiae* (Albert et al., 2007) and higher eukaryotes (Bruce et al., 2005; Barski et al., 2007; Gevry et al., 2007; Zilberman et al., 2008). The deposition of H2A.Z is catalysed by Swr1 in *S. cerevisiae*, which replaces H2A-H2B dimers with H2A.Z-H2B in a stepwise manner (Luk et al., 2010). Swr1 forms the catalytic part of the SWR complex that is evolutionarily conserved across many eukaryotic species (Lu et al., 2009).

However, the correlation between H2A.Z and gene expression is not simple. In higher eukaryotes, H2A.Z is enriched at promoters of active genes and its levels are correlated with gene expression, H3K4me3 and Pol II (Barski et al., 2007; Hardy et al., 2009) and are anti-correlated with the repressive marks H3K9me2/3 and DNA methylation (Zilberman et al., 2008; Dryhurst et al., 2009). Knockdown of H2A.Z causes a reduction of Pol II that is recruited to the promoter (Hardy et al., 2009). H2A.Z is also depleted from stably repressed genes that are marked only by
CHAPTER 1. INTRODUCTION

Figure 1.1: Histone modifications and variants at promoters of genes. (A) Histone marks in actively transcribed genes. Promoters of actively transcribed genes have well positioned nucleosomes at -1 and +1 positions around the NDR. H2A.Z is deposited at both nucleosomes. Histone modifications enriched at active promoters include the H3 and H4 acetylations (H3K9ac, H3K14ac, H3K18ac, H3K56ac, H4K5ac, H4K8ac, H4K12ac and H4K16ac), H3K4me3. In higher eukaryotes, active promoters are also enriched for H3K4me2 (*). Coding regions of active genes are enriched for H3K36me3, H2B monoubiquitylation and H3K79 methylations. (B) Inactive promoters tend to be occupied by nucleosomes but some inactive promoters also have stable NDRs (**). Inactive promoters in higher eukaryotes are enriched for H3K27me3 (*). Coding regions are enriched for H3K9me2 and H3K9me3 in higher eukaryotes (*) and enriched for H3K4me1 and H3K4me2 in _S. cerevisiae_.

H3K27me3 but is enriched at poised genes with both H3K27me3 and H3K4me3 (Ku et al., 2012). Consistently, H2A.Z is required for the induction of androgen receptor regulated genes (Draker et al., 2011) and differentiation-related genes (Creyghton et al., 2008). These evidence supported a predominantly activating role of H2A.Z on transcription in higher eukaryotes.

In _S. cerevisiae_, H2A.Z is found at the promoters of both active and inactive genes and some studies even reported anti-correlations between the promoter occupancy and gene expression (Guillemette et al., 2005; Li et al., 2005; Zhang
et al., 2005; Millar et al., 2006). However, the consensus of H2A.Z is still a transcriptional activator since it is required for the activation of many stress-response genes (Santisteban et al., 2000; Adam et al., 2001; Larochelle and Gaudreau, 2003). Moreover, deletion of H2A.Z affects the induction of genes but not the repression in response to heat shock (Zhang et al., 2005). H2A.Z may also promote transcription elongation since the deletion mutant results in slower rate of elongation and increased nucleosome density (Santisteban et al., 2011). Despite the differences in correlation with expression, both *S. cerevisiae* and higher eukaryotes highlighted the importance of H2A.Z in gene activation during transcriptional responses.

1.3.2 Remodelling of promoter during gene activation

During gene activation, promoter organisation undergoes significant changes from tightly packed to a relatively open state that can be accessed by DNA-binding proteins. The first barrier to overcome is the nucleosome present at the promoter (Fig 1.2 A). During gene activation, promoter nucleosomes are acetylated at many lysine residues, resulting in weakened interactions between DNA and the nucleosome. In addition, promoters of active genes often contain an AT-rich region which also disfavours nucleosome binding (Drew and Travers, 1985). The weakened nucleosome-DNA interaction leads to the removal of promoter nucleosome by the ATP-dependent chromatin remodeller RSC (Hartley and Madhani, 2009; Ganguli et al., 2014), creating a nucleosome-depleted region (NDR) of $\sim 140$ bp within the promoter region (Fig 1.2 B). NDRs and histone acetyltransferase complexes both facilitate the SWR-mediated H2A.Z incorporation at the -1 and +1 nucleosomes flanking the NDR (Fig 1.2 C; Luk et al., 2010; Yen et al., 2012; Ranjan et al., 2013). The incorporation of H2A.Z at the +1 nucleosome is essential as it promotes transcription initiation by recruiting chromatin remodellers and other transcription factors to the gene to be activated (Fig 1.2 D; Draker et al., 2012). In higher eukaryotes, the NDR also creates binding sites for the histone chaperone HIRA, which facilitates the incorporation of histone H3.3 to facilitate the destabilisation of promoter nucleosomes (Tagami et al., 2004; Loppin et al., 2005; Jin and Felsenfeld, 2007). In addition, the H2A.Z at +1 nucleosome appears to weaken the nucleosome barrier that prevents Pol II from entering elongation phase and H2A.Z knockdown results in decreased the protein level of the gene (Weber et al., 2014).
Figure 1.2: Chromatin remodelling at promoters during transcription initiation. (A) Promoters of inactive genes tend to have DNA bound by unacetylated nucleosomes, forming a structure that is inaccessible by DNA-binding proteins. (B) During gene activation, promoter nucleosomes are acetylated at many lysine residues. The nucleosome that is bound to the AT-rich region is destabilised, leading to its removal by the ATP-dependent chromatin remodeller RSC and the establishment of NDR. (C) NDR along with acetylation facilitates the incorporation of histone variants H2A.Z and H3.3 at -1 and +1 nucleosomes by SWR1 and HIRA respectively. (D) Incorporation of histone variants facilitate the formation of pre-initiation complex and lead to transcriptional activation.

1.3.3 Chromatin organisation inside coding regions

With high-resolution genomic mapping, patterns of histone marks inside coding regions of genes have been distinguished from promoters. The coding regions of actively transcribed genes are enriched for H3K36me3, H2B mono-ubiquitylation (H2Bub1) and H3K79 methylations in all eukaryotes (Fig 1.1 A; Smolle and Workman, 2013). The co-localisation between H2Bub1 and H3K79 methylations is not unexpected because H2Bub1 was known to be a prerequisite for the methylations on H3K79 in S. cerevisiae (Ng et al., 2002). The coding regions of inactive genes in higher eukaryotes are enriched for H3K27me3, H3K9me2 and H3K9me3 (Fig 1.1 B). H3K27me3 is associated with silenced genes in euchromatic regions and H3K9me2/3 are associated with constitutive heterochromatin (Ho et al., 2014).

H2A.Z is depleted from the coding regions of actively transcribed genes in both S. cerevisiae (Albert et al., 2007) and higher eukaryotes (Barski et al., 2007; Zilberman et al., 2008; Hardy et al., 2009), which is likely due to the fact that
CHAPTER 1. INTRODUCTION

H2A.Z-containing nucleosomes are actively replaced with H2A-containing ones by the INO80 complex (Papamichos-Chronakis et al., 2011). In *S. cerevisiae*, H2A.Z is enriched in the coding regions of subtelomeric genes (Guillemette et al., 2005), which acts as a barrier to protect the subtelomeric euchromatin from the spread of telomeric heterochromatin (Meneghini et al., 2003). In human, H2A.Z is abundant in the coding regions of genes within facultative heterochromatin (Hardy et al., 2009). However, whether H2A.Z contributes towards shaping the heterochromatin or is merely a consequence of the absence of transcription remains to be established. Interestingly, some genome-wide studies in *S. cerevisiae* observed genes outside subtelomeric regions with H2A.Z enrichment in their coding regions (Guillemette et al., 2005; Raisner et al., 2005), the functions of which have not been addressed much by previous studies. We will discuss the significance of these coding-region H2A.Z in the result chapters.

1.3.4 Resetting chromatin after elongation

The chromatin marks inside coding regions are essential in maintaining the genome integrity during transcription elongation. The passage of elongating Pol II leaves chromatin in an open state that is prone to undesired transcription from cryptic promoters. Cryptic transcription is suppressed by various resetting mechanisms to ensure the chromatin returning to the normal state after the passage of Pol II. As Pol II passes through the coding region, ubiquitylated H2A-H2B dimers are re-deposited into the chromatin by the FACT complex (Fig 1.3 A; Pavri et al., 2006). Redeposited ubiquitylated H2B interacts with the chromatin remodelling complex Chd1 to restore the spacing between nucleosomes (Radman-Livaja et al., 2012). H3K36me3 is also co-transcriptionally added to the coding region by the histone methyltransferase Set2 (Strahl et al., 2002; Pokholok et al., 2005). H3K36me3 inhibits the ability of histone chaperones such as Asf1 to exchange deacetylated nucleosomes with acetylated ones, leaving the coding region in a hypoacetylated state (Fig 1.3 B; Venkatesh et al., 2012). The H3K36me3-containing nucleosomes also bind the chromatin remodelling complex Isw1b via a PWWP domain, which promotes the retention of nucleosomes and facilitates proper spacing (Smolle et al., 2012; Maltby et al., 2012). In addition, Set2 interacts with the histone deacetylase complex Rpd3S to actively deacetylate the coding region, further ensuring the hypoacetylation (Fig 1.3 C; Govind et al., 2010). The H3K36me3-enriched, hypoacetylated and properly spaced nucleosomes are of
Figure 1.3: Restoring chromatin structure after transcription. (A) H2B is co-transcriptionally ubiquitylated (Ub). Ubiquitylation facilitates the FACT-mediated reassembly of nucleosome and the recruitment of the chromatin remodeler Chd1 to restore normal nuclear spacing. (B) H3K36me3 is co-transcriptionally added to the coding region by the methyltransferase Set2. H3K36me3 inhibits the histone chaperone Asf1’s ability to incorporate acetylated histones into chromatin. H3K36me3 also recruits the chromatin-remodelling complex Isw1b, which functions with Chd1 to restore nucleosome spacing. (C) H3K36me3 activates the histone deacetylase complex Rpd3S, which cooperate with Set2 to remove any acetylation marks from the coding region. (D) Failure in restoring the coding-region nucleosomes to a H3K36me3-enriched, hypoacetylated and properly spaced organisation results in production of transcripts from cryptic promoters.

vital importance in suppressing cryptic transcription since the Set2 and Rpd3S mutants both result in hyperacetylation of coding regions and increased levels of non-coding transcripts originated from cryptic promoters (Fig 1.3 D; Smolle and Workman, 2013). However, whether H2A.Z is involved in the regulation of non-coding RNA remains largely unclear. Given that H2A.Z is actively removed from the coding regions of actively transcribed genes, it is reasonable to speculate that H2A.Z eviction may also contribute towards the suppression of cryptic transcription.

1.4 Non-coding RNA Transcripts

It was originally thought that eukaryotic transcriptomes are primarily composed of protein-coding transcripts. With the advancement of detection techniques,
it is becoming clear that the majority of eukaryotic transcriptome is made of non-coding RNAs (ncRNAs). Characterised ncRNAs include the RNAs comprising translation machinery (i.e., ribosomal RNA, tRNA, RNase P, SRP-7S), small regulatory RNAs, such as microRNAs, small nucleolar RNAs (snoRNAs) and long non-coding RNAs (lncRNAs). lncRNA is an interesting class since they are present in all eukaryotic organisms. Although they were originally considered as transcriptional noises, in recent years more and more evidence suggests them being functional. Similar to mRNAs, most lncRNAs are 5’-capped and 3-polyadenylated. In addition, lncRNA promoters also have strikingly similar chromatin features to protein-coding genes, namely that active lncRNA promoters are enriched for H3K4me3 and histone acetylations and inactive promoters are enriched for H3K27me3 (Derrien et al., 2012). lncRNA promoters are evolutionarily conserved in human and mouse (Guttman et al., 2009; Derrien et al., 2012) and the number of lncRNAs in each species correlates with the organism’s complexity better than the numbers of protein-coding genes do (Mattick, 2010), which implies the biological importance of lncRNAs. lncRNAs are originated from various genomic locations such as promoters, coding regions of genes, intergenic regions and enhancers. Their genomic location and configuration with nearby genes are closely related to their functions. In this section, we will discuss different types of lncRNAs and how chromatin organisation is related to their transcription.

### 1.4.1 Intergenic lncRNA

Enhancers are cis-DNA elements located at intergenic regions or introns that regulate the expression of genes from distance in higher eukaryotes. Non-coding RNA transcripts, or enhancer RNAs (eRNAs) are produced from enhancer elements. The eRNAs comprise 15-20% of all lncRNA and their levels are correlated with the mRNA of their gene targets (Kim et al., 2010; De Santa et al., 2010). The eRNAs are important in cell signalling, differentiation and development in higher eukaryotic organisms (Rinn et al., 2007; Wang et al., 2011b; Trimarchi et al., 2014). lncRNAs transcripts generated from non-enhancer intergenic regions are known as long intergenic non-coding RNAs or lincRNAs. Most lincRNAs are polyadenylated and tend to overlap with repetitive elements (Ulitsky et al., 2011; Kelley and Rinn, 2012). While the functions of lincRNAs are largely unknown, there have been well-characterised functional lincRNAs such as the X-inactive
specific transcript and the telomerase reverse transcriptase (Feng et al., 1995; Sun et al., 2006).

In *S. cerevisiae*, well-characterised intergenic IncRNAs are usually transcribed in *cis* from upstream of the target gene. These IncRNA transcripts tend to overlap with the promoter of their target gene and repress the expression of their target by various mechanisms. Examples of the IncRNA:gene pairs include *SRG1:SER3* (Martens et al., 2004), *ZRR1:ADH1* (Bird et al., 2006), *ICR1:FLO11* (Bumgarner et al., 2009) and *IRT1:IME1* (van Werven et al., 2012). We will readdress this type of IncRNA in Chapter 4 and discuss the mechanisms of how they regulate their target genes.

### 1.4.2 Promoter-associated IncRNA in higher eukaryotes

The largest and most well-studied class of IncRNAs are those associated with the promoter of protein-coding genes. Most eukaryotic promoters are intrinsically bidirectional but surprisingly, only transcripts in the sense direction are stably produced. It was recently revealed in higher eukaryotes that transcripts in the antisense (AS) direction are in fact transcribed but the majority of the nascent transcripts are rapidly degraded due to the polyadenylation sites enriched in the AS direction (Ntini et al., 2013). These promoter-associated AS transcripts are termed upstream antisense RNA or uasRNA (Grzechnik et al., 2014). The uasRNAs are classified into short uasRNAs, which are typically <100 nt, and promoter upstream transcripts or PROMPTs, which are >200 nt. Short uasRNAs are products of degradation of the nascent transcript protected by paused Pol II (Valen et al., 2011) whereas PROMPTs are produced by Pol II that has escaped from pausing (Kapranov et al., 2007). PROMPTs are 5’ capped, polyadenylated and are rapidly degraded by exosome (Preker et al., 2008, 2011).

### 1.4.3 Promoter-associated IncRNA in *S. cerevisiae*

In *S. cerevisiae*, antisense IncRNAs are also produced from the promoters of genes and their levels are restricted by various mechanisms. Nascent antisense RNAs are targeted by the Nrd-Nab3-Sen1 (NNS) complex, which recruit the TRAMP complex to the transcript (Creamer et al., 2011). The TRAMP complex adds poly(A) to the AS transcript and also stimulates the 3’-5’ degradation by exosome (Fig 1.4 A; Vanacova et al., 2005; Callahan and Butler, 2010). Some
of the transcripts that escaped the NNS-triggered degradation can stably exist in wild-type cells and are known as stable unannotated transcripts (SUTs). A genome-wide study reported 847 SUTs with median length of 761 nt (Xu et al., 2009). When deleting the catalytic subunit Rrp6 of exosome that carries out the degradation nascent RNA transcripts, Xu et al. (2009) observed an additional 925 transcripts, also known as cryptic unstable transcripts (CUTs). CUTs are shorter than SUTs, with the median length of 440 nt. Most of the SUTs and CUTs (73%) are produced from bidirectional promoters at either 5’ or 3’ ends of protein-coding genes. The borderline between SUTs and CUTs is not 100% strict as there are 102 SUTs whose levels are slightly increased in the \( rrp6\Delta \) which could be also classified as CUTs. A separate study showed that depletion of Nrd1 (a component of the NNS complex) from the nucleus results in 1526 new transcripts originated from gene promoters (Schulz et al., 2013). They denoted these transcripts as the Nrd1-dependent transcripts or NUTs. The vast majority (\(~80\%) of NUTs are overlapping with CUTs or SUTs and are extensions of CUTs or SUTs (Fig 1.4 D).

Non-coding transcripts that escaped from degradation in the nucleus are subject to the 5’-3’ degradation by exonuclease Xrn1 in the cytoplasm (Fig 1.4 B; Thompson and Parker, 2007). In a mutant lacking the exonuclease Xrn1, 1658 Xrn1-dependent unstable transcripts or XUTs were detected (van Dijk et al., 2011). XUTs are usually originated from the 3’ end of protein-coding genes and overlap with the entire ORF of the gene (Fig 1.4 D). XUT levels anti-correlates with their cognate protein-coding genes, suggesting their potentially silencing role on sense transcripts. Genes with expressed XUTs can be de-repressed when the methyltransferase Set1 for H3K4 methylations is also deleted. Based on these observation, van Dijk et al. (2011) proposed that XUTs may contribute towards the demethylation at gene promoters by some unknown silencing factors to achieve the repression of sense transcripts. Also, It is worth noting that XUTs are not mutually exclusive from SUTs or CUTs.

Another type of promoter-associated lncRNA is suppressed by the Ssu72-mediated gene looping (Fig 1.4 C; Tan-Wong et al., 2012). In the Ssu72 inactivation mutant, gene loops are disrupted which gives rise to increased levels of Ssu72-restricted transcripts or SRTs. SRT transcription is suppressed in Ssu72\(^{+}\) cells and SRTs are mutually exclusive from SUTs, CUTs or NUTs. SRTs expressed in \( ssu72\Delta \) mutants are susceptible to exosome digestion and can be stabilised
CHAPTER 1. INTRODUCTION

Figure 1.4: Types of long non-coding RNAs in S. cerevisiae. (A) Termination and degradation of promoter-associated antisense (AS) transcripts in nucleus. Nascent AS transcripts are recognised by the Nrd1-Nab3-Sen1 (NNS) complex. NNS recruits the TRAMP complex, which poly-adenylates the RNA tail and also triggers 3'-5' exosomic activity. (B) In cytoplasm, decapped non-coding RNAs are degraded by the 5'-3' exonuclease Xrn1. (C) Antisense transcripts are suppressed by Ssu72-mediated gene looping. (D) Relative location of non-coding RNA transcripts to protein-coding genes. Note that the diagram only shows the most likely locations. For example, most CUTs are upstream of genes. There are CUTs within the coding regions but few in number.

by RRP6 deletion. Other lncRNAs are condition-specific, such as the meiotic unannotated transcripts or MUTs and the respiring or sporulating associated SUTs or rsSUTs (Lardenois et al., 2011). MUTs and rsSUTs can be originated from either 5' promoters or 3' promoters (Fig 1.4 D). Although large numbers of lncRNAs have been identified from high-throughput transcriptomic data, very few have been functionally characterised. Identification of lncRNA functions is still an ongoing area for current research.
1.4.4 Promoter bidirectionality

It is clear that the levels of promoter-associated AS transcripts are reduced post-transcriptionally by RNA degradation. This leads to the question whether their transcription is restricted at chromatin level. It has been recently shown that two pre-initiation complexes (PICs) flanking the NDR are assembled in opposite direction at most bidirectional promoters (Fig 1.5; Murray et al., 2012; Rhee and Pugh, 2012). Interestingly, the occupancy levels of the two PICs do not correlate well with each other and the sense and AS transcripts sharing the same NDR only show moderate correlations, which leads to the speculation that the initiation of sense and AS transcription are independently regulated (Churchman and Weissman, 2011; Rhee and Pugh, 2012). In a recent study conducted in Drosophila, the H2A.Z at the +1 nucleosome has been shown to anti-correlate with stalled Pol II, indicating that H2A.Z may contribute towards promoter directionality by reducing the energy required to overcome the +1 nucleosome barrier (Weber et al., 2014). Therefore, it would be interesting to investigate the asymmetry of H2A.Z occupancies at the -1 and +1 nucleosomes and their relationship with AS and sense transcripts.
1.5 Project Aims

It is clear that one evolutionarily conserved feature of H2A.Z is to mark the promoters of protein-coding genes. H2A.Z is required for gene activation and is generally considered as a transcriptional activator in all eukaryotic species. However, it is unclear why H2A.Z does not correlate with transcript levels in *S. cerevisiae*. With emerging understanding of long non-coding transcription in eukaryotic cells and their similarity to the transcription of protein-coding genes, it would be interesting to study whether H2A.Z also regulates the transcription of ncRNAs. Eukaryotic promoters are inherently bidirectional. Although two PICs are assembled (possibly independently) at bidirectional promoters, it is still unclear whether the chromatin organisation of -1 and +1 nucleosomes and the transcriptional initiation of sense and AS transcripts are regulated independently. Examining the relative levels of H2A.Z at -1 and +1 nucleosomes and its correlation with transcripts may provide useful information. To address these questions, we chose the budding yeast *S. cerevisiae* as the model for various reasons. First, the *S. cerevisiae* genome contains only one copy of H2A.Z gene (Htz1) and the Htz1 gene is non-essential. This means that deletion strains can be generated to study the effect of H2A.Z on the transcriptome of non-coding RNA. Moreover, H2A.Z and the Swr1 and Ino80 that are required for its localisation are highly conserved in eukaryotic species, which implies that functions of H2A.Z are likely to be conserved.

To reveal non-coding RNA transcripts (in this case CUTs), we first generated the strain that lacks the exosome catalytic subunit Rrp6. We deleted *HTZ1* in both Rrp6+ and Rrp6− backgrounds and profiled the H2A.Z and transcriptome by ChIP-seq and strand-specific RNA-seq respectively. Our first objective is to examine the impact of *HTZ1* deletion on the steady-state levels of IncRNAs and whether this have any effect on the transcripts of protein-coding genes. Since it was known that Htz1 becomes essential when the cell is under environmental stresses, we next aim to investigate the relationship between H2A.Z and IncRNAs during stress response. We introduced transcriptional disturbance by caffeine treatment and measured the H2A.Z and transcriptome profiles over a time course of 120 minutes. Unfortunately, the *htz1Δ* mutant is not viable under caffeine stress. To seek for alternative ways of changing the H2A.Z levels in chromatin, we generated the lysine-to-arginine and overexpression mutants that are non-lethal in caffeine.
Chapter 2

Methods

Only methods that are general to all analyses are described in this chapter. Methods that are specific to individual analysis will be described in the corresponding result section.

2.1 Strains

Yeast strains were generated by Yanin Naiyachit and Catherine Millar.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Description</th>
<th>Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMY772</td>
<td>Wild type</td>
<td>Leu2-730-HTZ1-HpH</td>
</tr>
<tr>
<td>CMY773</td>
<td>rrp6Δ</td>
<td>Leu2-730-HTZ1-HpH RRP6::KanMX</td>
</tr>
<tr>
<td>CMY884</td>
<td>htz1Δ</td>
<td>Leu2-730-htz1D::HpH</td>
</tr>
<tr>
<td>CMY885</td>
<td>htz1Δrrp6Δ</td>
<td>Leu2-730-htz1D::HpH RRP6::KanMX</td>
</tr>
<tr>
<td>CMY666</td>
<td>Wild type</td>
<td>YDS2 + bar1::HIS htz1::TRP + p[HTZ1]</td>
</tr>
<tr>
<td>CMY569</td>
<td>4KR</td>
<td>YDS2 + bar1::HIS htz1::TRP + p[htz1K3,8,10,14R-URA3]</td>
</tr>
<tr>
<td>CMY1057</td>
<td>OE</td>
<td>YDS2 + bar1::HIS htz1::TRP + p[HTZ1]-2µm-URA3</td>
</tr>
</tbody>
</table>

Table 2.1: Yeast strains used in this study.

2.2 Cell culture and DNA library

Yeast culture was prepared by Yanin Naiyachit. For steady-state samples, cells of CMY772 (WT), CMY773 (rrp6Δ), CMY884 (htz1Δ) and CMY885 (htz1Δrrp6Δ) growing in SD medium were harvested from log-phase. For caffeine samples, cells of CMY666 (WT), CMY569 (4KR) and CMY1057 (OE) strains were exposed to 10mM caffeine for 2 hours. Samples were harvested from logarithmic growing cultures in YPD at 0, 30 and 120 minutes. RNA libraries were purified
using the Ribopure Purification Kit, selecting poly-adenylated transcripts only. Contaminating genomic DNA was removed by DNaseI digestion. DNA libraries were prepared by the core facility of University of Manchester. Htz1 ChIP was performed using affinity purified custom αHtz1 (α660) antibodies (Millar et al., 2006). DNA fragments were sonicated to lengths of 150-300 bp. ChIP DNA were amplified by 18 cycles using the Illumina TruSeq ChIP and sequenced by Illumina GAIIx or HiSeq2000. Control datasets were generated from samples not treated by the α660 antibody.

2.3 ChIP-seq Data Analysis

2.3.1 ChIP-seq data processing

Bioinformatic analyses were performed by the author. Single-ended reads were mapped to the *S. cerevisiae* genome assembly sacCer1 using Bowtie2 version 0.12.9 (Langmead and Salzberg, 2012), allowing up to 2 mismatches per sequence and no ambiguously mapped reads. ChIP-seq mappings are summarised in Table 2.2. Ratios between ChIP samples and the input DNA background were calculated at each base pair. Extreme values were removed from both the high and low ends such that the remaining ratios have standard deviation (σ) less than 0.2 of the mean (μ). The background level was set to μ + 1.2 × σ of the remaining intermediate ChIP-to-input ratios and subtracted from the ChIP-seq signal. In the case where background was higher than the ChIP-signal, the value was set to zero. The final ChIP-signal was normalised to the number of reads per million uniquely mapped reads.
<table>
<thead>
<tr>
<th>ChIP-seq Sample</th>
<th>Strain</th>
<th>Len/nt</th>
<th>Tot.Reads</th>
<th>Uniq.Map</th>
<th>% Uniq</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT input DNA</td>
<td>CMY772</td>
<td>100</td>
<td>17.9M</td>
<td>15.3M</td>
<td>85.1</td>
</tr>
<tr>
<td>WT - Rep1</td>
<td>CMY772</td>
<td>100</td>
<td>17.8M</td>
<td>14.7M</td>
<td>82.5</td>
</tr>
<tr>
<td>WT - Rep2</td>
<td>CMY772</td>
<td>100</td>
<td>16.0M</td>
<td>13.2M</td>
<td>82.3</td>
</tr>
<tr>
<td>rrp6Δ - Rep1</td>
<td>CMY884</td>
<td>100</td>
<td>13.7M</td>
<td>12M</td>
<td>87.1</td>
</tr>
<tr>
<td>rrp6Δ - Rep2</td>
<td>CMY884</td>
<td>100</td>
<td>16.8M</td>
<td>14.4M</td>
<td>86.0</td>
</tr>
<tr>
<td>WT caffeine input DNA</td>
<td>CMY666</td>
<td>101</td>
<td>14.8M</td>
<td>12.1M</td>
<td>81.8</td>
</tr>
<tr>
<td>WT caffeine 0min - Rep1</td>
<td>CMY666</td>
<td>101</td>
<td>14.1M</td>
<td>10.6M</td>
<td>75.6</td>
</tr>
<tr>
<td>WT caffeine 0min - Rep2</td>
<td>CMY666</td>
<td>101</td>
<td>15.8M</td>
<td>11.5M</td>
<td>72.6</td>
</tr>
<tr>
<td>WT caffeine 30min - Rep1</td>
<td>CMY666</td>
<td>101</td>
<td>14.7M</td>
<td>12.1M</td>
<td>82.4</td>
</tr>
<tr>
<td>WT caffeine 30min - Rep2</td>
<td>CMY666</td>
<td>101</td>
<td>12.9M</td>
<td>10.5M</td>
<td>80.9</td>
</tr>
<tr>
<td>WT caffeine 120min - Rep1</td>
<td>CMY666</td>
<td>101</td>
<td>13.0M</td>
<td>10.4M</td>
<td>80.2</td>
</tr>
<tr>
<td>WT caffeine 120min - Rep2</td>
<td>CMY666</td>
<td>101</td>
<td>11.7M</td>
<td>9.5M</td>
<td>81.5</td>
</tr>
<tr>
<td>4KR caffeine 0min - Rep1</td>
<td>CMY569</td>
<td>101</td>
<td>13.0M</td>
<td>9.8M</td>
<td>75.4</td>
</tr>
<tr>
<td>4KR caffeine 0min - Rep2</td>
<td>CMY569</td>
<td>101</td>
<td>11.4M</td>
<td>8.5M</td>
<td>74.7</td>
</tr>
<tr>
<td>4KR caffeine 30min - Rep1</td>
<td>CMY569</td>
<td>101</td>
<td>12.9M</td>
<td>10.6M</td>
<td>82.1</td>
</tr>
<tr>
<td>4KR caffeine 30min - Rep2</td>
<td>CMY569</td>
<td>101</td>
<td>12.6M</td>
<td>10.1M</td>
<td>80.4</td>
</tr>
<tr>
<td>4KR caffeine 120min - Rep1</td>
<td>CMY569</td>
<td>101</td>
<td>9.8M</td>
<td>7.8M</td>
<td>79.9</td>
</tr>
<tr>
<td>4KR caffeine 120min - Rep2</td>
<td>CMY569</td>
<td>101</td>
<td>10.6M</td>
<td>8.4M</td>
<td>79.8</td>
</tr>
<tr>
<td>OE caffeine 0min - Rep1</td>
<td>CMY1057</td>
<td>101</td>
<td>11.0M</td>
<td>8.6M</td>
<td>78.3</td>
</tr>
<tr>
<td>OE caffeine 0min - Rep2</td>
<td>CMY1057</td>
<td>101</td>
<td>13.0M</td>
<td>9.2M</td>
<td>71.3</td>
</tr>
<tr>
<td>OE caffeine 30min - Rep1</td>
<td>CMY1057</td>
<td>101</td>
<td>13.8M</td>
<td>10.9M</td>
<td>79.5</td>
</tr>
<tr>
<td>OE caffeine 30min - Rep2</td>
<td>CMY1057</td>
<td>101</td>
<td>12.3M</td>
<td>9.5M</td>
<td>77.8</td>
</tr>
<tr>
<td>OE caffeine 120min - Rep1</td>
<td>CMY1057</td>
<td>101</td>
<td>14.0M</td>
<td>10.9M</td>
<td>77.7</td>
</tr>
<tr>
<td>OE caffeine 120min - Rep2</td>
<td>CMY1057</td>
<td>101</td>
<td>12.6M</td>
<td>9.8M</td>
<td>77.5</td>
</tr>
</tbody>
</table>

Table 2.2: Summary of ChIP-seq mapping. Numbers of total sequenced reads (Tot.Reads) and uniquely mapped reads (Uniq.Map) are rounded down to the nearest 0.1 million.

2.3.2 Peak identification

Based on background-subtracted data, regions with non-zero ChIP signals were isolated. Region containing multimodal ChIP signals were then disassembled by repeatedly raising the base level by 5% of maximum ChIP height at one time. The procedure of raising the base level stopped until the multimodal region is disassembled into at least two regions. For each of the resultant region, the above procedure was repeated until all regions were unimodal. The final set of peaks was selected for those longer than 150 bp.
2.4 Strand-specific RNA-seq Analysis

2.4.1 RNA-seq data processing

Single-ended reads were mapped to the sacCer1 assembly using TopHat2 version 0.10 (Langmead et al., 2009). Parameters allow up to 2 mismatches per read and no ambiguously mapped reads. Anchor length for splice junctions were set to 10 bp. Mappings for RNA-seq data were summarised in Table 2.3. Mapped reads were then segregated into + and - strands. Normalisation was performed such that the total amount of sense-strand RNA was adjusted to 108 arbitrary units and antisense RNA-seq levels were adjusted by the same factor.
### CHAPTER 2. METHODS

<table>
<thead>
<tr>
<th>ChIP-seq Sample</th>
<th>Strain</th>
<th>Len/nt</th>
<th>Tot.Reads</th>
<th>Uniq.Map</th>
<th>%.Uniq</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT - Rep1</td>
<td>CMY772</td>
<td>36</td>
<td>40.7M</td>
<td>31.4M</td>
<td>77.0</td>
</tr>
<tr>
<td>WT - Rep2</td>
<td>CMY772</td>
<td>100</td>
<td>26.5M</td>
<td>23.0M</td>
<td>86.8</td>
</tr>
<tr>
<td>htz1Δ - Rep1</td>
<td>CMT773</td>
<td>36</td>
<td>30.0M</td>
<td>23.1M</td>
<td>77.0</td>
</tr>
<tr>
<td>htz1Δ - Rep2</td>
<td>CMT773</td>
<td>100</td>
<td>37.6M</td>
<td>32.5M</td>
<td>86.5</td>
</tr>
<tr>
<td>rrp6Δ - Rep1</td>
<td>CMY884</td>
<td>36</td>
<td>82.4M</td>
<td>57.0M</td>
<td>69.2</td>
</tr>
<tr>
<td>rrp6Δ - Rep2</td>
<td>CMY884</td>
<td>100</td>
<td>52.8M</td>
<td>39.5M</td>
<td>74.9</td>
</tr>
<tr>
<td>htz1Δrrp6Δ - Rep1</td>
<td>CMY885</td>
<td>36</td>
<td>64.9M</td>
<td>47.6M</td>
<td>73.3</td>
</tr>
<tr>
<td>htz1Δrrp6Δ - Rep2</td>
<td>CMY885</td>
<td>100</td>
<td>35.0M</td>
<td>27.9M</td>
<td>79.8</td>
</tr>
<tr>
<td>WT caffeine 0min - Rep1</td>
<td>CMY666</td>
<td>101</td>
<td>25.7M</td>
<td>20.2M</td>
<td>78.8</td>
</tr>
<tr>
<td>WT caffeine 0min - Rep2</td>
<td>CMY666</td>
<td>101</td>
<td>22.9M</td>
<td>18.6M</td>
<td>81.3</td>
</tr>
<tr>
<td>WT caffeine 30min - Rep1</td>
<td>CMY666</td>
<td>101</td>
<td>13.5M</td>
<td>10.9M</td>
<td>80.2</td>
</tr>
<tr>
<td>WT caffeine 30min - Rep2</td>
<td>CMY666</td>
<td>101</td>
<td>25.9M</td>
<td>21.0M</td>
<td>81.0</td>
</tr>
<tr>
<td>WT caffeine 120min - Rep1</td>
<td>CMY666</td>
<td>101</td>
<td>17.5M</td>
<td>14.2M</td>
<td>81.1</td>
</tr>
<tr>
<td>WT caffeine 120min - Rep2</td>
<td>CMY666</td>
<td>101</td>
<td>16.3M</td>
<td>13.6M</td>
<td>83.5</td>
</tr>
<tr>
<td>4KR caffeine 0min - Rep1</td>
<td>CMY569</td>
<td>101</td>
<td>18.8M</td>
<td>15.0M</td>
<td>80.0</td>
</tr>
<tr>
<td>4KR caffeine 0min - Rep2</td>
<td>CMY569</td>
<td>101</td>
<td>20.9M</td>
<td>16.9M</td>
<td>81.1</td>
</tr>
<tr>
<td>4KR caffeine 30min - Rep1</td>
<td>CMY569</td>
<td>101</td>
<td>19.1M</td>
<td>15.2M</td>
<td>79.8</td>
</tr>
<tr>
<td>4KR caffeine 30min - Rep2</td>
<td>CMY569</td>
<td>101</td>
<td>33.1M</td>
<td>26.7M</td>
<td>80.7</td>
</tr>
<tr>
<td>4KR caffeine 120min - Rep1</td>
<td>CMY569</td>
<td>101</td>
<td>26.2M</td>
<td>20.9M</td>
<td>79.9</td>
</tr>
<tr>
<td>4KR caffeine 120min - Rep2</td>
<td>CMY569</td>
<td>101</td>
<td>19.0M</td>
<td>15.7M</td>
<td>82.6</td>
</tr>
<tr>
<td>OE caffeine 0min - Rep1</td>
<td>CMY1057</td>
<td>101</td>
<td>18.2M</td>
<td>14.7M</td>
<td>81.0</td>
</tr>
<tr>
<td>OE caffeine 0min - Rep2</td>
<td>CMY1057</td>
<td>101</td>
<td>9.6M</td>
<td>7.8M</td>
<td>81.7</td>
</tr>
<tr>
<td>OE caffeine 30min - Rep1</td>
<td>CMY1057</td>
<td>101</td>
<td>21.4M</td>
<td>17.3M</td>
<td>81.1</td>
</tr>
<tr>
<td>OE caffeine 30min - Rep2</td>
<td>CMY1057</td>
<td>101</td>
<td>26.1M</td>
<td>21.2M</td>
<td>81.4</td>
</tr>
<tr>
<td>OE caffeine 120min - Rep1</td>
<td>CMY1057</td>
<td>101</td>
<td>14.8M</td>
<td>11.9M</td>
<td>80.6</td>
</tr>
<tr>
<td>OE caffeine 120min - Rep2</td>
<td>CMY1057</td>
<td>101</td>
<td>22.0M</td>
<td>18.1M</td>
<td>82.1</td>
</tr>
</tbody>
</table>

**Table 2.3: Summary of strand-specific RNA-seq mapping.** Numbers of total (Tot.Reads) and uniquely mapped (Uniq.Map) reads are rounded down to the nearest 0.1 million.

#### 2.4.2 Differential expression

Differentially expressed genes were identified by DESeq2 (Anders and Huber 2010), using corrected p-value < 0.05 and minimum fold-change of 2 fold. Differentially expressed non-coding transcripts were computed by two-tailed t-tests using p-value < 0.1 and minimum fold-change of 2 fold.
2.4.3 Transcriptome assembly

Regions with non-zero RNA-seq signals were first identified for each strand of the strand-specific RNA-seq data. Adjacent regions with less than 20 bp apart were concatenated. Transcripts that are shorter than 100 bp were discarded as transcriptional noises. To identify long non-coding RNA transcripts, assembled transcripts were compared with SGD and Ensembl gene annotations. Transcripts that have >20 bp overlap with annotated protein-coding genes or snoRNAs were filtered out.

2.4.4 Hierarchical clustering

Hierarchical clustering were generated from data matrices by taking the Euclidean distances between vectors.

2.5 Acquisition of Published Data

2.5.1 ChIP data

<table>
<thead>
<tr>
<th>Target</th>
<th>Coverage</th>
<th>Type</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K4me3</td>
<td>Genome</td>
<td>ChIP-chip</td>
<td>Guillemette et al. (2011)</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>Chromosome 3</td>
<td>ChIP-chip</td>
<td>Liu et al. (2005)</td>
</tr>
<tr>
<td>H3K9ac</td>
<td>Genome</td>
<td>ChIP-seq</td>
<td>Cai and Tu (2011)</td>
</tr>
<tr>
<td>H3K9ac</td>
<td>Chromosome 3</td>
<td>ChIP-chip</td>
<td>Liu et al. (2005)</td>
</tr>
<tr>
<td>H3K14ac</td>
<td>Chromosome 3</td>
<td>ChIP-chip</td>
<td>Liu et al. (2005)</td>
</tr>
<tr>
<td>H3K18ac</td>
<td>Genome</td>
<td>ChIP-chip</td>
<td>Xue-Franzen et al. (2013)</td>
</tr>
<tr>
<td>H3K18ac</td>
<td>Chromosome 3</td>
<td>ChIP-chip</td>
<td>Liu et al. (2005)</td>
</tr>
<tr>
<td>H4K12ac</td>
<td>Chromosome 3</td>
<td>ChIP-chip</td>
<td>Liu et al. (2005)</td>
</tr>
<tr>
<td>Pol II</td>
<td>Genome</td>
<td>ChIP-chip</td>
<td>Lefrancois et al. (2013)</td>
</tr>
<tr>
<td>Pol II</td>
<td>Chromosome 3</td>
<td>ChIP-seq</td>
<td>Liu et al. (2005)</td>
</tr>
</tbody>
</table>

Table 2.4: ChIP-chip and ChIP-seq data imported from published studies.

2.5.2 Data processing

ChIP-seq data were processed using the same methods as in-house ChIP-seq datasets. For ChIP-chip data, microarray probe sets with sequence were downloaded from individual study. Probes were mapped to the sacCer1 genome assembly using Bowtie2 version 0.12.9 (Langmead and Salzberg, 2012), allowing up to
2 mismatches per probe sequence. ChIP-chip signals for each probe were assigned to the mapped genomic regions. In the case of overlapping probes, signals were averaged within the overlapping regions.

2.5.3 Gene and transcript annotations

SGD annotation, containing 5750 protein-coding genes, was downloaded from the yeast genome database (www.yeastgenome.org) on Jan 2014. Ensembl annotation, containing 7065 protein-coding genes, dubious ORFs and snoRNAs, was downloaded from the Ensembl database (http://www.ensembl.org/) on Jan 2014. Transcript annotation was downloaded from Xu et al. (2009). SUTs and CUTs were excluded to obtain the 5143 protein-coding transcripts. Due to the missing sequence in the sacCer1 assembly, annotated transcripts on Chromosome 10 from 116,000 bp onwards were off by 210 bp compared to our RNA-seq data. An offset of 210 bp was manually added to these Chromosome 10 genes to match our data. Annotation for TATA-containing and TATA-less genes was downloaded from Basehoar et al. (2004). Gene IDs were then matched to the IDs of transcripts.
Chapter 3

H2A.Z Positively Regulates Antisense Transcripts

3.1 Introduction

Previous studies generally supported the view that H2A.Z (or Htz1 in *S. cerevisiae*) is a transcriptional activator. However, this could not explain why H2A.Z levels at gene promoters anti-correlate with transcription rates in *S. cerevisiae*. In this chapter, we will review the methods and conclusions of previous studies conducted in *S. cerevisiae* and discuss possible explanations for this discrepancy. In due course, we will highlight the key factors that need to be taken into account while studying the relationship between H2A.Z and transcription.

3.1.1 H2A.Z stops spread of telomeric heterochromatin

Meneghini et al. (2003) reported a positive role of H2A.Z in regulating subtelomeric genes. In *S. cerevisiae*, H2A.Z is enriched at the euchromatic regions next to the heterochromatin located at mating-type loci and telomeres. The deletion of *HTZ1* results in down-regulation of genes that are within 35 kb of telomeres, indicating its activating effects on these genes. They further showed that the down-regulation can be recovered by also deleting the component required for heterochromatin formation *SIR2*. Based on these observations, Meneghini et al. (2003) proposed that H2A.Z is an activator of subtelomeric genes by preventing the transcriptional silencing caused by the spread of telomeric heterochromatin.
3.1.2 H2A.Z is required for gene activation

More evidence for its positive effects on transcription came from studies on stress-related genes. Santisteban et al. (2000) and Adam et al. (2001) demonstrated that the \( htz1\Delta \) mutant partially disrupts the activation of \( GAL7, GAL1/10 \) and \( PHO5 \) after galactose induction. Larochelle and Gaudreau (2003) also demonstrated that the loss-of-function C-terminal mutant of H2A.Z disrupts the activation of \( GAL1/10 \) and \( GAL7 \) in response to galactose and \( PUR5 \) in response to 6-azauracil. The effects on gene activation appeared to be related to the SWI/SNF complex, since further deleting the catalytic unit Snf2 of SWI/SNF exacerbates the impaired gene inductions caused by \( htz1\Delta \). On the genomic level, Zhang et al. (2005) revealed that greater H2A.Z loss tends to occur at promoters of highly induced genes during heat shock and the \( htz1\Delta \) mutant affects the expression of induced genes but not repressed genes. Taken together, these studies also supported the activating role of H2A.Z on the expression of genes.

3.1.3 H2A.Z occupancy negatively correlates with transcription rates in \( S.cerevisiae \)

However, the activating role of H2A.Z does not explain why its promoter occupancy anti-correlates with transcription rate. The steady-state transcription rate measured by microarray (Holstege et al., 1998) shows negative correlations with the H2A.Z occupancies at gene promoters (Guillemette et al., 2005; Li et al., 2005; Zhang et al., 2005; Millar et al., 2006) and inside coding regions (Li et al., 2005; Millar et al., 2006). Raisner et al. (2005) further demonstrated that the negative correlation may be attributed to differences in nucleosome occupancy since nucleosome-normalised H2A.Z levels show no correlation with transcription rates. However, why there is no positive correlation between H2A.Z and gene expression remains largely unclear. The lack of correlation could be contributed by the fact that H2A.Z is usually absent from both highly active genes (e.g. ribosomal components) and and repressed genes (e.g. many TATA genes in steady states, Albert et al., 2007).
3.1.4 Data resolution is important for quantification

The discrepancy between its activating role and negative correlation with transcription could be due many factors. First, the resolution of genomic profiling techniques could affect the quantification of H2A.Z. The resolution of ChIP-chip datasets in previous studies ranges from entire ORFs or IGRs (Li et al., 2005; Zhang et al., 2005; Millar et al., 2006; Zanton and Pugh, 2006) to \(~\)50 bp (Guillemette et al., 2005; Raisner et al., 2005). Highest resolution were obtained by Liu et al. (2005), which uses overlapping oligomer probes on Chromosome 3 to reach a 10-bp resolution. ORF/IGR probes could not distinguish individual peaks within the region, which can be overcome by tiling arrays with \(~\)50mer probes. However, the major limitation of \(~\)50mer probes is its limited accuracy in measuring the span of H2A.Z peaks, especially in the \textit{S.cerevisiae} genome where the closest non-overlapping transcripts are separately by merely few bp. In our study, we used ChIP-seq to obtain H2A.Z profiles at single-nucleotide resolutions, which will improve the quantification of H2A.Z in relation to genes. In addition, ChIP-seq also has general advantages such as no hybridisation and saturation issues, over DNA microarrays (Park, 2009).

3.1.5 Accurate gene annotation is essential for studying H2A.Z-transcript relationship

Previous studies implicated the need to dissect H2A.Z occupancies at various genomic features (e.g. promoters and CDSs of genes and IGRs), which requires accurate gene annotations. Early annotations of the protein-coding genes in \textit{S.cerevisiae} were constructed from ORFs, which consists of the genomic coordinates from the start to stop codon. However, recent works on transcriptome assembly revealed that most protein-coding genes contain 5’ and/or 3’ untranslated regions (UTRs) that are up to few hundred base pairs (Nagalakshmi et al., 2008; Xu et al., 2009). The drawback of using the ORF rather than transcript annotation is that H2A.Z could be assigned to the wrong genic feature. The use of transcript annotations will enable us to accurately define the promoter, CDS, 3 end of genes. In addition, accurate annotations will also improve the understanding of genomic contexts of genes such as the distance between adjacent genes, the importance of which will be discussed in this chapter.
3.1.6 Dissecting direct and indirect effects of H2A.Z on transcription

The most common method to study the effect of H2A.Z on transcripts is to generate a knockout strain. However, one of the main difficulties is to distinguish between direct and indirect effects. Direct effect is for example, when H2A.Z contributes towards certain chromatin organisations and in turn affects the transcriptional process of the gene. Indirect effect refers to the situation where the deletion of the HTZ1 gene triggers the induction of repression of certain pathways. Previous studies reported both up- and down-regulated genes in the \(htz1\Delta\) mutant (Meneghini et al., 2003). It is difficult to dissect direct and indirect effects because they occur simultaneously when \(HTZ1\) is deleted. However, if a gene with H2A.Z-occupied promoter is differentially regulated in \(htz1\Delta\), the effect is likely to be directly caused by H2A.Z whereas the effects on genes unoccupied by H2A.Z are likely to be indirect. Dissecting the direct and indirect effects would provide new insights into the mechanism of H2A.Z in transcriptional regulation.

3.1.7 Objectives

In this chapter, we will focus on the function of H2A.Z in steady states rather than during transcriptional disturbance. It is still a mystery why H2A.Z does not correlate with transcript levels, even though being a transcriptional activator. In addition, Guillemette et al. (2005) estimated that \(\sim74\%\) of the H2A.Z is located at promoters of protein-coding genes, which leads to the question how the rest of the H2A.Z is distributed in the genome. Could the H2A.Z outside promoters be an important factor in this puzzle? Using high-resolution genomic profiling technique (ChIP-seq) and accurate annotations of protein-coding transcripts (Xu et al., 2009), we were able to pinpoint the exact locations of H2A.Z across the genome. The term "gene" in this chapter always refers to the entire transcribed region of the gene, which includes both 5’ and 3’ UTRs, rather than the ORF between the start and stop codon. We also integrated with strand-specific RNA-seq data to study the relationship between H2A.Z and transcripts.
3.2 H2A.Z is enriched at 3’ ends of genes

To search for H2A.Z enrichments in relation to genes, we aligned all 5143 protein-coding transcripts annotated by Xu et al. (2009) by their transcription start sites (TSSs) and transcription end sites (TESs). As expected, high abundances of H2A.Z were located around the nucleosome-depleted region (NDR) near TSSs (Fig 3.1 A, left panel). Notably, H2A.Z is more abundant downstream of TSS than upstream (Fig 3.1 A, left panel), which is reflecting the lower nucleosome occupancy at -1 position upstream of the NDR. We estimated that 5’ H2A.Z consists of ~60% of the total abundance (Fig 3.1 B), which largely agrees with previous genome-wide studies (Guillemette et al., 2005; Albert et al., 2007). Surprisingly, we also found an H2A.Z enrichment upstream of the TES of protein-coding genes (Fig 3.1 A, right panel). The TES level is significantly enriched compared to the genomic level and consists of ~6.7% of the total H2A.Z in the genome (Fig 3.1 B, C). However, the H2A.Z peak downstream of TES is likely due to signals at TSSs of adjacent genes in the compact S.cerevisiae genome. To search for enrichments inside coding-regions of genes (CDSs), we also stretched all CDSs, defined as the genic regions that are >300 bp from TSS and TES, to the same length and calculated the average H2A.Z levels at each relative location. We observed no enrichment of H2A.Z inside CDSs (Fig 3.1 A, middle panel). However, CDSs are much longer than TSSs or TESs and still contain ~11% of total H2A.Z in the genome (Fig 3.1 B).

In summary, we observed that in addition to the well-known 5’ enrichment, H2A.Z is also enriched in a 200-bp window immediately upstream of the 3’ ends of protein-coding genes. Although the 3’ ends only contain 6.7% of the total genomic H2A.Z, the H2A.Z signal per bp is significantly higher than in the level in CDSs and IGRs (Fig 3.1 C). An overview of the 5’, CDS and 3’ occupancies of H2A.Z is shown in the heatmap, highlighting clusters of genes with 3’ H2A.Z enrichment (Fig 3.1 D). This leads us to investigate the significance of these H2A.Z signals marking the 3’ ends of genes.
Figure 3.1: Distribution of H2A.Z at genic and intergenic regions. (A) H2A.Z profiles around transcription start sites (TSSs), transcription end sites (TESs) and inside coding regions (CDSs) of protein-coding genes. The 5143 protein-coding genes were aligned at their TSSs/TESs and the average H2A.Z signal per gene was calculated at each base relative to the TSS/TES. For the CDS profile, all coding regions (i.e. >300 bp away from TSS or TES) were stretched to the same length before alignment. (B) H2A.Z abundance at various genomic locations. Windows of 300 bp were used to define the up- or downstream of TSS or TES. CDSs are defined as the ORF regions that are >300 bp away from TSS or TES. Intergenic regions are defined as regions that are >300 bp away from genes. (C) Distribution of length-normalised H2A.Z signals at various genomic locations. The 5'-upstream, 5'-downstream, 3'-upstream and 3'-downstream occupancies are significantly than intergenic levels (two-tailed t-test, p-values: < 1.8 × 10^{-308}, < 1.8 × 10^{-308}, = 5.5 × 10^{-163}, = 1.9 × 10^{-287} respectively) (D) Hierarchically clustered data of H2A.Z profiles. Each column vector represents a protein-coding gene. Vectors consist of three 100-bp windows upstream of TSS, three 100-bp windows downstream of TSS, five stretched windows covering the CDS and three 100-bp windows upstream of TES. Clusters of 3' H2A.Z enrichment were highlighted.
3.3 H2A.Z outside TSS marks active antisense promoters

ChIP-seq data revealed that, in addition to the well-known promoter enrichment, significant amount of H2A.Z is also located at the 3’ end and inside CDSs of genes. These observations lead us to investigate what chromatin feature these H2A.Z signify. In recent years, pervasive non-coding RNA transcription has been revealed in most eukaryotic genomes. Knowing that H2A.Z marks the promoter of protein-coding genes, we next speculated whether H2A.Z is also marking the promoters of non-coding transcripts. To reveal the non-coding transcripts in \textit{S. cerevisiae}, we generated stains lacking the catalytic component of exosome Rrp6 so that the cryptic non-coding transcripts or CUTs are stabilised. However, other types of non-coding transcripts such as the NUTs or XUTs were not detectable in this study.

3.3.1 Cryptic transcripts are revealed in \textit{rrp6\textDelta} strain

As a quality control, we first examined the effect of \textit{rrp6\textDelta} on \textit{S. cerevisiae}. By comparing gene expression between \textit{rrp6\textDelta} and WT, we confirmed that the knock-out has very little effect on the transcriptome, with less than 5% of the genes being differentially regulated by >2 folds (Fig 3.2 A). Certain Gene Ontology (GO) categories such as TM transporter and oxidoreductase were differentially regulated but no down-regulation of transcriptional or translational machinery was observed, indicating that \textit{RRP6} deletion has little impact on cell stability (Fig 3.2 A). As expected, the \textit{rrp6\textDelta} strain shows significantly higher levels of non-coding transcripts on the antisense (AS) strand of DNA compared to WT (Fig 3.2 B). Using >5 normalised units as threshold, we identified an additional 549 AS transcripts (649 in \textit{rrp6\textDelta} versus 100 in WT) by \textit{rrp6\textDelta} (Fig 3.2 C). Interestingly, the level of AS transcripts near 3’ end of genes tends to be higher than near the 5’ end (Fig 3.2 D and E), which is likely due to early termination by other factors such as Nrd1 (Castelnuovo et al., 2014).
Figure 3.2: Cryptic transcripts are revealed in rrp6Δ. (A) Effect of rrp6Δ on gene expression. Pearson correlation between gene expression in WT and rrp6Δ $R = 0.97$. Enriched GO categories (Bonferroni-corrected $p$-value $< 0.01$) are shown on the plot. (B) Heatmap showing the increase in transcript levels on the antisense (AS) strand of DNA in rrp6Δ strain. Box-and-whisker plot shows the significance of the increase (two-tailed $t$-test, $p$-value $= 1.2 \times 10^{-37}$). (C) Number of genes with AS transcripts present (defined as $>5$ normalised units per bp) in WT (100 genes) and rrp6Δ (649 genes). (D) Profiles of AS transcripts around the TSS. (E) Profiles of AS transcripts around the TES. Note: the peak downstream of TES is due to up-regulation of snoRNAs in rrp6Δ and is not investigated further (**).
3.3.2 H2A.Z at 3’ ends of genes marks start sites of AS transcripts

By examining the H2A.Z ChIP-seq and strand-specific RNA-seq data, we identified many examples where the H2A.Z peak at 3’ end of a gene is marking the start site of an AS transcript (Fig 3.3 A). On the genomic level, we identified 954 genes with H2A.Z-occupied 3’ ends, of which an overrepresented proportion (354 genes) also have highly expressed AS transcripts (Fig 3.3 B, C; Fig 3.3 D solid blue bar). Consistently, genes having no AS transcripts tend to have unoccupied 3’ ends by H2A.Z (Fig 3.3 D dashed grey bar). To test the significance of AS-H2A.Z association, we randomly generated 954 150-bp regions from all 3’ ends of genes and tested how many of those are associated with AS transcripts. We repeated this randomisation 100 times and observed that the actual H2A.Z-AS association (354 genes) is significantly higher than random chance (Fig 3.3 E). In addition, the profiles of AS transcripts indicate that genes with H2A.Z-occupied 3’ ends tend to have higher AS transcripts than genes without 3’ H2A.Z (Fig 3.3 F) and the difference is significant (Fig 3.3 G). These observations suggest a strong association between the 3’ H2A.Z and the start sites of AS transcripts.
CHAPTER 3. H2A.Z ACTIVATES AS TRANSCRIPTS

Figure 3.3: Co-localisation between 3’ H2A.Z peaks and AS transcripts. (A) Examples of 3’ H2A.Z peaks marking the start site of AS transcripts. (B) Number of genes with high, low and zero levels of AS transcripts. The 2223 convergent overlapping genes were excluded. AS levels > 5 normalised units are classified as High AS; between 0 and 5 normalised units are classified as Low. (C) Number of genes with high, low and zero levels of 3’ H2A.Z. Genes with 3’ H2A.Z > 100 RPKM are classified as High 3’ H2A.Z; between 0 and 100 RPKM are classified as low 3’ H2A.Z. (D) Concurrence of 3’ H2A.Z and AS transcripts. High 3’ H2A.Z and high AS transcripts (solid blue, 354 genes, overrepresented, p-value = 3.7 × 10^{-44}); High 3’ H2A.Z and no AS transcripts (solid black, 69 genes, underrepresented, p-value = 5.6 × 10^{-89}); No 3’ H2A.Z and high AS transcripts (dashed blue, 150 genes, underrepresented, p-value = 1.4 × 10^{-47}); No 3’ H2A.Z and no AS transcripts (dashed grey, 732 genes, overrepresented, p-value = 1.6 × 10^{-102}). P-values were computed by Fisher’s Exact test. (E) Significance of association between 3’ H2A.Z and AS transcripts. A set of 954 random 150-bp regions were drawn from all 3’ ends and tested for AS association. A region is regarded as AS-associated if the AS RNA level in a 50-bp window downstream is >3 fold higher than upstream. The randomisation was repeated 100 times and the frequencies of AS-associations are plotted as histogram. Dashed line shows the observed 354 AS-associated 3’ H2A.Z peaks. False discovery rate is approximately zero. (F) AS profiles of genes with and without 3’ H2A.Z. Genes are aligned at TES and average AS transcript levels were calculated for each base pair. (G) Significantly higher AS transcript level for genes with high 3’ H2A.Z than genes without 3’ H2A.Z (two-tailed t-test, p-value = 5.4 × 10^{-12}).
3.3.3 H2A.Z-marked 3’-end regions are enriched for active chromatin marks

The co-localisation of 3’ H2A.Z and the start sites of AS transcripts leads us to investigate whether these H2A.Z-marked 3’-end regions show features of active promoters. Chromatin marks at gene promoters have been studied to great extents in *S.cerevisiae*. Liu et al. (2005) conducted a comprehensive set of ChIP-chip experiments on yeast Chromosome 3 genes and revealed distinct patterns of histone modifications at active and inactive promoters. Active promoters are generally enriched for H3K4me3, H3K9ac, H3K14ac, H3K18ac, H4K12ac and PolII and depleted for H3K4me1 and H3K4me2.

To test whether H2A.Z-marked 3’-end regions contain characteristic marks of active promoters, we imported the ChIP-chip datasets from Liu et al. (2005). Since the ChIP-chip data only cover Chromosome 3, we extracted the 125 Chromosome 3 genes from our datasets and focused on these genes. Out of the 125 genes, the ones with H2A.Z-occupied 3’ ends (21 genes) are also enriched for all active histone marks compared to the 37 genes without H2A.Z at their 3’ ends (Fig 3.4 A-L). To compute the significance of the differences, we quantified the levels of each histone modification using a 300-bp window upstream of TES and observed significant differences for all histone marks (Figure 3.4 B, D, F, H, J, L). The ChIP enrichment tends to reach a peak at ~200 bp upstream of the TES, which is consistent with the position of H2A.Z peaks (Fig 3.4 A, C, E, G, I, K; Fig 3.1 A, right panel). Using similar methods, we also tested for histone marks that are depleted from active promoters (H3K4me1 and H3K4me2). As expected, genes with high 3’ H2A.Z have significantly less H3K4me1 at their 3’ ends that gene with zero 3’ H2A.Z (Fig 3.5 A, B). One exception is that H3K4me2 did not show any significant difference (Fig 3.5 C, D). Nevertheless, these findings suggest that the local chromatin organisation of H2A.Z-enriched 3’-end regions is very similar to the features of active promoters.
Figure 3.4: H2A.Z-occupied 3’ ends are enriched for active histone marks. ChIP-on-chip data for Chromosome 3 of *S. cerevisiae* were obtained from Liu et al. (2005). Genes with high levels of 3’ H2A.Z (21 genes) and genes with zero 3’ H2A.Z (37 genes) were examined separately. For each group, genes were aligned according to their 3’ TES and the average level of each histone mark was calculated for each base pair. Solid and dashed lines represent genes with and without 3’ H2A.Z respectively. Windows of 300 bp upstream of TES were used to quantify the levels of histone marks. Significance was tested by two-tailed t-tests. (A) (B) H3K4me3, *p*-value = 0.024. (C) (D) Pol II, *p*-value = 0.017. (E) (F) H3K9ac, *p*-value = 0.0047. (G) (H) H3K14ac, *p*-value = 0.014. (I) (J) H3K18ac, *p*-value = 0.0033. (K) (L) H4K12ac, *p*-value = 0.012.
CHAPTER 3. H2A.Z ACTIVATES AS TRANSCRIPTS

Figure 3.5: Comparison between 3’ H2A.Z and histone marks that are depleted from active promoters. ChIP-on-chip data for Chromosome 3 of S.cerevisiae were obtained from Liu et al. (2005). Genes with high levels of 3’ H2A.Z (21 genes) and genes with zero 3’ H2A.Z (37 genes) were extracted. Windows of 300 bp upstream of TES were used to quantify the levels of histone marks. Significance was tested by two-tailed t-tests. (A) H3K4me1, \( p\text{-value} = 0.012 \). (C) H3K4me2, \( p\text{-value} = 0.46 \).

3.3.4 Abundance of 3’ H2A.Z correlates with AS transcript levels

Previous studies reported either negative or no correlation between 5’ H2A.Z and gene expression (Guillemette et al., 2005; Zhang et al., 2005; Raisner et al., 2005; Millar et al., 2006), which is also observed in our datasets (Fig S3.1 A, B). Here, we are interested in how the 3’ H2A.Z correlates with transcript levels. We extracted the 2574 genes with H2A.Z occupied 3’ ends and divided them into 5 quantile groups of 514 or 515 genes according to their 3’ H2A.Z levels from low to high. For each group, we aligned the genes at their TESs and plotted the average AS profile at each base pair upstream of TES (Fig 3.6 A). To our surprise, we observed a positive correlation indicated by higher H2A.Z abundance accompanied by higher levels of AS transcripts (Fig 3.6 A, B). We also integrated with published genome-wide histone modification data and observed that the 3’ H2A.Z occupancy also positively correlates with other active histone marks such as H3K4me3, H3K9ac and H3K18ac at the 3’ end of genes (Fig 3.6 C-E).
CHAPTER 3. H2A.Z ACTIVATES AS TRANSCRIPTS

Figure 3.6: H2A.Z at 3’ ends of genes correlates with gene activity. Genes with non-zero 3’ H2A.Z occupancy (2574 genes) were divided into five quantiles (∼515 genes) according to their 3’ H2A.Z levels. Windows of 300 bp upstream of TES were used to quantify H2A.Z levels. The boundary limits are 0.0029, 9.67, 36.7 and 104.2 RPKM respectively. (A) AS profiles upstream of TES for the H2A.Z quantile groups. (B) Distributions of the AS transcript levels. (C) H3K4me3 levels at 3’ ends for the five quantiles. ChIP-seq data were obtained from Guillemette et al. (2011). (D) H3K9ac levels at 3’ ends, quantified from ChIP-seq data of Cai and Tu (2011). (E) H3K18ac levels at 3’ ends, quantified from ChIP-on-chip data of Xue-Franzen et al. (2013).

3.3.5 H2A.Z has stronger association with cryptic than stable transcripts

The non-coding RNA detected in our datasets consists of two different types: the stable unannotated transcripts (SUTs), which can stably exist in WT cells and the cryptic unstable transcripts (CUTs), which are revealed in rrp6Δ. We next looked into the two types of ncRNAs separately. Comparing the RNA-seq data of WT and rrp6Δ, we identified 264 genes with expressed CUTs which are defined as having AS transcripts in rrp6Δ only and 110 genes with SUTs which
are defined as having AS transcripts in both WT and \textit{rrp6}\Delta (Fig 3.7 A, B). The majority of CUTs and SUTs are also present in Xu et al. (2009)’s microarray datasets (Fig 3.7 B). Notably CUTs tend to be more highly expressed than SUTs (Fig 3.8 B). We also identified 275 ambiguous AS transcripts which can stably exist in WT but the levels are higher in \textit{rrp6}\Delta (Fig 3.7 B, E). These ambiguous AS transcripts could be partially digested by exosome. Figure 3.7 C, D and E show examples of the CUTs, SUTs and ambiguous AS transcripts we identified from our datasets. Promoters of CUTs tend to have higher H2A.Z occupancy than promoters of SUTs (Fig 3.8 A). Notably, the association between high level of 3’ H2A.Z and CUTs is stronger than the association between high 3’ H2A.Z and SUTs (Fig 3.8 C).

![Figure 3.7: Classification of cryptic and stable transcripts.](image)

- **(A)** Criteria for defining SUTs and CUTs. SUTs are defined as: AS level >5 normalised units in both WT and \textit{rrp6}\Delta; CUTs are defined as: AS level <0.5 normalised units in WT and >5 normalised units in \textit{rrp6}\Delta; Ambiguous AS transcripts are defined as: AS level between 0.5 and 5 normalised units in WT and >5 normalised units in \textit{rrp6}\Delta.
- **(B)** Numbers and percentages of CUTs and SUTs.
- **(C)** Examples of CUTs shown in Genome Browser.
- **(D)** Examples of SUTs.
- **(E)** Examples of ambiguous AS transcripts.
3.4 H2A.Z at 3’ promoters affects expression by regulating AS transcripts

Up to now, we have demonstrated that the H2A.Z is enriched at 3’ ends of protein-coding genes and is likely part of active promoters for AS transcription. We next investigated whether the 3’ H2A.Z is required for the function of AS promoters. To study this, we next generated HTZ1 knockout strains in both Rrp6+ and Rrp6− backgrounds and quantified their transcriptomes by strand-specific RNA-seq. HTZ1 deletion has been studied previously by Meneghini et al. (2003), who demonstrated down-regulation of subtelomeric genes because H2A.Z is an activator to these genes by stopping the spread of telomeric heterochromatin. For the genes outside subtelomeric regions, there has been a debate of whether H2A.Z is an activator or repressor. In the next part of the study, we will focus on the effects of H2A.Z on genes outside subtelomeric regions.
3.4.1 Effects of H2A.Z on transcriptome

As a quality control, we first examined the effect of \textit{HTZ1} deletion on the sense transcriptome. In the Rrp6\textsuperscript{+} background, \textit{HTZ1} deletion results in only 12 up-regulated and 128 down-regulated genes (< 3% of genome) with no enriched Gene Ontology categories (Fig 3.9 A). In the Rrp6\textsuperscript{−} background, \textit{HTZ1} deletion causes more stress-related genes to be differentially regulated compared to Rrp6\textsuperscript{+} (Fig 3.9 B). Notably, a similar set of genes are down-regulated in the Rrp6\textsuperscript{+} and Rrp6\textsuperscript{−} backgrounds (Fig 3.9 C). Also, the changes in gene expression are generally correlated in both Rrp6\textsuperscript{+} and Rrp6\textsuperscript{−} backgrounds (Fig 3.9 D). These results indicate that \textit{HTZ1} deletion have similar effects on the sense transcriptome in both Rrp6\textsuperscript{+} and Rrp6\textsuperscript{−} backgrounds.
Figure 3.9: Effect of HTZ1 deletion on transcriptome in Rrp6+ and Rrp6− backgrounds. (A) Effect of HTZ1 deletion on transcriptome in Rrp6+ background. Pearson correlation = 0.98. No Gene Ontology is significantly enriched (Bonferroni-corrected p-value < 0.01). (B) Effect of HTZ1 deletion on transcriptome in Rrp6− background. Pearson correlation = 0.97. Significantly enriched GO categories (Bonferroni-corrected p-value < 0.01) are shown. (C) Significant overlap between up-regulated genes (Fisher’s Exact test, p-value = 3.9×10−7) and down-regulated genes (Fisher’s Exact test, p-value = 2.7×10−36) in Rrp6+ and Rrp6− backgrounds. (D) Changes of gene expression by HTZ1 deletion in Rrp6+ and Rrp6− are correlated (Pearson correlation = 0.51).
3.4.2 H2A.Z is required for expression of AS transcripts

We next studied the effect of \textit{HTZ1} deletion on AS transcripts. By comparing the AS transcriptome of \textit{htz1}\textsuperscript{Δ}rrp6\textsuperscript{Δ} with \textit{rrp6}\textsuperscript{Δ}, we identified 94 up-regulated and 115 down-regulated AS transcripts (Fig 3.10 B). Out of the 115 down-regulated AS transcripts, a significant proportion (83 genes, 72\%) also have highly occupied 3' promoters (Fig 3.10 A, solid blue bar). On the genomic level, down-regulated AS transcripts tend to have high occupancy of H2A.Z at 3' ends (Fig 3.10 C, D, blue) and conversely, AS transcripts with highly occupied promoters by H2A.Z tend to be down-regulated upon \textit{HTZ1} deletion (Fig 3.10 E). This indicates that the direct effect of H2A.Z on AS transcripts is activating. On the other hand, up-regulated genes tend to have unoccupied promoters by H2A.Z, which indicates that the up-regulation may be caused by indirect effects (Fig 3.10 C, red). These results suggest that the presence of H2A.Z at 3' promoters is required for the expression of AS transcripts and is therefore an activator of AS transcripts.

For the effect of 5' H2A.Z on sense transcripts, we also observed a positive but much weaker effect of H2A.Z. Genes that are down-regulated in \textit{HTZ1} deletion also have higher 5' H2A.Z than genes that are not differentially expressed (Fig 3.11 A, B) and genes with high 5' H2A.Z tend to be more down-regulated than genes with unoccupied 5' ends (Fig 3.11 C). However, the \textit{p-values} are much less significant than those of 3' H2A.Z (Fig 3.11 B, C).
CHAPTER 3. H2A.Z ACTIVATES AS TRANSCRIPTS

Figure 3.10: Positive effect of 3’ H2A.Z on AS transcript levels. (A) Numbers and percentages of up- and down-regulated genes in \( htz1 \Delta rrp6 \Delta \) compared to \( rrp6 \Delta \) strain. Differentially expressed AS transcripts were identified by >1.5 fold in expression and <0.1 in \( p\)-value of t-test. The 738 genes with AS transcripts in either \( htz1 \Delta rrp6 \Delta \) or \( rrp6 \Delta \) strain are shown. (B) Concurrence between down-regulated AS and highly occupied 3’ ends by H2A.Z (solid blue, 83 genes, overrepresented, \( p\)-value = \( 9 \times 10^{-19} \)); down-regulated AS and no 3’ H2A.Z (dashed blue, 18 genes, underrepresented, \( p\)-value = \( 9.3 \times 10^{-14} \)); up-regulated AS and high 3’ H2A.Z (solid red, 17 genes, underrepresented, \( p\)-value = 0.012); up-regulated AS and no 3’ H2A.Z (dashed red, 43 genes, not significant). (C) Fold-change of AS expression upon \( HTZ1 \) deletion against 3’ H2A.Z levels. Genes with low AS transcript levels (<0.5 normalised units) in both \( htz1 \Delta rrp6 \Delta \) and \( rrp6 \Delta \) strains are excluded. Up- and down-regulated genes are indicated in red and blue respectively. (D) Genes with down-regulated AS transcripts have significantly higher 3’ H2A.Z than the rest of the genome (two-tailed t-tests \( p\)-value = \( 1.7 \times 10^{-14} \)). (E) Genes with highly occupied 3’ ends by H2A.Z (954 genes) have significantly more down-regulated AS transcripts than gene with unoccupied 3’ ends (1154 genes) upon \( HTZ1 \) deletion (two-tailed t-tests \( p\)-value = \( 8.6 \times 10^{-46} \)).
Figure 3.11: Positive effect of 5’ H2A.Z on sense transcripts of genes. (A) Fold-change of sense expression upon HTZ1 deletion against 5’ H2A.Z levels. Up-regulated (151 genes) and down-regulated (237 genes) are indicated in red and blue respectively. (B) Down-regulated genes have higher 5’ H2A.Z (two-tailed t-tests p-value $= 8.0 \times 10^{-5}$) and up-regulated genes have lower 5’ H2A.Z (two-tailed t-tests p-value $= 7.1 \times 10^{-12}$) than the rest of the genome. (C) Genes with highly occupied 5’ promoters by H2A.Z (450 genes) are more down-regulated than gene with unoccupied 5’ promoters (528 genes) upon HTZ1 deletion (two-tailed t-tests p-value $= 5.2 \times 10^{-12}$).

3.4.3 AS has negative effects on sense transcripts

Having established that HTZ1 deletion causes AS transcripts to be down-regulated, we next investigated whether down-regulation of AS affects their cognate sense transcripts. Out of the 954 genes with H2A.Z occupied 3’ promoters, 81 have down-regulated AS transcripts by >2 fold (Fig 3.12 A). We focused on these 81 down-regulated AS transcripts and examined their corresponding fold-changes of sense transcripts. It turns out that down-regulated AS tend to be accompanied by up-regulated sense transcripts and the fold-change is significant compared to genes whose AS are not differentially expressed (Fig 3.12 B, blue box). Consistent with this, the overlap between down-regulated AS and increased sense transcripts is significantly overrepresented (Fig 3.12 C, blue bar), whereas the overlap between down-regulated AS and down-regulated sense is much less frequent. These observations indicate that the down-regulation of AS caused by HTZ1 deletion is linked with the up-regulation of their cognate sense transcripts.
Figure 3.12: AS transcripts have negative effects on sense transcripts. (A) Number and percentage of genes with 3’ H2A.Z within those having up- and down-regulated AS transcripts. (B) Genes with 3’ H2A.Z and down-regulated AS (81 genes) have more up-regulated sense and genes with 3’ H2A.Z and down-regulated AS (17 genes) have more down-regulated sense transcripts than those with 3’ H2A.Z and non-differentially-expressed AS (856 genes). P-values from two-tailed t-tests = 0.0030 and 0.044 respectively. (C) Overlaps between up-regulated AS and increased sense (left grey, 3 genes, not significant), up-regulated AS and decreased sense (red, 12 genes, overrepresented, p-value = 0.0098), down-regulated AS and increased sense (blue, 48 genes, overrepresented, p-value = 1.2 × 10^{-4}), down-regulated AS and decreased sense (right grey, 19 genes, underrepresented, p-value = 0.0039). Decreased and increased sense transcripts were defined as sense expression decreased or increased by >0.2 on log2 scale. Significance was tested using Fisher’s Exact test.
So far, we have demonstrated that the H2A.Z at 3’ end of genes has positive effects on the AS transcript, which in turn have negative effects on sense transcripts. Also, the 5’ H2A.Z have weak positive effects on the sense transcript. Based on these observations, we proposed a model that the overall effect of H2A.Z on the sense transcript is dependent on the relative 5’/3’ distribution of H2A.Z on the gene. To test for this model, we divided the genome into 4 classes according to H2A.Z occupancies at their 5’ and 3’ ends. Class 1 genes have H2A.Z-occupied 5’ ends but not 3’ ends; Class 2 genes have H2A.Z at both 5’ and 3’ ends; Class 3 genes have unoccupied promoters and Class 4 genes only have H2A.Z-occupied 3’ ends (Fig 3.13 A). We would expect H2A.Z to have strongest positive effects on Class 1 genes and strongest negative effects on Class 4 genes. By examining the fold-changes of expression upon HTZ1 deletion, we observed Class 4 genes being significantly up-regulated, confirming the negative effects of 3’ H2A.Z on sense transcripts (Fig 3.13 B, blue box). Interestingly, we observed no difference between Class 1 and Class 2, suggesting that the negative effect of 3’ H2A.Z is only significant when the 5’ promoter is unoccupied by H2A.Z.

Figure 3.13: Combinatorial effects of 5’ and 3’ H2A.Z on sense transcripts. (A) Genome was divided into 4 groups according to their relative 5’/3’ H2A.Z occupancies: ones with high 5’ H2A.Z and zero 3’ H2A.Z (Class 1, 83 genes), ones with high H2A.Z at both 5’ and 3’ ends (Class 2, 236 genes), ones with H2A.Z at neither 5’ or 3’ ends (Class 3, 61 genes) and ones with high H2A.Z only at 3’ ends (Class 4, 33 genes). (B) Fold-changes of sense expression in htz1Δrrp6Δ compared to rrp6Δ. The up-regulation of Class 4 genes is significantly higher than Class 3 (p-value = 0.046), Class 2 (p-value = 4.9 × 10^{-6}) and Class 3 (p-value = 0.0010). Significance was computed by two-tailed t-test.
3.5 H2A.Z-related AS are coregulated with adjacent sense transcripts

The H2A.Z located at 3’ ends of protein-coding genes has a predominantly positive effect on AS transcripts. Since the budding yeast *S. cerevisiae* has a very compact genome, which means that 3’ H2A.Z peaks could be located in proximity to the regulatory element of the adjacent genes and may affect their transcription. In the next part of the study, we focused on genes with H2A.Z-marked 3’ promoters and examined their relationship the gene located downstream of their 3’ ends. We first examined the orientation of the downstream gene.

3.5.1 H2A.Z-related AS transcripts predominantly occur at tandem genes

The genomic context of 3’ promoters may fall into two configurations. If a 3’ promoter is adjacent to a neighbouring 5’ promoter, we denote this configuration as ”tandem” (because two genes are tandemly arranged) whereas if a 3’ promoter is adjacent to the neighbour’s 3’ promoter, we denote this as ”convergent”. Also taking the distance between two genes into account, we further divided the each configuration into three categories: overlapping, close (<300 bp apart) and far (>300 bp apart). The majority of gene pairs fall into the tandem-close category (1747 genes) and convergent-overlapping (2223 genes) and no tandem-overlapping pairs are present in the *S. cerevisiae* genome (Fig 3.14 A). We plotted the H2A.Z profiles upstream of TES and observed very high abundance of H2A.Z at the 3’ promoter in tandem-close gene pairs (Fig 3.14 B, red line). The level is significantly higher than any other configuration (Figure 3.14 C). The AS transcripts originated from tandem-close promoters are also higher than other configurations (Fig 3.14 D).
CHAPTER 3. H2A.Z ACTIVATES AS TRANSCRIPTS

3.5.2 AS and adjacent sense transcripts are coregulated by H2A.Z

Previous studies have reported promoter bidirectionality in almost all eukaryotic organisms. The observation that the H2A.Z at 3’ ends tends to be located near a 5’ promoter leads to the speculation whether the AS and sense transcripts are initiated from the same promoter. To test for this, we examined the correlation of the AS transcript with the sense transcript of the adjacent gene. We focused on the 781 tandem-close gene pairs that are less than 300 bp apart and with H2A.Z-occupied 3’ promoters. By examining the fold-changes of AS and sense expression in HTZ1 deletion, we observed that the AS:sense pairs tend to be simultaneously up-regulated (Fig 3.15 A) or down-regulated (Fig 3.15 B). The fold-changes of
AS and sense expression are positively correlated with Pearson correlation = 0.36 (Fig 3.16 A, B). Out of the 781 AS:sense pairs, we identified 292 AS and 117 sense transcripts that are decreased by >1.5 fold upon HTZ1 deletion (Fig 3.16 C) and their overlap is significantly overrepresented (Fig 3.16 C, D, blue). Consistent with this, the overlap between increased AS and increased sense is also overrepresented (Fig 3.16 C, D, red). These results indicate that the AS:sense pairs originated from H2A.Z-occupied promoters are coregulated, which could be the consequence of being produced from the same bidirectional promoter.

Figure 3.15: Examples of H2A.Z-related AS transcripts coregulated with adjacent sense transcript. (A) Genes whose AS (light blue) and sense (dark blue) transcripts are both up-regulated in htz1Δrrp6Δ compared to rrp6Δ strain. Coregulated sense and AS pairs are indicated by red arrows. (B) Genes whose AS (light blue) and sense (dark blue) transcripts are both down-regulated in htz1Δrrp6Δ compared to rrp6Δ strain.
Figure 3.16: Coregulation between H2A.Z-related AS and adjacent sense transcripts. (A) Hierarchically clustered data showing the positive correlation between fold-changes of AS and sense transcripts within the 781 expressed AS:sense pairs with H2A.Z-occupied promoters. (B) Positive correlation (Pearson correlation = 0.36) between AS and sense transcripts upon HTZ1 deletion. (C) Numbers and percentages of increased/decreased AS/sense transcripts within these 781 AS:sense pairs. Increased and decreased transcripts are defined as $>1.5$ fold change in expression and with t-test $p$-value $<0.1$ in $htz1\Delta rrp6\Delta$ compared to $rrp6\Delta$. (D) Concurrence of decreased AS and increased sense (11 pairs, underrepresented, $p$-value $= 1.2 \times 10^{-4}$), decreased AS and decreased sense (blue, 61 pairs, overrepresented, $p$-value $= 5.7 \times 10^{-4}$), increased AS and increased sense (red, 20 pairs, overrepresented, $p$-value $= 1.8 \times 10^{-7}$), increased AS and decreased sense (6 pairs, not significant).
3.6 H2A.Z peaks inside coding regions negatively correlate with AS transcripts

In our H2A.Z ChIP-seq data, we also observed \( \sim 11\% \) of the total H2A.Z signals located inside the coding regions (CDSs) of genes, despite that their length-normalised levels are not enriched as 5’ or 3’ peaks. We also would like to know whether H2A.Z peaks in CDSs are related to AS transcripts. In the entire genome, we identified 12027 H2A.Z peaks, of which 3012 are located inside the CDSs of genes which are defined as the ORF regions that are >300 bp away from both TSS and TES. We next checked whether these H2A.Z peaks are associated and correlated with AS transcripts.

3.6.1 H2A.Z peaks in CDS are associated with AS transcripts

By examining the locations of H2A.Z peaks and AS transcripts, we observed many examples of H2A.Z inside CDS marking the start site of an AS transcript (Fig 3.17 A). To test for the significance of this association, we first obtained 718 peaks (24% out of the 3012) that are associated with the start sites of AS transcripts (Fig 3.17 B). We randomly generated 3012 150-bp regions (approximately the median length of all CDS peaks) from all the coding regions of genes. We repeated the randomisation 100 times and observed that the frequency of AS association due to random chances is significantly lower than the actual number (Fig 3.17 C). H2A.Z peaks inside CDS also have significantly higher levels of histone marks for active promoters such as H3K4me3, H3K9ac and H3K18ac than CDS regions without H2A.Z (Fig 3.17 D-F). In addition, H2A.Z peaks that are associated with AS transcripts tend to have higher active-promoter marks than those are not associated with AS transcripts (Fig 3.17 D-F). These observations suggest that the H2A.Z peaks inside CDS are also associated with cryptic promoters and AS transcripts, which is similar to the H2A.Z at 3’ ends of genes.
Figure 3.17: H2A.Z peaks inside coding regions of genes are associated with AS transcripts. (A) Examples of H2A.Z peaks inside CDS that are marking the start sites of AS transcripts. (B) Number and percentage of H2A.Z peaks in CDS (>300 bp from ends of genes) that are associated with AS transcripts. Inside CDS, 3012 H2A.Z peaks were found. Peaks are regarded as AS-associated if the RNA level in a 50 bp window downstream of it is more than 3 fold higher than the upstream window on the AS strand. (C) Significance of the association between H2A.Z peaks in CDS and AS transcripts. A set of 3012 150-bp regions was generated from all CDSs and tested for AS association. Randomisation was repeated 100 times and the frequencies of AS-associations were plotted as histogram. Dashed line shows the observed 718 AS-associated peaks and the false discovery rate is approximately zero. (D) H3K4me3 levels (Guillemette et al., 2011) at AS-associated H2A.Z peaks inside CDS is higher than at CDS H2A.Z peaks that are not associated with AS (two-tailed t-test, p-value = $2.5 \times 10^{-31}$), which is higher than CDS regions with no H2A.Z (two-tailed t-test, p-value < $1.8 \times 10^{-308}$). (E) H3K9ac (Cai and Tu, 2011) showing similar pattern to D (two-tailed t-test, p-values = $3.0 \times 10^{-14}$ and $1.1 \times 10^{-70}$ respectively). (F) H3K18ac (Xue-Franzen et al., 2013) showing similar pattern to D (two-tailed t-test, p-values = $3.7 \times 10^{-11}$ and < $1.8 \times 10^{-308}$ respectively).
3.6.2 CDS H2A.Z has positive effects on AS and negative effects on sense transcripts

To test whether CDS peaks are also activators of AS transcripts, we next examined the effect of HTZ1 deletion on the 718 AS transcripts marked by CDS H2A.Z. Upon HTZ1 deletion, we observed 70 up-regulated and 101 down-regulated AS transcripts by >1.5 fold (Fig 3.18 A). AS transcripts whose promoters are occupied by high levels of H2A.Z tend to be down-regulated in HTZ1 deletion, suggesting that H2A.Z is required for the expression of AS transcripts (Fig 3.18 A, B, blue). The down-regulation of AS transcripts is linked with up-regulation of their cognate sense transcripts (Fig 3.18 C). Therefore, we conclude that the H2A.Z inside CDSs are likely marking cryptic promoters and also has positive effects on the AS and negative effects on the sense transcripts of protein-coding genes, which is very similar to the effects of 3’ H2A.Z.

Figure 3.18: Effects of CDS H2A.Z on AS and sense transcripts. (A) Fold-changes of AS transcripts upon HTZ1 deletion against the length-normalised signals of the 718 AS-associated peaks inside CDS. Up- and down-regulated AS transcripts are defined as >1.5 fold and with \( p\)-value < 0.1. (B) Down-regulated AS transcripts (blue) are marked by significantly higher H2A.Z peaks than those are not differentially expressed (two-tailed t-test \( p\)-value = 4.1 \times 10^{-8}). (C) Down-regulated AS transcripts are accompanied by more up-regulated sense transcripts compared to up-regulated AS (two-tailed t-test, \( p\)-value = 0.0061).
3.7 Discussion

3.7.1 ChIP-seq improves quantification of H2A.Z

A main motive of this study is to answer why previous studies reported both activating and repressive roles in \textit{S.cerevisiae}. To look for answers, we first used high-resolution technique (ChIP-seq) and accurate transcript annotations to improve the quality of genome-wide H2A.Z profiling. Our results agree with previous data in the sense that H2A.Z is primarily located at promoters of protein-coding genes (Guillemette et al., 2005; Raisner et al., 2005; Li et al., 2005). We estimated that 60\% of H2A.Z is located at the promoters of genes which is consistent with the \textasciitilde 74\% estimated by Guillemette et al. (2005). We also observed a significant but not enriched proportion inside coding regions of genes, which is also visible from the results of previous studies (Fig.2 Raisner et al., 2005; Fig.1 B Guillemette et al., 2005). Not many previous studies have addressed these H2A.Z peaks outside gene promoters. Interestingly, we observed an enrichment of H2A.Z at the 3’ ends of protein-coding genes that consists of \textasciitilde 6.7\% of the total abundance. This 3’ signal may have not been detected or distinguished from 5’ peaks using low-resolution ChIP-chip or ORF annotations. In general, our data give an improved resolution and more complete view of H2A.Z distribution at various genomic features in relation to genes.

3.7.2 H2A.Z at promoter is required for both sense and AS transcripts

By comparing H2A.Z localisation with transcripts, we established the co-localisation between the H2A.Z at 3’ ends of genes and the start sites of AS transcripts. The H2A.Z peaks outside 5’ promoters tend to co-localise with the start sites of AS transcripts. These regions are enriched for all histone modifications that have been known to mark active promoters (Liu et al., 2005), indicating the likelihood of being active AS promoters. Our findings support a predominantly activating role of H2A.Z on transcripts since \textit{HTZ1} deletion causes down-regulation of transcripts (both sense and AS) with H2A.Z-occupied promoters. Up-regulated transcripts in \textit{htz1}$\Delta$ tend to have promoters unoccupied by H2A.Z, indicating that the up-regulation is likely due to indirect effects of \textit{HTZ1} deletion. Notably, the impact of \textit{HTZ1} deletion is especially prominent
for AS transcripts, implying that the regulatory role of H2A.Z on AS transcripts may be more potent than on sense transcripts. This is consistent with a recent study that shows the accumulation of H2A.Z at 3’ ends of genes is accompanied by increased AS transcripts in a loss-of-function mutant of INO80 (Alcid and Tsukiyama, 2014). In mouse embryonic stem cells, it was recently shown that H2A.Z is required for the expression of AS transcripts generated at bidirectional 5’ promoters (Rege et al., 2015). However, HTZ1 deletion does not diminish the AS transcripts but its levels, which implies that H2A.Z is not the only factor that regulates AS transcripts. Indeed, studies have reported other chromatin components that are involved in regulating AS transcript levels, including the ATP-dependent remodeler Isw2 (Whitehouse et al., 2007), the histone methyltransferase Set1 (Margaritis et al., 2012), acetyltransferase Rtt109 (Rege et al., 2015), the histone deacetylase complexes Rpd3S (Churchman and Weissman, 2011) and Set3C (Kim et al., 2012). Therefore, it is likely that H2A.Z is cooperating with other chromatin components in regulating steady-state expression of transcripts.

3.7.3 Overall effect of H2A.Z on sense transcripts

Most genome-wide studies on the H2A.Z-transcript relationship in *S.cerevisiae* were conducted before the scale of non-coding RNA was fully realised. Our results indicate that the H2A.Z at 3’ ends and inside CDSs of genes are marking and probably activating AS transcripts. The anti-correlation between AS and sense transcripts suggests that the 3’- and CDS-H2A.Z are likely negative regulators of the gene’s sense transcript. This partially explains why the 5’ occupancy of H2A.Z alone does not correlated with the sense transcript levels. We emphasise the importance of dissecting H2A.Z at various genic locations when analysing its effect on gene’s sense transcripts. The overall effect depends on the relative levels of 5, CDS and 3’ occupancies. The multi-functionality of H2A.Z on transcripts has also been suggested by Rege et al. (2015) who demonstrated that H2A.Z, in addition to promote CUTs and mouse promoter-proximal ncRNAs, also represses the Ssu72 repressed transcripts or SRTs in *S.cerevisiae*. 
3.7.4 H2A.Z is likely marking bidirectional promoters

H2A.Z-marked AS promoters are preferably produced from tandemly arranged gene pairs, which resemble the bidirectional promoters reported by Neil et al. (2009) and Xu et al. (2009). H2A.Z is enriched at tandem gene pairs that are <300 bp apart but far less enriched at tandem pairs that are >300 bp apart, which supports the case of a single bidirectional promoter rather than two separate promoters. We also observed a stronger association between H2A.Z with CUTs than with SUTs, which is similar to the previous finding that most CUTs are initiated from bidirectional promoters (Neil et al., 2009). In addition, the AS and its adjacent sense transcripts are coregulated in HTZ1 deletion, which also suggests that the H2A.Z at 3' ends of genes is likely part of bidirectional promoters between two tandemly arranged genes.

Recent studies reported that bidirectional promoters consist of two pre-initiation complexes (PICs) independently assembled in opposite directions (Murray et al., 2012; Rhee and Pugh, 2012). It is still unclear whether the two PICs are differentially regulated or have differential functions in the regulation of transcripts. In our results, we observed coregulation between the AS:sense pairs produced from bidirectional promoters, which indicates that the two PICs may be coregulated. On the other hand, we also observed the difference that the H2A.Z at -1 nucleosome strongly correlates with AS transcripts, whereas the H2A.Z at +1 nucleosome only weakly correlates with sense transcripts. This indicates an asymmetrical role of H2A.Z at bidirectional promoters, which could indicate differential regulation of PIC assembly in sense and AS directions. It would be interesting to study the extent of differential regulation at bidirectional promoters in the future.

3.7.5 H2A.Z’s correlation with transcript levels

In spite of the predominantly activating function of H2A.Z on transcripts, our data are not contradicting the previously reported negative correlation between promoter H2A.Z and expression. We observed that H2A.Z is depleted from promoters of both lowly and highly expressed genes (Fig S3.1 A). Also, genes with unoccupied promoter by H2A.Z (enriched for ribosomal-component genes) tend to be highly expressed (Fig S3.1 B), which is consistent with the negative correlation observed by Guillemette et al. (2005); Li et al. (2005); Zhang et al. (2005) and
Zanton and Pugh (2006). Therefore, our results suggest that promoter H2A.Z occupancy is required for the normal level of sense transcripts but does not determine the level of transcript expression in steady state. However, it is interesting that the 3’ H2A.Z shows a positive correlation with AS transcript levels and HTZ1 deletion affects AS more than sense transcripts (Fig 3.10 D, Fig 3.11 B).

3.7.6 Subtelomeric H2A.Z does not mark AS transcripts

H2A.Z in subtelomeric regions has been known to act as a positive regulator of genes by limiting the spread of heterochromatin (Meneghini et al., 2003). We observed relatively high H2A.Z occupancy in the CDSs of subtelomeric genes (Fig S3.2 B). Given that many H2A.Z peaks inside CDSs are required for the expression of AS transcripts which negatively affects sense transcripts (Fig 3.18 C, S3.2 A), we wondered whether CDS H2A.Z in subtelomeric regions are also repressing sense transcripts. In fact, our data showed that AS transcripts are depleted from subtelomeric regions, indicating the AS-mediated repression is nonexistent for subtelomeric genes (Fig S3.2 C). The accumulated H2A.Z in the CDS of subtelomeric regions is likely due to the lack of transcription (Fig S3.2; Hardy et al., 2009). Therefore, the regulation of H2A.Z on subtelomeric genes does not involve AS transcripts.

3.7.7 Repressive effect of H2A.Z on AS transcripts in S. pombe

In contrast to our findings, a previous study conducted in the fission yeast S. pombe reported H2A.Z as a repressor of AS transcripts (Zofall et al., 2009). Zofall et al. (2009) focused on the AS transcripts that are produced by transcriptional read-through between convergent gene pairs and demonstrated that H2A.Z is suppressing these AS transcripts in a synergistic manner with RNAi factors such as Clr4 and Ago1. However, this may be a different scenario to S. cerevisiae, as the majority of AS transcripts in S. pombe are generated by transcriptional read-through at convergent genes (Rhind et al., 2011) whereas most AS transcripts in S. cerevisiae are initiated from bidirectional promoters between tandem genes. Due to the limitation of our experimental technique, we could not distinguish between transcriptional read-through and the AS transcripts initiated from 3’ promoters between convergent gene pairs. For future studies, this can be improved
by techniques such as NET-seq. However, it is unlikely that the mechanism is conserved since most of the RNAi components are missing in *S. cerevisiae*.

### 3.7.8 Steady state versus stress response

It is essential to distinguish between the roles of H2A.Z in steady states and during stress response. In this chapter, we have established that promoter H2A.Z is likely to be a transcriptional activator in steady states, where the NDR at active promoters have been established and transcription machinery is constantly assembled and released into elongation. Mechanistically, H2A.Z could contribute towards destabilising the +1 nucleosome barrier to achieve its activating role (Weber et al., 2014). However, this does not necessarily mean that H2A.Z is also an activator during transcriptional disturbance. Studies have reported negative correlations between the changes in promoter H2A.Z and gene expression during stress responses (Zhang et al., 2005; Zanton and Pugh, 2006), which could be due to nucleosome loss at induced genes (Shivaswamy and Iyer, 2008). Since induced and repressed genes undergo vigorous changes in promoter organisation, the role of H2A.Z during transcriptional response could be different from steady states, which will be discussed in Chapter 5.

![Figure 3.19: Summary of the effect of H2A.Z on transcripts.](image)

The H2A.Z located at 3' ends of tandemly transcribed genes is required for the expression of AS transcripts. AS transcripts have negative effects on their cognate sense transcripts. The AS and adjacent sense transcripts are likely originated from a shared promoter that is marked by H2A.Z. The AS:sense pairs are coregulated by H2A.Z.
3.8 Summary

Based on our observations, we conclude that H2A.Z is a general marker of promoters and is required for the normal expression of transcripts in steady states. The overall effect of H2A.Z on sense transcripts is a combination of activation and repression depending on the relative levels of H2A.Z at the 5’, CDS and 3’ ends of genes and the involvement of AS transcripts (Fig 3.19). The H2A.Z at 5’ end of genes acts as a direct positive regulator of sense transcripts. The H2A.Z at 3’ end or inside CDS positively regulates AS transcripts and in turn represses the sense transcript of the gene. The repression through AS transcripts explains some of its apparently conflicting effects on the sense transcripts of genes. In addition, AS transcripts tend to share promoter with the adjacent gene and are coregulated with the adjacent gene. Therefore, our findings highlight the importance to study all transcripts derived from a locus in order to fully understand how genes are regulated in eukaryotic genomes.
3.9 Supplementary Figure

Figure S3.1: Correlation between 5’ H2A.Z and expression. (A) Gene were divided into 10 quantiles based on expression from low to high. The 5’ H2A.Z occupancy is plotted for each quantile. Lowly expressed genes are enriched for TATA genes and highly expressed gene are enriched for ribosomal components. (B) Expression for quantiles of 5’ H2A.Z occupancy.

Figure S3.2: CDS H2A.Z in subtelomeric regions. (A) CDS H2A.Z is negatively correlated with sense transcripts. Genes are divided into 10 quantiles based on CDS H2A.Z occupancy and the sense expression for each quantile is plotted. (B) Subtelomeric genes have significantly higher CDS H2A.Z than non-subtelomeric genes (two-tailed t-test *p-value* = 1.5 × 10\(^{-6}\)). Subtelomeric genes are defined as genes that are within 20 kb from either end of chromosome. (C) Subtelomeric genes have significantly lower AS transcripts than non-subtelomeric genes (two-tailed t-test *p-value* = 0.031). (D) Subtelomeric genes have significantly lower sense transcript levels than non-subtelomeric genes (two-tailed t-test *p-value* = 3.4 × 10\(^{-7}\)).
Chapter 4

Non-coding RNAs Between Tandem Genes Are Induced During Stress Response

4.1 Introduction

In the previous chapter, we established the link between H2A.Z and the transcription of long non-coding RNAs (lncRNAs) under non-stress conditions. In *S. cerevisiae*, the gene HTZ1 becomes essential when yeast is exposed to environmental stresses. The mechanism is linked with the fact that some stress-related genes fail to be activated in the *htz1*Δ strain (Santisteban et al., 2000; Adam et al., 2001; Larochelle and Gaudreau, 2003; Farris et al., 2005). In recent years, it has become clear that lncRNAs make up a significant proportion of the *S. cerevisiae* transcriptome and could be important in regulating the transcription of protein-coding genes. Therefore, it would be interesting to investigate how the non-coding transcriptome changes during stress response. Insights into lncRNAs could also shed light on why the *htz1*Δ mutant has severely impaired growth under certain stressed conditions.

4.1.1 Transcriptomic changes in stress response

The budding yeast *S. cerevisiae* is able to respond to various stresses such as changes in nutrients, temperature, osmolarity and the presence of toxins or radiations. The environmental stress responses (ESRs) to different types of stresses
tend to share a common set of induced and repressed genes (Gasch et al., 2000; Causton et al., 2001). Repressed genes include ones that are required for ribosome structures, RNA polymerase-dependent transcription, and protein translation, indicating a need to pause protein synthesis in order to conserve energy while the cell is adapting to the new environment (Ashburner et al., 2000; Ball et al., 2000). Induced genes are involved in carbohydrate metabolism, protein folding and degradation, oxidative stress defense, autophagy, cytoskeletal reorganization and DNA-damage repair. This is thought to protect internal energy reserves, osmolarity balance and integrity of cellular structures (Schuller et al., 1994; Lewis et al., 1995). Genome-wide studies have also reported differential expression of lncRNAs during stress response. Xu et al. (2009) and Xu et al. (2011) observed that in response to the switch of growth medium, promoter-proximal lncRNAs tend to be coregulated with the sense transcript of gene whereas intragenic lncRNAs tend to anti-correlate gene expression. The anti-correlation was also observed when yeast undergoes heat shock (Yassour et al., 2010). There has been a debate on whether lncRNAs are functional in regulating gene expression or are merely transcriptional noises. In recent years, more and more functions of individual lncRNAs have been characterised, which has changed the view of how eukaryotic transcriptomes are regulated.

### 4.1.2 Upstream cis-lncRNAs as repressor

lncRNA can be produced from different locations in relation to protein-coding genes. The relative location is closely related to the mechanism of how it regulates the target gene. A major class of well-characterised lncRNAs are located upstream and in cis to genes. Upstream lncRNAs generally have repressive roles on the expression of their downstream target. Two different and possibly interrelated mechanisms have been characterised. The first mechanism is promoter occlusion, in which the transcription of lncRNA extends into the promoter of the downstream gene, physically preventing the access of general TFs and chromatin remodelers and resulting in a closed chromatin organisation at the gene promoter (Fig 4.1 A). Well characterised examples of such lncRNA:gene pairs include \textit{SRG1:SER3} (Martens et al., 2004), \textit{ZRR1:ADH1} (Bird et al., 2006) and \textit{ICR1:FLO11} (Bumgarner et al., 2009).
Figure 4.1: Locations of lncRNAs relative to genes and mechanisms of regulatory functions. (A) Upstream cis-lncRNAs tend to have repressive roles on the gene downstream. Repression is achieved by preventing the access of transcription factors to the promoter of gene (i.e. promoter occlusion) or by creating H3K36me3 and histone deacetylation around promoter regions with the involvement of Set2 and Rpd3S. (B) lncRNAs spanning the ORF of genes tend to repress the gene by blocking the transcription elongation with uncharacterised mechanism or by causing deacetylation at the gene promoter by the histone deacetylases Hda1/2/3. (C) The lncRNA produced from the 3’ end of CDC28 positively correlates with the expression of CDC28. Transcription of lncRNA is triggered by the binding of Hog1 at the 3’ end of CDC28. During osmostress, Hog1 is relocated to the 5’ end by a gene loop and induces the expression of CDC28.

Upstream lncRNAs can also repress the downstream target by lncRNA-mediated chromatin remodelling. The transcription or transcript of lncRNA can recruit chromatin factors that alter the local chromatin organisation to achieve repression of the target gene. In the pair of IRT1:IME1 for example, as the lncRNA IRT1 is being transcribed, RNA Pol II brings Set2 into contact with chromatin and causes H3K36me3 at the promoter of the downstream gene (Fig
4.1 A). H3K36me3 also leads to deacetylation on many lysine sites by the Rpd3S histone deacetylase. Deacetylated chromatin suppresses histone exchange and contributes towards increased nucleosome density at promoter that silences the IME1 transcription (van Werven et al., 2012; Kim et al., 2012). The same mechanism was observed for the lncRNA of GAL10 produced from the 3’ end of GAL10 in repressing the downstream GAL1 gene (Houseley et al., 2008; Pinskaya et al., 2009).

4.1.3 lncRNA in *trans* as repressor

Many lncRNAs are produced from the 3’ ends of protein-coding genes in *trans* and usually span the entire ORF of the gene (Fig 4.1 B). These lncRNAs tend to repress their cognate by elongation block or chromatin remodelling. Elongation block is the proposed mechanism that the transcription of lncRNA interferes with the elongation of sense transcription without affecting TF bindings at its promoter. Examples of lncRNA-gene pairs include RME2:IME4 (Hongay et al., 2006) and RME3:ZIP2 (Gelfand et al., 2011). In the RME2:IME4 pair of transcripts, deleting 224-675 bp from the IME4 ORF disrupts the RME2-mediated repression, indicating that the site for repression is within the ORF of gene rather than at the promoter (Gelfand et al., 2011). However, the exact mechanism of how lncRNA transcripts or transcription block the elongation of the cognate gene is still unclear. A different mechanism that involves chromatin remodelling at the promoter of the target gene was observed for the antisense transcript of PHO84. Stabilisation of the PHO84 antisense transcript induces H3/H4 deacetylation by recruiting the histone deacetylases Hda1/2/3 to the PHO84 promoter (Camblong et al., 2007). The deacetylated promoter disfavours nucleosome exchange and leaves the gene in a repressed state (Fig. 4.1 B).

4.1.4 lncRNA as gene activator

In a more recent study, positive correlations between lncRNAs and sense transcripts were observed in *S.cerevisiae* in response to osmostress (Nadal-Ribelles et al., 2014). Upon osmostress, the stress-induced protein kinase Hog1 binds to the 3’ ends of many protein-coding genes and mediates the expression of their antisense lncRNAs (Fig 4.1 C). One of the Hog1 target is CDC28 whose 3’ UTR and 5’ promoter forms a gene loop while the lncRNA is produced, which causes
relocation of Hog1 to the 5’ promoter. Hog1 at the 5’ promoter, along with the lncRNA transcript, leads to nucleosome eviction by the RSC remodelling complex and induces the sense expression of \( CDC28 \). In addition to \( CDC28 \), Hog1 also induces the AS lncRNA of other genes. These lncRNAs are strictly transcribed from the 3’ end and usually span the entire ORF of their cognate gene. Another similar example observed by Nadal-Ribelles et al. (2014) is the co-activation of \( MMF1 \) and its lncRNA.

### 4.1.5 Objective

While studies on individual transcripts have reported functional lncRNAs, recent genome-wide studies revealed over 1000 lncRNA transcripts in any condition in \( S.\text{cerevisiae} \), the majority of which are of unknown function (Xu et al., 2009; Yassour et al., 2010). During stress responses, the levels of lncRNAs also change in both positive and negative correlation with nearby genes, indicating their interleaved regulatory patterns with protein-coding genes (Yassour et al., 2010; Xu et al., 2011). In this chapter, we used caffeine as the stress condition and investigated how the lncRNA transcriptome responses to caffeine. We used strand-specific RNA-seq to profile the stable lncRNA transcripts in over a 120-minute time course of caffeine response. First, we would like to ask whether lncRNAs are likely to be induced or repressed during stress response? Given that lncRNAs at different locations relative to genes tend to carry out different functions, we would also like to know the locations of lncRNAs that are differentially regulated. Moreover, it would be important to study the correlation between lncRNA and the sense transcripts of genes for different types of lncRNAs. Insights into the non-coding transcriptome during transcriptional responses will contribute towards understanding the interleaved nature of gene regulatory system in eukaryotic genomes.
4.2 Caffeine Causes Significant Transcriptomic Changes

Caffeine (1, 3, 7-trimethylxanthine) is an analogue of purine bases and has pleiotropic effects on eukaryotic organisms. Pathways that were previously known to be affected by caffeine include the activation of cAMP-dependent protein kinase pathway (Tsuzuki and Newburgh, 1975), inhibition of phosphoinositide-3-kinase-related protein kinases (Saiardi et al., 2005), affecting DNA-repair (Kaufmann et al., 2003; Cortez, 2003) and inhibiting TOR signalling pathway (Kuranda et al., 2006; Reinke et al., 2006). In this study, we treated yeast cells with 10mM of caffeine for 2 hours and measured its transcriptome before and after treatment using strand-specific RNA-seq. As a quality control, we first examined the sense transcriptome and checked whether our data agree with previously published datasets obtained from microarrays. Despite the differences in technique, we would also like to know how similar caffeine response is compared to other environmental stress responses (ESRs).

4.2.1 Induced and repressed genes

To quantify gene expression, we first downloaded the annotation of the 5143 protein-coding transcripts (Xu et al., 2009) and calculated the length-normalised RNA-seq signal for each gene. We identified 1144 activated and 778 repressed genes whose sense transcript levels are up- or down-regulated by more than 2 fold in the first 30 minutes of caffeine response (Fig 4.2 A). Activated genes have lower-than-average expression at 0 minute (Fig 4.2 B and C, red) and are enriched for GO terms such as oxidoreductases, pyridoxal phosphate binding proteins, transaminases amino peptidases and transmembrane transporters, which are enzymes required to tackle the environmental stress (Fig 4.2 D). Notably, an overrepresented set of 335 activated genes are TATA genes, which were known to be related to stress response (Basehoar et al., 2004). Repressed genes are usually highly expressed at 0 minute (Fig 4.2 B and C, blue) and are enriched for transcriptional and translational machineries, indicating a temporary shut-down of protein synthesis (Fig 4.2 E).
Figure 4.2: Activated and repressed genes at 30 minutes of caffeine response.

(A) 1144 genes are activated and 778 genes are deactivated out of the 5143 genes in total. An overrepresented set of 335 activated genes contain TATA promoters. 

(B) Fold-changes of expressions versus the ground level of expression at 0 minute. Activated and deactivated genes are coloured in red and blue respectively.

(C) Activated genes have significantly lower expression (two-tailed t-test $p$-value $= 7.1 \times 10^{-7}$) and deactivated genes have significantly higher expression (two-tailed t-test $p$-value $= 2.3 \times 10^{-9}$) than genes whose expression levels are not changed.

(D) Enriched Gene Ontology (GO) categories in activated genes. GO categories with Bonferroni-corrected $p$-values $< 1.0 \times 10^{-5}$ are shown. Upper panel shows the significance indicated by $-\log_{10}(p$-value$)$ and lower panel shows the number of genes in the category.

(E) Enriched GO categories in deactivated genes with Bonferroni-corrected $p$-values $< 1.0 \times 10^{-5}$. 
4.2.2 Caffeine response is similar to general ESRs

To establish whether our data agree with previously published data of general ESRs, we performed hierarchical clustering using Euclidean distances of the data matrix containing log fold-changes of our caffeine response data combined with Gasch et al. (2000)'s data of heat shock at 20 minutes, 1M sorbitol at 30 minutes, 0.32mM H₂O₂ at 30 minutes, 1.5mM diamide at 30 minutes, 1mM menadione at 30 minutes and amino acid starvation for 1 hour. Clustering results suggest that caffeine response is highly similar to all these ESRs and is closest to heat shock (Fig 4.3 A). This is confirmed by high Pearson correlations between fold-changes of gene expression in caffeine response and in general ESRs (Fig 4.3 B). Gasch et al. (2000) also identified the core sets of induced and repressed genes that are common to all the ESRs. We found that nearly all the core stress-response genes are within the genes activated or deactivated by caffeine (Fig 4.3 C). Notably, we observed a much larger set of genes that are induced or repressed by caffeine than general stresses, indicating that the effects of caffeine on cell is more severe than general environmental changes.

4.2.3 Gene expression recovers after 30 minutes

To study how the transcriptome behaves after the initial response, we next examined the gene expression at 120 minutes in relation to 30 minutes. We found that 430 out of the 1144 activated genes at 30 minutes are down-regulated by 120 minutes and 220 out of the 778 deactivated genes are up-regulated (Fig 4.4 C). Globally, the fold-changes of gene expression between 0 and 30 minutes are negatively correlated with the changes from 30 to 120 minutes (Fig 4.4 A and B). The number of genes that recovered from induction and repression are both significantly overrepresented (Fig 4.4 D). All these results suggest that the activated and deactivated genes at 30 minutes are recovering by 120 minutes. However, the expression of activated genes are recovered by \(\sim 72\%\) and the deactivated genes are recovered by \(\sim 65\%\) (Fig 4.4 E).

So far, we have demonstrated that the sense transcriptomic response of our caffeine data shows good agreement with previous published ESR data. In both cases, protein synthesis and RNA metabolism are paused and stress-related enzymes such as oxidoreductases, protein kinases and TM transporters are induced.
This indicates the good quality of our data and suitability for analysis of non-coding RNAs.

Figure 4.3: Caffeine response is similar to other environmental stress responses. (A) Hierarchically clustered data showing the similarities between caffeine response and other ESRs. The data matrix consists of the log₂ fold-changes of gene expressions between 0 and 30 minutes in our caffeine response RNA-seq data and other ESR microarray data from Gasch et al. (2000). Datasets include heat shock 20 minutes, 1M sorbitol 30 minutes, 0.32 mM $H_2O_2$ 30 minutes, 1.5mM diamide 30 minutes, 1mM menadione 30 minutes and amino acid starvation 1 hour. The 1922 genes that are differentially expressed at 30 minutes of caffeine response are shown. (B) Pearson correlations between the expression change of the 1922 differentially regulated genes during the first 30 minutes of caffeine response and various ESRs. (C) Nearly the entire core set of genes that are induced in all ESRs (identified by Gasch et al., 2000) are up-regulated at 30 minutes of caffeine response (Fisher’s Exact test, $p$-value $= 6.5 \times 10^{-48}$). Similar for the core set of repressed genes (Fisher’s Exact test, $p$-value $= 2.3 \times 10^{-78}$).
CHAPTER 4. NON-CODING RNA DURING STRESS

Figure 4.4: Recovery of expression from 30 to 120 minutes in caffeine response.
(A) Hierarchical clustering data on fold-changes of gene expressions from 0 to 30 minutes and from 30 to 120 minutes. Only the 2007 differentially regulated genes at 30 minutes are included.
(B) Negative correlation between the fold-changes of expressions from 0 to 30 minutes and from 30 to 120 minutes. Genes recovered from activation (significantly up-regulated from 0 to 30 minutes and down-regulated from 30 to 120 minutes) and from deactivation (significantly down-regulated from 0 to 30 minutes and up-regulated from 30 to 120 minutes) are coloured in red and blue respectively. Number of genes in each quadrant are shown on the plot. (C) Summary of number of genes that are recovered from activation or deactivation. (D) Activated and deactivated genes at 30 minutes are likely to recover by 120 minutes (Fisher’s Exact test p-values = 6.1 × 10^{-219} and 1.4 × 10^{-118} respectively) (E) The degree of expression recovery by 120 minutes.
4.3 Long Non-coding RNAs Are Significantly Induced by Caffeine

IncRNAs in the *S. cerevisiae* transcriptome can stably exist in wild-type cells or can be actively degraded by exosomes. Due to the nature of our experimental procedure, we were only able to detect the stable IncRNAs or SUTs in our caffeine datasets. Therefore, in the rest of this chapter, the term "IncRNA" only refers to SUTs (i.e. excluding CUTs, XUTs and etc). If a IncRNA is produced from the antisense (AS) strand in the vicinity of a protein-coding gene, it is also denoted as the AS transcript of the gene. However, IncRNAs produced from the sense strand of protein-coding genes cannot be distinguished from the sense transcripts in strand-specific RNA-seq and are therefore not discussed in this study.

4.3.1 IncRNA levels are increased in during caffeine response

To quantify the levels of IncRNA transcripts, we first performed *de novo* transcriptome assembly from our strand-specific RNA-seq data. We took the union set of transcripts that are expressed in at least one of the 0, 30 or 120-minute data, filtered out protein-coding transcripts (annotated by Xu et al., 2009) and obtained 1703 IncRNA transcripts. We then compared these 1703 IncRNAs to the SGD and Ensembl gene assemblies. Out of the 1703 transcripts, 148 have been annotated by SGD (also by Ensembl) as putative protein-coding genes or dubious ORFs and 31 have been annotated only by Ensembl as predicted ORFs or snoRNAs (Fig 4.5 A). Next, we focused on the 1524 unannotated IncRNA transcripts and observed that their levels are significantly increased at 30 minutes of caffeine response (Fig 4.5 B). Out of the 1524 IncRNA transcripts, the majority (886) are up-regulated in the first 30 minutes of caffeine response whereas only 60 are down-regulated (Fig 4.5 C). By 120 minutes, both up- and down-regulated IncRNAs are recovering towards their ground levels (Fig 4.5 B, box plot; 4.5 D, E) which is similar to the pattern of sense expression during caffeine response.
Figure 4.5: ncRNAs are increased during caffeine response. (A) 1703 ncRNA transcripts were assembled from the union sets of the strand-specific RNA-seq data at 0, 30 and 120 minutes. The transcripts were compared with SGD genes to filter out known dubious ORFs. The rest were further compared with Ensembl genes to filter out snoRNAs and predicted transcripts. (B) Clustering data of ncRNA transcripts at 0, 30 and 120 minutes. Only the 1524 unannotated ncRNAs are shown. Box plot shows that ncRNAs are significantly increased from 0 to 30 minutes (two-tailed t-test, \( p\text{-value} = 8.1 \times 10^{-80} \)) and significantly decreased from 30 to 120 minutes (two-tailed t-test, \( p\text{-value} = 2.0 \times 10^{-10} \)). (C) Numbers of up- and down-regulated ncRNA transcripts at 30 minutes compared to 0 minute. (D) Fold-changes of ncRNA expression from 0 to 30 minutes and from 30 to 120 minutes, showing recovery of ncRNA levels from 30 to 120 after caffeine treatment. (E) Numbers of ncRNA transcripts that are changed from 0 to 30 and from 30 to 120 minutes of caffeine response. The 552 and 28 transcripts that are recovered from up- and down-regulation at 30 minutes are significantly overrepresented (Fishers Exact test, \( p\text{-value} = 1.1 \times 10^{-122} \) and \( 5.0 \times 10^{-15} \) respectively).
4.3.2 Induced lncRNAs are related to stress-response genes

An important question here is why lncRNAs are more likely to be activated than repressed? Also, do they have any functional significance? To investigate potential functions of these unannotated lncRNAs, we first examined their location in relation to protein-coding transcripts. We define a lncRNA transcript as genic if it overlaps with the ORF of any protein-coding gene extended by 300 bp. Transcripts that are >300 bp away from protein-coding genes are regarded as intergenic. Under these criteria, we identified 982 lncRNAs that are in the vicinity of protein-coding genes on the AS strand and 542 lncRNAs that are intergenic (Fig 4.6 A). The 982 lncRNAs are linked with 912 genes since more than one lncRNA transcripts are found in some genes. Here, we denote these 912 lncRNAs as AS transcripts to genes. Interestingly, genes with induced AS transcripts during caffeine response tend to have lower expression at 0 minute (Fig 4.6 C), indicating being suppressed in non-stress conditions. The 518 genes with increased AS transcripts are highly enriched in stress-response genes including transporters, phosphatases and oxidoreductases (Fig 4.6 D).
Figure 4.6: ncRNAs are related to stress-response genes. (A) Numbers of unannotated ncRNAs in the vicinity of genes (overlapping with extended ORFs) and in intergenic regions (>300 bp away from ORFs). (B) Numbers of genes associated with up- and down-regulated ncRNAs. (C) Genes with up-regulated AS transcripts are more repressed at 0 minute (two-tailed t-test \( p\)-value = 0.00042). (D) Enriched Gene Ontology categories (Bonferroni-corrected \( p\)-value < 0.05) for genes with up-regulated AS transcripts, sorted by significance of overrepresentation (indicated by \(-\log_{10}(p\text{-value})\)).
4.3.3  Induced IncRNA is accompanied by gene activation

Having established that induced AS transcripts are likely produced from the vicinity stress-response genes, we next studied the relationship between their expression levels during caffeine response. Out of the 518 genes with up-regulated AS transcripts in 30 minutes of caffeine response, 172 (33%) also have up-regulated sense transcript levels, which is significantly overrepresented (Fig 4.7 A). The 33% is significantly higher than the 20% amongst genes with no associated AS transcripts (Fig 4.7 C, red bars left vs right). Similarly, the overlap between up-regulated AS and down-regulated sense transcripts is significantly underrepresented (Fig 4.7 B). We also observed a weakly positive correlation between the fold changes of sense and AS expression between 0 and 30 minutes, with Pearson correlation = 0.11 (Fig 4.7 D). It is notable that genes with down-regulated AS transcripts also tend to be up-regulated (Fig 4.7 C, middle bars), indicating that the positive correlation between sense and AS is mainly contributed by co-upregulation rather than co-downregulation.
Figure 4.7: Increased lncRNAs are accompanied by increased sense transcripts. (A) Overrepresented overlaps between up-regulated AS and sense transcripts (Fisher’s Exact test, \( p\)-value = 1.3 \( \times \) 10^{-9}). (B) Overlap between up-regulated AS and down-regulated sense transcripts are less than expected (Fisher’s Exact test, \( p\)-value = 2.4 \( \times \) 10^{-11}). (C) Genes with and without increased AS transcripts show different degree of up-regulation. The 518 genes with up-regulated AS, 29 genes with down-regulated AS and are compared to the 4269 genes with no AS transcripts at any time points during caffeine response. The percentages of the cognate gene’s sense transcripts that are up-, down-regulated or not differentially expressed are shown in bar chat. (D) Fold-changes of AS and sense expressions from 0 to 30 minutes of caffeine response. Pearson correlation = 0.11.
4.4 LncRNA Links Co-activated Tandemly Arranged Gene Pairs

Next, we would like to investigate why AS transcripts tend to be co-activated with stress-response genes? To answer this question, it is essential to understand the relative location of the AS transcripts to the protein-coding genes since the AS transcripts produced from the 3’ ends of genes are likely to have different regulatory roles on the sense transcript to the AS originated from the TSS of genes. Therefore, we first classified the AS transcripts according to the location of their start sites in relation to protein-coding genes and then examined how each class correlates with sense transcripts.

4.4.1 Classification of lncRNA by relative location to gene

We used 300-bp windows up- and down-stream of the TSS of genes to classify the start sites of AS transcripts. Out of the 1542 lncRNA transcripts, we identified 221 AS transcripts that start within 300 bp upstream of TSS and are transcribed in the AS direction. We also found 200 AS transcripts whose start sites are downstream of the TSS of genes (Fig 4.8 A). We also observed 561 AS transcripts originated within the CDSs of genes that are >300 bp away from TSSs or TESs. We also identified 542 intergenic lncRNAs that are >300 bp away from protein-coding genes (Fig 4.8 A). Despite that most transcripts are located in the CDSs and IGRs, length-normalised number of transcripts showed highest density of lncRNAs initiated from around the TSS of genes compared to CDSs or IGRs (Fig 4.8 B, dark and light red). Genes with AS lncRNA transcripts at different locations tend to have enrichment in different GO categories (Fig 4.8 C, D, E).
Figure 4.8: Classification of AS transcripts by relative location to genes. (A) Classification of unannotated lncRNAs. Assembled lncRNA transcripts are classified into: 5’-upstream (whose 5’ end is within 300 bp upstream of a protein-coding gene), 5’-downstream (whose 5’ end is within 300 bp downstream of a protein-coding gene), CDS (whose 5’ end lies within the ORF >300 bp from TSS or TES) and intergenic (whose 5’ end is >300 bp from genes). (B) Density of each class of lncRNA, which is the number of transcripts normalised to the length of the genomic element. (C) Enriched GO categories (Bonferroni-corrected p-values < 0.05) for the 221 genes with AS transcripts starting upstream of their TSSs, sorted by significance levels indicated by -log_{10}(p-value). (D) Enriched GO categories for the 200 genes with AS transcripts starting downstream of their TSSs. (E) Enriched GO categories for the 561 genes with AS transcripts within their CDSs.
4.4.2 Promoter-associated AS lncRNAs are coregulated with genes

We then examined the correlation of AS transcripts with their cognate sense transcripts for each class. The 221 AS transcripts starting within 300 bp upstream of the TSSs of protein-coding genes show significant positive coregulation with their cognate sense transcripts (Pearson $R = 0.42$, Fig 4.9 A, B). Out of these 221 AS transcripts, 146 are significantly up-regulated at 30 minutes of caffeine response, of which 60 are accompanied by up-regulated sense transcripts which is significantly overrepresented (Fig 4.9 C, red bar). Consistently, up-regulated AS are unlikely to have down-regulated sense transcripts (Fig 4.9 C, left panel grey bar). Since the AS:sense configuration is analogous to divergently transcribed gene pairs, we also compared the 882 divergent gene pairs whose TSSs are less than 200 bp apart. Similar to AS-sense pairs, we observed a relatively strong coregulation between divergent gene pairs, with Pearson $R = 0.39$ (Fig 4.9 D). Similarly, gene pairs that are co-upregulated or co-downregulated are overrepresented whereas differential regulation is underrepresented (Fig 4.9 E).

In contrast, the 200 AS transcripts starting downstream of the TSSs of genes show much weaker coregulation with their sense cognates with Pearson $R = 0.10$ (Fig 4.10 A, B) and the overlap between up-regulated AS and sense are not significantly overrepresented (Fig 4.10 C). Similarly, the AS transcripts located within CDSs of genes also show very little coregulation with the sense transcripts of genes (Fig 4.10 D, E, F).
Figure 4.9: AS transcripts upstream of genes are coregulated with the gene. (A) AS transcripts upstream of genes are coregulated with the gene. (A) Clustered data of the log₂ fold-changes of AS and sense expressions between 0 and 30 minutes of caffeine response, showing the 251 AS transcripts starting within 300 bp upstream of the TSS of protein-coding genes only. (B) Fold-changes of the 5'-upstream AS transcripts against sense transcripts. Pearson correlation ($R = 0.42$) indicates positive correlation. (C) Up-regulated AS have overrepresented overlap with up-regulated sense transcripts (60 genes versus 32 expected, Fisher’s Exact test, $p-value = 2.5 \times 10^{-7}$) and underrepresented overlap with down-regulated sense transcripts (6 genes versus 22 expected, Fisher’s Exact test, $p-value = 3.0 \times 10^{-6}$). Down-regulated AS are too few in numbers to test for significance. (D) Fold-changes of expressions of divergent gene pairs that are <200 bp apart from 0 to 30 minutes. Pearson correlation ($R = 0.39$) indicates positive correlation. (E) Within the 882 divergent gene pairs, both up-regulation (49 pairs versus 32 expected, Fisher’s Exact test, $p-value = 0.0017$) and both down-regulation (29 pairs versus 9 expected, Fisher’s Exact test, $p-value = 2.4 \times 10^{-9}$) are significantly overrepresented. Inversely regulated pairs are underrepresented (7 pairs versus 18 expected, Fisher’s Exact test, $p-value = 0.00083$).
Figure 4.10: AS transcripts downstream of TSS and inside CDSs are less coregulated with sense transcripts. (A) Clustering data of AS and sense expression fold changes between 0 and 30 minutes of caffeine response, showing only the 200 AS transcripts starting within the 300 bp window downstream of TSS. (B) Fold-changes of the 5'-upstream AS transcripts against sense transcripts for the AS transcripts starting at 5'-downstream of TSS. A slight positive correlation is indicated by Pearson correlation. (C) Overlaps between up-regulated AS and sense cognates are not significantly overrepresented for these 200 AS transcripts. (D) Clustering data showing AS transcripts starting inside the CDSs of genes (i.e. >300 bp away from TSS or TES). (E) Fold-changes of the 5'-upstream AS transcripts against sense transcripts for the AS starting inside CDSs. No correlation is observed. (F) Overlaps between up-regulated AS and sense cognates are not significantly overrepresented for these 200 AS transcripts.

4.4.3 AS IncRNAs occur between tandem gene pairs

It has been previously reported that the AS:sense pairs of transcripts are likely to be produced from bidirectional promoters with the NDR length of ~131 bp (Xu et al., 2009). Given that these 221 AS transcripts are within 300 bp upstream of TSSs of genes, it is likely that they are sharing the same bidirectional promoter with the sense transcript of the protein-coding gene. Xu et al. (2009) and Yassour et al. (2010) reported hundreds of AS transcripts produced from bidirectional promoters and the function of most of them were still unknown. Therefore, we next would like to study whether these AS transcripts have functional importance.
in gene regulation or are merely an artefact of gene induction. To address this question, we examined the genomic contexts of these AS transcripts. Surprisingly, we found that these AS transcripts are preferably located between two tandemly arranged protein-coding genes that are >300 bp apart, but not <300 bp apart (Fig 4.11 B). Out of the 221 AS transcripts, 83 (37%) are located between tandemly arranged gene pairs that are >300 bp apart, which is much higher than the expected number due to random chance, which is 12% in the genome (Fig 4.11 A, B). Also, tandem gene pairs that are >300 bp apart have higher AS transcript levels in the intergenic region between them, compared to gene pairs in any other arrangement (Fig 4.11 C).

4.4.4 Tandem gene pairs with intergenic lncRNA tend to be coregulated

We next examined the expression patterns between the AS transcript and the two tandem genes flanking it. Out of the 83 AS transcripts, 20 are terminated within 150 bp around the TES of the upstream adjacent gene (Fig 4.12 A). Interestingly, these 20 AS transcripts show a strong positive coregulation with the sense transcript of the upstream gene (Pearson $R = 0.69$, Fig 4.12 B). For the 63 AS transcripts that are not extending to the 3’ end of the upstream gene, their
levels are not coregulated with the upstream gene (Fig 4.12 C). Given that the AS transcript and the downstream gene itself are coregulated, the sense transcripts of the two genes in this arrangement are also coregulated, which is as expected (Fig 4.12 D). As a negative control, we also examined the 352 tandem gene pairs with no intergenic AS transcripts. Interestingly, these gene pairs show much less coregulation (Pearson $R = 0.11$, Fig 4.12 E), indicating that the AS transcript could be important in facilitating the coregulation between tandemly arranged gene pairs.

Figure 4.12: Sense:AS:sense cluster shows coregulation during caffeine response. (A) Out of the 63 AS transcripts between tandem gene pairs that are $>300$ bp apart, 20 are terminated in the vicinity of the upstream gene’s 3’ end (from -150 bp to +150 bp of TES). (B) For these 20 AS transcripts, fold-changes of AS and the sense of the upstream gene are highly correlated (Pearson $R = 0.69$). (C) For the 63 AS transcripts not ending in the vicinity of the upstream gene’s TES, fold-changes of AS and the sense of upstream gene show little correlation (Pearson $R = 0.69$). (D) Coregulation between the sense transcripts of tandemly gene pairs with intergenic AS transcript (Pearson $R = 0.31$). (E) For the 352 tandem gene pairs (300-1000 bp apart) with no intergenic AS transcripts, fold-changes of their sense transcripts show a much weaker correlation (Pearson $R = 0.11$).
4.5 Discussion

In this chapter, we quantified the transcriptome of \textit{S.cerevisiae} during caffeine response and observed both up- and down-regulation of protein-coding genes in the first 30 minutes of caffeine response. Surprisingly, lncRNA transcripts (specifically the SUTs, due to our experimental protocol) are overwhelmingly up-regulated, which leads to us to investigate the cause of this up-regulation. We first classified the lncRNAs according to their relative organisation to the nearest to protein-coding genes and examined their correlation with gene expression for different classes of lncRNA transcripts.

4.5.1 Organisation of lncRNAs in relation to genes

The majority of lncRNAs are found at promoters of protein-coding genes and are transcribed in the AS direction to the ORF, which is consistent with previous genome-wide studies (Xu et al., 2009; Yassour et al., 2010; Xu et al., 2011). These promoter-associated AS transcripts tend to be located at stress-response genes that are repressed at ground state and induced during caffeine response (Fig 4.5 C). The stress causes simultaneous up-regulation of AS and sense transcripts which is consistent with the patterns of response to growth medium (Xu et al., 2009, 2011). We detected relatively fewer stable lncRNA transcripts (per bp) produced from within CDSs of genes (Fig 4.8 B). Previous studies have shown that the lncRNAs initiated from CDS or 3’ ends, including many cryptic transcripts, tend to anti-correlate their sense cognates (Xu et al., 2009; Yassour et al., 2010; Camblong et al., 2007; van Werven et al., 2012). Unfortunately, due to the limitation of our experimental technique, we could not detect cryptic transcripts in our caffeine datasets. This is probably why we observed a predominantly positive correlation between sense and AS transcripts. Therefore, future study on the cryptic transcripts during stress response would be useful to complete the picture of the non-coding transcriptome in response to stresses \textit{S.cerevisiae}.

4.5.2 AS:sense coregulation is likely due to promoter sharing

Next, we demonstrated that a large and overrepresented proportion (83) of the 221 AS transcripts starting upstream of TSSs of genes are located between
tandemly arranged gene pairs that are >300 bp apart. For the 20 AS transcripts that extend into the 3' end of the adjacent upstream gene, their expression also correlates with the adjacent gene. These three transcripts form a coregulated sense:AS:sense cluster. The pair of AS:sense transcripts is likely to be produced from the same bidirectional promoter due to their close proximity to the TSS. Deducing from previous knowledge, it is likely that the coregulation between the AS:sense pair is due to the promoter sharing (i.e. any chromatin remodelling that occurs during the induction of sense transcripts is likely to affect the AS transcript). It has been previously shown that an essential feature of promoter sharing is that both AS and sense transcripts are in close proximity to the NDR (Neil et al., 2009; Xu et al., 2009). This is why the AS transcripts initiated from the downstream of TSS, which are not close to the NDR, show much less coregulation with the sense transcript of the gene. Furthermore, divergent protein-coding genes are also coregulated in the \textit{S.cerevisiae} genome, indicating that coregulation at bidirectional promoter is probably a general phenomenon of transcription (Fig 4.9 D).

### 4.5.3 Regulation between convergent transcripts

While bidirectionality can explain the coregulation of the divergent AS:sense pairs, the coregulation between the upstream sense transcript and the AS transcript (the convergent sense:AS pair) seems more difficult to explain. The sense:AS pair resembles convergent overlapping protein-coding gene. Coregulations between convergent genes have been noted by Arnone et al. (2014) where they demonstrated that the coregulation of gene pair \textit{MPP10-MRX12} can be disrupted by inserting a long highly expressed gene but not a short tRNA gene, indicating that the overlap is essential for coregulation. This is analogous to our finding that the AS transcripts not extending to the 3' end of the upstream ORF show much less coregulation with the upstream gene (Fig 4.12 C). As for whether extension into the ORF of adjacent convergent gene is sufficient for coregulation, we examined all convergent overlapping protein-coding gene pairs and observed no global coregulation (Fig S4.1). This indicates that more factors than overlap are required for convergent transcripts to be coregulated. Furthermore, our results highlighted the importance of intergenic lncRNAs since tandem gene pairs without intergenic lncRNAs show less degrees of coregulation (Fig 4.12 E). Therefore, it would be interesting for future studies to investigate how lncRNAs are...
involved in the coregulation of tandemly arranged genes.

4.5.4 Possible mechanisms of sense:AS coregulation

Based on previous knowledge, we hereby proposed a few possible mechanisms by which the AS transcripts may regulate the upstream gene. Many AS transcripts that are transcribed from the AS strand of their cognate gene are capable of repressing the gene by blocking gene’s elongation or triggering chromatin remodelling. Well-known examples are the $GAL10$ lncRNA:$GAL10$, $RME2$:IME4,$RME3$:ZIP2 and $PHO84$ lncRNA:$PHO84$, where the AS transcripts interact with chromatin remodellers to silence the promoter of their cognate genes. For the AS transcripts we observed in our sense:AS:sense clusters, their true lengths may be longer than detected since our experiments were conducted the an Rrp6$^+$ background. If the full-length of AS transcripts extended into the coding region of the upstream gene, this would resemble the case of $PHO84$. However, the difference is that the AS transcripts we observed are co-activated with the upstream gene, which presumably causes acetylation rather than deacetylation of the upstream promoter whereas the lncRNA of $PHO84$ causes deacetylation and gene repression (Camblong et al., 2007). Further works can be done to characterise whether AS transcripts or transcription can recruit histone acetyltransferases to chromatin to achieve gene activation.

The coregulation may also signify a particular 3D structure of local chromatin. Nadal-Ribelles et al. (2014) revealed that stress-activated protein kinases induce the AS transcript of $CDC28$ from its 3’ end and cause the 3’ end to form a loop with its 5’ promoter. The loop triggers the spread of 3’ chromatin marks to the 5’ promoter and activates the sense transcription. Therefore, it is reasonable to speculate that our sense:AS:sense coregulation could be due to gene looping such that the promoters of these 3 transcripts are brought into close proximity in the 3D space. We explored this possibility by examining the Hi-C data generated by (Duan et al., 2010) but found no indication of contact (data not shown). Surprisingly, Duan et al. (2010)’s data showed no contact between the 5’ and 3’ ends of $CDC28$ either. A possible explanation is that the long-range chromatin interactions may be too transient or are prone to biological variation.
4.5.5 Comparison to other organisms

Convergent pairs of AS and sense transcripts have been previously reported in the fission yeast S.pombe. A recent study has shown that the AS transcripts that overlap with a convergently arranged gene tend to be negatively coregulated with the gene’s sense transcript during osmostress (Leong et al., 2014), which is different from the positive correlation we observed in the sense:AS pairs in S.cerevisiae. However, the ncRNA system in S.pombe is probably different to S.cerevisiae since AS transcripts in S.pombe tend to be produced by transcription read-through downstream of the 3’ ends of protein-coding genes (Zofall et al., 2009; Leong et al., 2014) whereas the AS transcripts in S.cerevisiae tend to be produced from gene promoters in the AS direction. It has also been shown that the silencing effect between convergent transcripts in S.pombe is related to RNAi (Gullerova and Proudfoot, 2012). The IncRNA antagonises the sense transcript by cooperating with the RNAi machinery (i.e. trans-acting) rather than directly interferes with the sense transcript (i.e. cis-acting). Gullerova and Proudfoot (2012) further demonstrated that the RNAi-mediated silencing by AS transcripts is also present in human. However, since the components of RNAi are missing in S.cerevisiae, the regulatory mechanism of convergent transcripts is likely to be different in S.cerevisiae.

4.5.6 Outstanding questions

1. What are the true lengths of AS transcripts? As mentioned before, our experimental protocol only detects stable transcripts or SUTs. The drawbacks are that we cannot detect cryptic transcripts and also the full lengths of transcripts may not be revealed. Since the degradation of IncRNAs by exosome could leave early-terminated transcripts, it is possible that these AS transcripts are longer than what RNA-seq revealed (Castelnuovo et al., 2014). Therefore, exosome-knockout strains would be helpful in providing extra information on the non-coding transcriptome. Alternatively, techniques such as native elongating transcript sequencing or NET-seq (Churchman and Weissman, 2011) could be used without altering the genome.

2. Are AS and sense transcripts expressed in the same cell? It also remains to be established whether the AS and sense transcripts in the
sense:AS:sense clusters are expressed in the same cell simultaneously. Previously, both simultaneous and mutually exclusive relationships between AS and sense pairs have been detected by single-cell RNA FISH. It is possible that the AS and sense transcripts are expressed in different cells and are both up-regulated in caffeine response. However, mutually exclusive AS:sense pairs have only been observed when AS is repressing the expression of sense. In the case of co-upregulated AS:sense pairs, the transcripts were known to be expressed in the same cell (Nadal-Ribelles et al., 2014).

Figure 4.13: Coregulated sense:AS:sense cluster. (A) Tandemly arranged gene pairs without intergenic antisense IncRNA are not coregulated. (B) Tandem gene pairs with AS transcripts extending into the 3’ end of adjacent genes are coregulated. Coregulation between AS and the downstream sense transcript is likely due to promoter sharing. The mechanism of coregulation between AS and the upstream sense transcript remains to be established. AS transcripts not extending into the 3’ end of upstream gene is not coregulated with the upstream gene.

4.6 Summary

In this chapter, we examined the patterns of non-coding transcriptome of *S. cerevisiae* in response to caffeine stress. We observed a global increase in IncRNAs, which predominantly occurs at the promoters of stress-induced genes and are transcribed in the antisense direction. The IncRNAs are positively coregulated with the sense transcript of their cognate genes. In particular, the ones
sharing the same promoter with the gene show the strongest positive correlation
with the sense transcript. These lncRNAs tend to be found between tandemly
arranged gene pairs that are > 300 bp apart and are also coregulated with the
gene upstream. These three transcripts form a coregulated sense:AS:sense cluster
that signifies a specific type of regulatory module in the \textit{S. cerevisiae} genome (Fig
4.13 B). The lncRNA appears to be required for the coregulation between tandem
genes (Fig 4.13 A, B), indicating that lncRNAs may act as a medium to spread
regulatory signals from an activated gene to nearby genomic locations.
4.7 Supplementary Figure

Figure S4.1: Convergent gene pairs are not coregulated in caffeine response.
Chapter 5

Delayed caffeine-induced H2A.Z Drop in Mutants Causes Slower Gene Recovery

5.1 Introduction

5.1.1 H2A.Z becomes essential under stress

In *S. cerevisiae*, H2A.Z is non-essential under normal growth conditions. However, the *htz1Δ* mutant shows impaired growth in various stressed conditions such as temperature changes (Santisteban et al., 2000), switch of nutrient switch from glucose to galactose (Adam et al., 2001), DNA damage (Morillo-Huesca et al., 2010; Mehta et al., 2010), microtubule destabilisation (Mehta et al., 2010), oleate stress (Wan et al., 2009) and caffeine stress (Wang et al., 2011a). Impaired growth is accompanied by failure in recruiting general transcription factors and RNA polymerase II to the promoters of stress response genes (Wan et al., 2009). Wan et al. (2009) further demonstrated that several stress-response genes in the *htz1Δ* mutant fail to disassemble their promoter nucleosome that is required for gene activation. However, the exact mechanism of how H2A.Z contributes towards gene activation has not been fully understood.
5.1.2 Cellular H2A.Z abundance decreases during caffeine response

Previously, our group (Yanin Naiyachit) has conducted western blot to probe the change of total cellular abundance of H2A.Z during caffeine response. Interestingly, our western blots showed that the total cellular abundance of H2A.Z (including both chromatin and cytoplasmic H2A.Z) is decreasing in WT cells during caffeine response (Fig 5.1 A, B). Here, it is important to understand the cause of the decrease in H2A.Z abundance. The decrease in cellular levels of H2A.Z could be either due to H2A.Z being expressed at a lower level or increased rate of degradation. Moreover, the level of chromatin-bound H2A.Z remains to be determined. It would be worth investigating whether chromatin-bound H2A.Z agrees with the patterns of cellular levels during caffeine response. To answer these question, we performed ChIP-seq to quantify H2A.Z levels that is bound to the chromatin. Since chromatin-bound H2A.Z is presumably directly related to transcription, we also integrated with the transcriptomic data to study the relationship between H2A.Z and transcription during stress response.

5.1.3 Mutant strains to mimic effect of \(_{htz1}\Delta\)

In the previous chapters, we established that H2A.Z is an activator for transcript levels in steady states. It would be interesting to study the roles of H2A.Z during stress response. Unfortunately, we could not generate viable \(_{HTZ1}\) knockouts for caffeine samples because the cell growth is severely impaired in the \(_{htz1}\Delta\) strain. To seek for alternatives, we generated H2A.Z mutants to simulate the effect of \(_{htz1}\Delta\). We first generated the 4KR mutant, which is an unacetylatable mutant of H2A.Z with all 4 lysines in the N-terminal tail mutated to arginine. From our previous work, we knew that the 4KR mutant moderately affects cell growth under caffeine treatment and the steady-state abundance of H2A.Z is largely unaffected (Fig 5.1 B). Surprisingly, we observed a slower decrease of H2A.Z levels in 4KR compared to WT, the reason of which was unknown. We also created an overexpression (OE) mutant, which provides excessive amount of H2A.Z in the cell. Despite having higher initial abundance of H2A.Z than WT and 4KR, the OE mutant also show slower rate of H2A.Z drop during the caffeine time course than WT (Fig 5.1 B).
5.1.4 Objective

Our western blots quantified the total cellular abundance of H2A.Z, which contains both cytoplasmic and chromatin-bound H2A.Z. In order to study its relationship with transcription, we need to measure the chromatin level of H2A.Z in each sample. To quantify the abundance, we performed ChIP-seq experiments on wild type, 4KR and OE mutants over the caffeine time course. The first objective is to examine whether chromatin-bound H2A.Z shows the same decreasing trend as cellular H2A.Z. With these data, we will also be able to pinpoint the exact location in the genome where the caffeine-induced H2A.Z drop is occurring. Our next objective is to study whether the H2A.Z drop is linked with gene activation or repression. We would also investigate whether chromatin-bound H2A.Z in 4KR and OE shows reduced drop rates compared to WT and whether H2A.Z have functional roles in transcriptional regulation during gene activation or repression.

5.1.5 Limitation of conventional normalisation methods

Since we are facing a situation where the genomic total of ChIP signals is varied between samples, it is crucial to use the right normalisation method to obtain
meaningful results. Pipelines for chromatin-probing next-generation sequencing analysis (e.g. ChIP-seq, MeDIP-seq, MNase-seq, DNase-seq, Faire-seq, ATAC-seq and etc.) generally fall in to two categories. The method commonly used by early studies usually involves "peak-calling", where peaks of ChIP signals are first identified from background noises. Genomic regions are regarded as either containing or not containing the signal in a binary fashion. Therefore, these methods are more suited for qualitative studies such as TF binding sites and well-positioned nucleosome occupancies. Due to its binary nature, normalisation is usually not required. The second method involves quantification of ChIP signals in the genomic locations of interest before comparisons between samples. Careful normalisation is important for this type of methods.

A good normalisation needs to counter both internal and external biases in sequencing samples. Internal biases are usually contributed by experimental techniques that favours or disfavours particular DNA sequences in one experiment. Well studied sources of internal biases include chromatin fragmentation (Teytelman et al., 2009), GC contents (Dohm et al., 2008), PCR amplification (Gilfillan et al., 2012) and read mappability (Lee and Schatz, 2012). To correct internal biases, non-linear normalisation methods are used to adjust the sequencing signals from different genomic locations to a comparable level and therefore, is essential for comparisons within a specific dataset. External biases are introduced by differences in experimental conditions such as antibody affinities, genomic noise levels, sequencing depths and many protocol-specific factors. To correct external biases, linear methods are often used to multiply the signal of each sample by a different factor such that all samples are adjusted to a comparable level. Widely used methods include depth normalisation (e.g. normalised to reads per kilobase per million mapped reads, RPKM) and quantile normalisation, which makes the signal distributions of signals between different samples identical.
5.2 Decreased Chromatin-bound H2A.Z is Revealed by Normalised Data

Both depth and quantile normalisations are based on the assumption that the genomic total of ChIP signals in all samples are the same. In this study however, we face a situation where the global levels of H2A.Z could be decreasing during caffeine time course, indicating that an alternative normalisation method is required. In this section, we will describe the new normalisation method we developed to suit our ChIP-seq data and demonstrate its good agreement with western blot results.

5.2.1 Signal-to-background ratios decrease during caffeine response

First, we assessed the likelihood of non-constant ChIP signals by examining the signal-to-background ratios across samples. To quantify signal-to-background ratios, we started with unnormalised data and separated ”signal regions” from ”background regions” using a preliminary set of H2A.Z peaks obtained from background-subtracted WT 0 minute data. Signal regions are usually around gene promoters whereas background regions are usually within CDSs or IGRs(Fig 5.2 A). For both signal and background regions, we quantified the ChIP signals within and obtained two distributions. Then the signal-to-background ratio was estimated by the median of signal distribution divided by the median of background distribution (Fig 5.2 B-H). A decreasing trend of signal-to-background ratios was observed during time course. The decreasing trend suggests two possibilities: either H2A.Z levels are decreased while the genomic background is unchanged or H2A.Z levels are unchanged while the genomic background is increased. Given the fact that all our samples were prepared in the same batch of experiments, it is likely that genomic backgrounds are relatively constant across samples.
Figure 5.2: Decreased signal-to-background ratios of ChIP-seq during caffeine response. (A) Genomic landscapes (a section of Chromosome 3) of unnormalised ChIP-seq data at 0, 30 and 120 minutes in caffeine response. (B) (C) Signal and background distributions of ChIP signal and background levels in 0-minute WT data. Bases located within peak regions are classified as signal and outside are classified as background. Distributions of raw reads for signal and background are shown in the density plots. (D) (E) Signal and background in 30 minutes wild type. (F) (G) Signal and background in 120 minutes wild type. (H) Decreasing signal-to-background ratios over time. The heights of bar represent the ratio between the median of signal and background. Upper error bar represents the median + 0.5 standard deviation of signal divided by the median - 0.5 standard deviation. Lower error bar represents the median -0.5 standard deviation of signal divided by the median +0.5 standard deviation.
5.2.2 Normalisation for WT and 4KR agrees with experimental data

Assuming the genomic background is constant between samples, we developed a customised normalisation method for WT and 4KR samples. We first identified the set of preliminary H2A.Z peaks from each unnormalised dataset and merged them into a union set. The genomic regions outside preliminary peaks were extracted and quantified as the background distribution. Datasets were then normalised such that all background distributions were on the same level and subsequently background subtraction was carried out using input DNA (Fig 5.3 A). Normalised data reveal decreasing trends of chromatin-bound H2A.Z over the caffeine time course in both WT (Fig 5.3 B, D) and 4KR (Fig 5.3 C, D). These results suggest that chromatin-bound H2A.Z follows similar decreasing patterns to the cellular abundance we observed in western blots (Fig 5.1 B).

5.2.3 OE samples are normalised to sequencing depth

In OE samples, we observed significantly higher H2A.Z abundance outside gene promoters, relative to the promoter level (Fig 5.4 B), suggesting that either large amount of H2A.Z have been deposited to CDSs and IGRs or promoter levels have been significantly reduced. In either case, the genomic total would not be the same to WT or 4KR, indicating that the background normalisation would not be suitable. For the OE samples, we adopted the conventional depth-normalisation where each dataset is normalised to the number of uniquely mapped reads. For convenience, we adjusted the OE samples by the same factor such that the promoter H2A.Z peaks in 0-minute OE are of similar heights to 0-minute WT (Fig 5.4 A). It is worth noting that H2A.Z levels in OE should not be directly compared to WT or 4KR. In agreement with western blots, normalised OE samples also show decreased levels of H2A.Z during the caffeine time course (Fig 5.4 C).
Figure 5.3: Normalised ChIP-seq data show decreased chromatin-bound H2A.Z during caffeine response. (A) Flow chart of normalisation. In this method, raw H2A.Z peaks were first identified for each dataset by subtracting the input DNA. Peaks with overlaps >20 bp were merged into a union set and genomic regions with pure background were isolated. Top and bottom 20% were removed from each background distribution and factors between the median of each dataset and wild type 0 minute data were obtained. All datasets were normalised to their factor such that all backgrounds are on the same level. Then the background were subtracted using the input DNA sample. (B) Genomic landscapes (a section of Chromosome 3) of normalised H2A.Z data. (C) Distribution of normalised H2A.Z per base for 0, 30 and 120 minutes after caffeine treatment. (D) Total chromatin levels of H2A.Z are decreasing.
Figure 5.4: Normalised data in OE strain show decreased chromatin levels of H2A.Z. (A) Normalisation pipeline for ChIP-seq data in OE strain. Unnormalised data at 0, 30 and 120 minutes were first normalised to sequencing depths (i.e. numbers of uniquely mapped reads). Depth-normalised data were then multiplied by the same factor such that 0 min OE matches 0 min WT in H2A.Z peak heights. Background subtraction was performed using the background level in 0 min WT data. (B) Genomic landscape of normalised data in OE strain on Chromosome 3. (C) Total normalised ChIP-seq signals at 0, 30 and 120 minutes of caffeine response in the OE strain.
5.3 H2A.Z Drop Occurs at Promoters

Having established that chromatin-bound H2A.Z is decreasing during caffeine responses, we next investigated where in the genome this decrease is occurring. Given that promoters contain the vast majority of H2A.Z, one would expect that the H2A.Z drop is happening at gene promoters. Since we have previously established that a significant proportion of H2A.Z is located outside gene promoters and may have different functions to the promoter H2A.Z, it would be worth studying the patterns of those H2A.Z separately. Therefore, we next examined the H2A.Z at promoters and CDSs of protein-coding genes separately and quantified their contribution towards the global H2A.Z drop during caffeine response.

5.3.1 Promoter H2A.Z decreases during caffeine response

To quantify the H2A.Z signals at the 5’ ends of genes, we summed up the normalised ChIP-seq signals within 150-bp windows downstream of TSS and normalised to the window length. The 5’ H2A.Z occupancy shows a decreasing trend over time course, which is similar to the chromatin-bound H2A.Z abundance (Fig 5.5 A). Out of the 5143 protein-coding genes, we identified 1027 genes with significantly decreased 5’ H2A.Z occupancy by >2 fold at 30 minutes and 383 genes with significantly increased 5’ H2A.Z (Fig 5.5 D). Genes with increased 5’ H2A.Z tend to have less occupied promoters by H2A.Z at 0 minute (Fig 5.5 B, E, red) whereas genes with decreased 5’ H2A.Z tend to have more highly occupied promoters by H2A.Z at 0 minute (Fig 5.5 B, E, blue). Consistent with this, genes with low 0-minute H2A.Z (0-50 RPKM) are more likely to have increased H2A.Z rather than decreased H2A.Z occupancy (Fig 5.5 C) whereas genes with 100-200 RPKM at 0 minute are more likely to have decreased H2A.Z (Fig 5.5 C).
Figure 5.5: Decrease of H2A.Z levels at the 5' end of genes during caffeine response. (A) Hierarchically clustered heatmap showing the decrease of H2A.Z occupancy at the 5' ends of genes during caffeine response. H2A.Z levels at 0 minute are significantly higher than 30 minutes (two-tailed t-test, \textit{p-value} = 7.6 \times 10^{-32}); H2A.Z at 30 minutes is significantly higher than 120 minutes (two-tailed t-test, \textit{p-value} = 5.8 \times 10^{-135}). (B) Fold-changes of 5' H2A.Z levels between 0 and 30 minutes against H2A.Z occupancy at 0 minute. Genes with significantly increased or decreased H2A.Z occupancy are coloured in red and blue respectively. (C) Genes were put into 5 bins according to their 5' H2A.Z occupancy at 0 minute (i.e. 0-50, 50-100, 100-200, 200-300 and >300 RPKM). For each bin, the significance of overlaps with increased / decreased H2A.Z was tested using Fisher’s Exact Tests. P-values for the 5 bins with increased H2A.Z (red) are 3.6 \times 10^{-5}, 0.078, 6.4 \times 10^{-4}, 0.032 and 0.012. P-values for the 5 bins with decreased H2A.Z (blue) are 6.9 \times 10^{-83}, 0.021, 6.0 \times 10^{-29}, 1.2 \times 10^{-5} and 0.049. (D) Numbers of genes with significantly increased or decreased 5' H2A.Z at 30 minutes. (E) Genes with increased 5' H2A.Z levels (red) have significantly lower 0-minute H2A.Z occupancy than genes whose 5' H2A.Z occupancies are not changed (two-tailed t-test, \textit{p-value} = 1.2 \times 10^{-4}); Genes with decreased 5' H2A.Z levels (blue) have significantly higher H2A.Z occupancy at 0 minute (two-tailed t-test, \textit{p-value} = 2.6 \times 10^{-54}).
We next examined the patterns of H2A.Z dynamics after the initial decrease. Between 30 and 120 minutes, we observed a general trend of H2A.Z recovery from the increase or decrease that occurred at 30 minutes (Fig 5.6 A). This is confirmed by the global negative correlation between fold-changes of 5’ H2A.Z from 0 to 30 minutes and from 30 to 120 minutes (Fig 5.6 B). Notably, almost all the genes whose promoter H2A.Z levels are increased by 30 minutes (317 genes out of 383) have decreased by 120 minutes (Fig 5.6 C) and the number is significantly higher than random chances (Fig 5.6 D, red bar). In contrast, out of the 1027 with decreased 5’ H2A.Z at 30 minutes, only 55 are significantly recovered by 120 minutes (Fig 5.6 C). However the 55 genes are also significantly higher than random chances (Fig 5.6 D, blue bar), which is due to the fact that very few genes (81) have increased 5’ H2A.Z from 30 to 120 minutes. Genes with increased H2A.Z by 30 minutes tend to recover to their original level by 120 minutes whereas the recovery from decreased H2A.Z is only by 30% of the magnitude of 30-minute drop (Fig 5.6 E, F). Despite the trend of H2A.Z recovery, global levels of 5’ H2A.Z are still decreasing at 120 minutes (Fig 5.5 A, box plot). Surprisingly, genes with continual decrease of 5’ H2A.Z over time course (85 genes) are significantly underrepresented (Fig 5.6 C, D, black bar).
Figure 5.6: Recovery of 5’ H2A.Z occupancy from 30 minutes to 120 minutes.

(A) Hierarchically clustered data showing the inverse relationship between 5’ H2A.Z occupancy changes from 0 to 30 minutes and from 30 to 120 minutes. Degrees of recovery by 120 minutes, as fractions of the differences between 0 and 30 minutes, are indicated for the genes with significant changes in 5’ H2A.Z occupancy. (B) Negative correlation between fold-changes of 5’ H2A.Z from 30 to 120 minutes and from 0 and 30 minutes. Numbers of genes in each quadrant were indicated on the plot. Genes with increased and decreased 5’ H2A.Z occupancy at 30 minutes are coloured red and blue respectively. Regression line is $y = -0.42x - 0.85$ and Pearson correlation $R = -0.65$. (C) Numbers of genes whose 5’ H2A.Z occupancies are significantly changed from 0 to 30 minutes and from 30 to 120 min. (D) Genes with increased 5’ H2A.Z from 0 to 30 minutes tend to have decreased 5’ H2A.Z from 30 to 120 minutes (red bar 317 genes, Fisher’s Exact test $p$-value $= 2.4 \times 10^{-157}$). Genes with decreased 5’ H2A.Z from 0 to 30 minutes tend to increase from 30 to 120 minutes (blue bar 55 genes, Fisher’s Exact test $p$-value $= 5.1 \times 10^{-21}$). Genes with continuously decreasing 5’ H2A.Z are underrepresented (black bar 85 genes, Fisher’s Exact test $p$-value $= 3.7 \times 10^{-37}$). (E) Extent of recovery of 5’ H2A.Z by 120 minute. The 317 genes whose 5’ H2A.Z levels are recovered from increase (red boxes) are recovered to original level whereas the 55 genes whose 5’ H2A.Z levels are recovered from decrease (blue boxes) are only recovered by $\sim 30\%$. (F) Recovery from increased H2A.Z is to a greater extent than recovery from H2A.Z drop (two-tailed t-test $p$-value $= 3.8 \times 10^{-28}$). Extent of recovery was defined as the change between 30 and 120 minutes as a fraction of the change between 0 and 30 minutes.
5.3.2 H2A.Z in coding regions increases during caffeine response

To investigate the pattern of H2A.Z inside CDSs of genes, we first defined CDSs as the ORF regions that are >300 bp from both TSS and TES. We ignored genes that are shorter than 600 bp and quantified the levels of H2A.Z inside CDS by normalising ChIP signals to the lengths of CDS and number of uniquely mapped reads. Surprisingly, we observed an overall increase of CDS H2A.Z levels at 30 minutes of caffeine response (Fig 5.7 A). In the genome, we observed 116 genes whose CDS H2A.Z levels are significantly increased whereas only 55 genes have decreased CDS H2A.Z at 30 minutes. The increase of H2A.Z tends to happen at CDSs that are relatively unoccupied by H2A.Z and decrease tends to happen at highly H2A.Z-occupied CDSs. (Fig 5.7 B). By 120 minutes, CDS H2A.Z is also recovering, show a negative correlation with the initial change at 30 minutes (Fig 5.7 C). The initial increase by 30 minutes is generally recovered to steady-state level by 120 minutes (Fig 5.7 D). However, the initial increase of CDS H2A.Z is uncorrelated with the change of sense or AS transcripts (Fig S5.2). It remains unclear what triggered this H2A.Z increase inside CDSs of genes.
Figure 5.7: Change of H2A.Z occupancy inside coding regions of genes during caffeine response. (A) Hierarchically clustered data showing the change of H2A.Z occupancy inside CDSs of genes during caffeine response in WT cells. CDS H2A.Z levels are significantly increased at 30 minutes (two-tailed t-test, \( p\)-value = 0.0013) and decreased by 120 minutes (two-tailed t-test, \( p\)-value = 0.0094). (B) Fold-changes of CDS H2A.Z between 0 and 30 minutes against the initial occupancy at 0 minute. Genes with decreased CDS H2A.Z have higher initial occupancy (two-tailed t-test, \( p\)-value = 3.9 \( \times \) 10\(^{-5}\)) and genes with increased CDS H2A.Z have lower initial H2A.Z (two-tailed t-test, \( p\)-value = 0.0017) than genes with unchanged CDS H2A.Z. (C) Negative correlation between fold-changes of CDS H2A.Z occupancy from 0 to 30 minutes and from 30 to 120 minutes with Pearson correlation = -0.44. (D) Degrees of recovery at 120 minutes for genes with increased occupancy at 30 minutes.
5.4 Promoter H2A.Z Drop is Linked with Gene Activation

During the transcriptional response to environmental stresses, a large proportion of the transcriptome is altered, which is accompanied by vigorous remodelling at chromatin level (Shivaswamy and Iyer, 2008; Weiner et al., 2015). Our next objective is to study whether the decreased promoter levels of H2A.Z is related to transcriptional changes. We integrated the RNA-seq data from Chapter 4 and examined the correlation between H2A.Z and transcripts during caffeine response.

5.4.1 Changes in 5’ H2A.Z levels anti-correlate changes in gene expression

We first investigated the relationship between changes in H2A.Z occupancy and gene expression during caffeine response. We took the 2233 genes that are either differentially expressed or with altered 5’ H2A.Z levels at 30 minutes of caffeine response and observed largely opposite patterns of the changes of H2A.Z and expression (Fig 5.8 A). On the global scale, the fold-changes of 5’ H2A.Z strongly anti-correlate the fold-changes in expression (Fig 5.8 B). Out of the 1027 genes with decreased 5’ H2A.Z at 30 minutes, an overrepresented proportion (499 genes) are up-regulated (Fig 5.8 C, red bar). Similarly, 292 out of the 383 genes with increased 5’ H2A.Z are significantly down-regulated between 0 and 30 minutes (Fig 5.8 C, blue bar). Between 30 and 120 minutes, the recovery of 5’ H2A.Z is also anti-correlated with the change of gene expression (Fig 5.8 E, F). Similarly, the overlaps between down-regulated genes and increased H2A.Z occupancy and between up-regulated genes and decreased H2A.Z occupancy are both overrepresented (Fig 5.8 D).
CHAPTER 5. DELAYED GENE RECOVERY IN H2A.Z MUTANTS

Figure 5.8: Negative correlation between changes of expression and 5’ H2A.Z occupancy from 0 to 30 minutes. (A) Hierarchically clustered data showing anti-correlation between the fold-changes of expression and 5’ H2A.Z occupancies from 0 to 30 minutes. The 2233 genes with significantly changed 5’ H2A.Z occupancies or gene expression are displayed. (B) Negative correlation between the fold-changes of expression and 5’ H2A.Z occupancy from 0 to 30 minutes. Activated genes with decreased 5’ H2A.Z are shown in red and deactivated genes with increased 5’ H2A.Z are shown in blue. Numbers of genes in each quadrant are shown on the graph. (C) Activated genes tend to have decreased 5’ H2A.Z (Fisher’s Exact test p-value = 2.7 × 10^{−178}) and deactivated genes tend to have increased 5’ H2A.Z (Fisher’s Exact test p-value = 1.1 × 10^{−100}). The overlaps between activated genes and increased 5’ H2A.Z and that between deactivated genes and decreased 5’ H2A.Z are both significantly underrepresented (Fisher’s Exact test p-values = 2.8 × 10^{−33} and 1.0 × 10^{−80} respectively). (D) Between 30 and 120 minutes, up-regulated genes tend to have decreased 5’ H2A.Z (Fisher’s Exact test p-value = 1.5 × 10^{−13}) and down-regulated genes tend to have increased 5’ H2A.Z (Fisher’s Exact test p-value = 1.0 × 10^{−50}). The overlaps between down-regulated genes and decreased 5’ H2A.Z is significantly underrepresented (Fisher’s Exact test p-values = 2.0 × 10^{−138}). (E) Negative correlation between the fold-changes of expression and 5’ H2A.Z occupancies from 30 to 120 min.
5.4.2 H2A.Z drop at 30 minutes is primarily contributed by gene activation

Having established the link between 5’ H2A.Z drop and gene activation at 30 minutes, we then quantified how much of the 5’ H2A.Z drop can be explained by gene activation. We observed that out of the 1027 genes with decreased 5’ H2A.Z at 30 minutes, the vast majority (962) are either highly activated or moderately activated (Fig 5.9 A). Consistently, activated or highly activated genes show significantly greater H2A.Z drop at promoters than genes that are not differentially regulated (Fig 5.9 B). We estimated that activated or highly activated genes contribute towards 96% of the H2A.Z drop between 0 and 30 minutes of caffeine response (Fig 5.9 C). From 30 to 120 minutes, the decrease of 5’ H2A.Z is also correlated with up-regulated genes, which are the ones recovered from deactivation at 30 minutes (Fig 5.9 D). We also observed 195 genes with both decreased 5’ H2A.Z and expression levels (Fig 5.9 D), which is likely a result of global H2A.Z drop between 30 and 120 minutes. In contrast, up-regulation only contributes towards 77% of the H2A.Z drop (Fig 5.9 E, F).

5.5 Delayed H2A.Z Drop Accompanies Reduced Gene Recovery

We have demonstrated that the caffeine-induced H2A.Z drop occurs at gene promoters and is predominantly happening at induced genes. We next would like to investigate the effect of H2A.Z levels on the induction of these genes. The ideal method is to compare the transcriptome of \textit{HTZ1} deletion strain with WT. Unfortunately, we could not generate viable \textit{htz1}\text{\textsuperscript{\textDelta}} strains that grow under caffeine stresses. To seek for alternatives, we used the 4KR and OE strains to provide decreased and increased 0-minute levels of H2A.Z respectively. Our next objective is to study how genes are induced (or repressed) differently in mutant strains compared to WT.
Figure 5.9: Promoter H2A.Z drop is contributed by gene activation. (A) Numbers of highly activated (significantly up-regulated), activated (up-regulated but not significant) and not differentially expressed genes at 30 minutes of caffeine response. (B) Highly activated genes have greater 5’ H2A.Z drop than activated genes (two-tailed t-test, p-value = $1.5 \times 10^{-112}$), which has greater 5’ H2A.Z drop than genes that are not differentially expressed (two-tailed t-test, p-value = 0.0031). (C) Contribution towards the 5’ H2A.Z drop (within the 1027 genes with decreased 5’ H2A.Z at 30 minutes) by highly activated, activated and not-differentially-expressed genes. Top panel shows the total H2A.Z drop and bottom panel shows the H2A.Z drop per gene from 0 to 30 minutes. (D) Numbers of highly up-regulated (significantly up-regulated), up-regulated (up-regulated but not significant), down-regulated and not differentially expressed genes at 120 minutes compared to 30 minutes of caffeine response. (E) Highly activated genes have greater 5’ H2A.Z drop than activated genes (two-tailed t-test, p-value = $2.3 \times 10^{-12}$). (F) Contribution towards the 5’ H2A.Z drop.
5.5.1 4KR and OE have negligible impact on steady-state transcriptome

To rule out the possibility that mutations disrupt the genome stability, we assessed the transcriptomes of mutant strains before caffeine treatment. At 0 minute, global patterns of gene expression are largely unaffected by either 4KR or OE (Fig 5.10 A, B, Pearson correlations). Using very tolerant thresholds, we were only able to identify 65 genes (∼1% of genome) that are differentially expressed in the 4KR strain (Fig 5.10 A). OE strain imposes slightly greater effects on the steady-state transcriptome, with 146 down-regulated genes compared to WT (Fig 5.10 B). Nevertheless, we observed no Gene Ontology enrichment or down-regulation of transcriptional or translational machinery in either strain. These results indicate that 4KR and OE mutations impose negligible effect on the stability of the cell.

Figure 5.10: 4KR and OE mutations have minor effects on 0-minute gene expression. (A) Expression of WT against 4KR at 0 minute. Up- and down-regulated genes were calculated using very tolerant thresholds: fold-changes >1.1 and p-values <0.5. Dotted lines indicates equal expression and Pearson correlation is shown on the top. (B) Expression of WT against OE at 0 minute.

5.5.2 Dynamic range of H2A.Z is reduced in 4KR and OE strains

Our previous western blot data have shown that the drop rates of celluar H2A.Z levels in both mutant strains are reduced compared to WT (Fig 5.1 B). We next investigated whether this is also true for the chromatin-bound levels of H2A.Z. To establish this, we compared the ChIP-seq for mutant samples with WT over
the time course. Although the patterns of H2A.Z during time course are largely unaffected by 4KR or OE, the dynamic ranges in both mutants are reduced compared to WT (Fig 5.11 A, B). Among the genes whose 5’ H2A.Z occupancies are decreased at 30 minutes, 654 genes have less drop in 4KR than in WT, which is much higher than the 132 having more drop in 4KR than in WT (Fig 5.11 A). Similar patterns were also observed in the OE strain with even greater reduction in dynamic ranges (Fig 5.11 B). To test for the significance, we plotted the dynamic ranges of 5’ H2A.Z between 0 and 30 minutes for WT, 4KR and OE and observed significantly reduced dynamic ranges for both increased and decreased H2A.Z occupancies (Fig 5.11 C). Given that more genes have decreased rather than increased 5’ H2A.Z during caffeine response, the reduced drop of H2A.Z in mutant strains are probably contributed by reduced dynamic ranges at gene promoters with decreased H2A.Z occupancy.

Figure 5.11: Reduced dynamic ranges of promoter H2A.Z in 4KR and OE strains. (A) Fold-changes of 5’ H2A.Z from 0 to 30 minutes of WT against 4KR. Solid grey line indicates $y = x$, which represents equal magnitude of changes. Genes whose dynamic ranges of 5’ H2A.Z in 4KR are greater/smaller than WT (by $>0.5 / <0.5$ on log$_2$ scale) are highlighted in red and blue respectively. (B) Fold-changes of 5’ H2A.Z from 30 to 120 minutes of WT against 4KR. (C) Reduced dynamic ranges of 5’ H2A.Z. For the 1027 genes with decreased 5’ H2A.Z at 30 minutes in WT, the drop in 4KR is significantly reduced compared to WT (two-tailed t-test, $p$-value $= 2.1 \times 10^{-32}$) and the drop in OE is significantly reduced compared to 4KR (two-tailed t-test, $p$-value $= 3.8 \times 10^{-17}$). For the 383 genes with increased 5’ H2A.Z at 30 minutes, the increase in 4KR is significantly reduced compared to WT (two-tailed t-test, $p$-value $= 6.2 \times 10^{-5}$).
5.5.3 4KR and OE share the same set of genes with reduced H2A.Z drop

We next focused on the genes with reduced drop rate of 5' H2A.Z in mutants than in WT. We identified 167 genes whose 30-minute drop of 5' H2A.Z in 4KR is reduced by over 2 times compared to WT and 310 genes with reduced 5' H2A.Z drop in OE. These two sets share a highly overrepresented overlapping set of 127 genes (Fig 5.12 C). We then took the union set that contains 350 of genes with reduced H2A.Z drop in either 4KR or OE. In an overview of H2A.Z for these 350 genes, reduced drop in 4KR and OE compared to is visible (Fig 5.12 A). Notably, the H2A.Z levels in OE is even more reduced than in 4KR at 30 minutes (Fig 5.12 B). Gene Ontology of these 350 genes highlights functional classes such as electron carrier, calmodulin binding and oxidoreductases (Fig 5.12 D).

5.5.4 Reduced H2A.Z drop is accompanied by slower H2A.Z recovery

We next focused on these 350 genes with reduced H2A.Z drop in either 4KR or OE and investigated their patterns from 30 to 120 minutes. We plotted the median levels of H2A.Z for these 350 genes and observed that in WT, H2A.Z has reached a minimum by 30 minutes and is recovering by 120 minutes whereas in 4KR and OE the H2A.Z is still decreasing from 30 to 120 minutes (Fig 5.13 A). This is confirmed by the fact that fold-changes of 5' H2A.Z from 30 to 120 minutes are recovering to much less extent in 4KR and OE compared to WT (Fig 5.13 B). In WT, we observed 207 genes with recovering H2A.Z occupancy and 91 genes with continuously decreasing H2A.Z from 30 to 120 minutes (Fig 5.13 C, left bars). In 4KR and OE, much fewer genes are recovering and more genes are decreasing compared to WT (Fig 5.13 C). These observations indicate that the reduced H2A.Z drop in mutant strains reflects delayed H2A.Z drop or less recovery by 120 minutes.
Figure 5.12: Reduced drop of 5’ H2A.Z at 30 minutes in mutant strains compared to WT. (A) Hierarchically clustered data of 5’ H2A.Z occupancies for WT, 4KR and OE strains at 0 and 30 minutes. Only the 350 genes with reduced 5’ H2A.Z drop at 30 minutes in either 4KR or OE strain compared to WT are shown. (B) Distributions of 5’ H2A.Z levels for the 350 genes in WT, 4KR and OE. In the OE strain, the drop is more reduced compared to 4KR (two-tailed t-test, \( p\)-value \( = 8.1 \times 10^{-8} \)). (C) Numbers of genes with reduced 5’ H2A.Z drop in 4KR and OE strains. The overlap is significantly overrepresented (Fisher’s Exact test, \( p\)-value \( = 5.6 \times 10^{-130} \)). (D) GO categories with Bonferroni-corrected \( p\)-values less than 0.005. Upper panel shows the significance indicated by \(-\log_{10}(p\text{-value})\) and lower panel shows the number of genes in the category.
CHAPTER 5. DELAYED GENE RECOVERY IN H2A.Z MUTANTS

Figure 5.13: Reduced H2A.Z drop reflects delayed H2A.Z drop during caffeine response. (A) Median levels of 5' H2A.Z occupancies for the 350 genes with reduced H2A.Z drops in WT, 4KR and OE strains over caffeine time course. (B) Hierarchically clustered heat map of the log2 fold-changes of 5' H2A.Z occupancies from 30 to 120 minutes. (C) Numbers of genes with recovering and continuously decreasing 5' H2A.Z by 120 minutes for WT, 4KR and OE strains. Genes with log2(fold-change) >0.1 are classified as recovering and <0.1 are classified as continuously decreasing.

5.5.5 Majority of reduced H2A.Z drop is independent of gene activation

We have identified 350 genes whose 5' H2A.Z drop is reduced in mutant strains compared to WT. It is important to distinguish whether the reduced H2A.Z drop is due to less gene induction or is independent of gene induction. Gene induction is generally accompanied by displacement of promoter nucleosome and hence decreased promoter H2A.Z occupancy. However, the reduced H2A.Z drop that is
independent of gene induction would be of interest. To establish the activation-independent ones, we quantified the degrees of gene induction by the fold-changes at 30 minutes compared to 0 minute. We defined the activation-independent genes as the ones with similar expression levels (i.e. less than ±0.05%) in mutant and WT strains at both 0 and 30 minutes. It turns out that the majority (86% in 4KR and 73% in OE) of the genes with reduced 5’ H2A.Z drop are independent of gene activation (Fig 5.14 A). Figure 5.14 C and D are two examples of the activation-dependent and independent genes with reduced H2A.Z drop in 4KR and OE. The 4KR and OE strains share a core set of 228 genes with reduced H2A.Z drop that are activation-independent in both 4KR and OE strains (Fig 5.14 B).

Figure 5.14: Majority of the reduced 5’ H2A.Z drop in 4KR and OE is independent of differences in gene activation. (A) Fractions of activation-independent genes in 4KR, defined as 4KR having similar expression levels to WT at both 0 and 30 minutes, within the 350 genes with reduced 5’ H2A.Z drop. The rest contains genes whose expression levels are affected at 0 minute of 4KR and gene whose induction at 30 minutes is affected. Fractions of activation-independent genes in OE were calculated similarly. (B) Overlap of activation-independent H2A.Z drop reduction between 4KR and OE. (C) Example of an activation-dependent. (D) Example of an activation-independent.
5.5.6 Reduced H2A.Z drop is accompanied by slower gene expression recovery

Having established that the reduced 5’ H2A.Z drop in 4KR and OE is largely independent of differences in gene activation, we next investigated whether the reduction in H2A.Z drop affects the recovery of gene expression from 30 to 120 minutes. To compare the recovery of gene expression in mutants with WT, we defined the term “expression recovery relative to WT” (ERRW) in mutant strains $ERRW(mut)$ as the difference between 120-to-30-minute fold-changes in mutant compared to that in WT or:

$$ERRW(mut) = \log_2\left(\frac{E(mut_{120})}{E(mut_{30})}\right) - \log_2\left(\frac{E(WT_{120})}{E(WT_{30})}\right)$$

where $E$ represents the gene expression in a strain at a time point. Since these genes are generally recovering from activation after 30 minutes, the log$_2$ fold-changes of expression between 30 and 120 minutes in WT are mostly negative (second term in formula). Therefore, positive ERRW means that the mutant is less negative than WT and therefore has reduced recovery of expression. Within the 228 genes, we observed significantly more genes with positive ERRW than negative ERRW in both 4KR (53 positive and 15 negative, Fig 5.15 B, left panel) and OE (101 positive and 29 negative, Fig 5.15 B, right panel). Both mutants have similar patterns of ERRW (Fig 5.15 A). In both mutants, the fractions of genes with reduced expression recovery are significantly higher in the set of 228 genes than in the genome (Fig 5.15 B), which also indicates the link between delayed H2A.Z drop and delayed gene recovery. To further demonstrate the significance, we divided the 228 genes into 5 quantiles according to fold-changes of expression between 30 and 120 minutes. In almost all quantiles, we observed significant less down-regulation in both mutants than in WT (Fig 5.15 C). As expected, we observed no difference in the magnitudes of initial activation at 30 minutes between mutants and WT (Fig 5.15 D).
Figure 5.15: Reduced drop of 5’ H2A.Z is accompanied by delayed recovery of gene expression. (A) Hierarchically clustered data of relative recovery of expression in 4KR and OE compared to WT, showing that the recovery of expression by 120 minutes tends to be reduced in both 4KR and OE. (B) Fractions of genes with reduced and increased expression recovery, defined as $ERRW > 0.1$ and $< -0.1$ respectively, within the 228 genes and genome. The 228 genes are more likely to have reduced expression recovery compared to the entire genome in both 4KR (Fisher’s Exact test $p$-value $= 1.8 \times 10^{-10}$) and in OE (Fisher’s Exact test $p$-value $= 8.8 \times 10^{-14}$). (C) Recovery of expression at 120 minutes are reduced in 4KR and OE compared to WT. The 228 genes with activation-independent reduced H2A.Z drop are divided into five quantiles. The fold-changes of expression between 30 minutes and 120 minutes in WT, 4KR and OE are plotted for each quantile. P-values of two-tailed t-test for each quantile between 4KR/OE and WT are: Quantile 1: 0.093, $1.1 \times 10^{-4}$; Quantile 2: $4.3 \times 10^{-4}$, $1.5 \times 10^{-5}$; Quantile 3: $5.8 \times 10^{-4}$, $4.8 \times 10^{-4}$; Quantile 4: 0.093, 0.23; Quantile 5: 0.19, 0.0013. (D) Induction at 30 minutes is hardly affected in 4KR and OE by reduced drop of 5’ H2A.Z.
5.5.7 Reduced 5’ H2A.Z drop affects highly responsive genes

The 4KR and OE mutants both show delayed H2A.Z drop at gene promoters, which has little effects on initial induction but affects the gene’s recovery from induction. These results lead us to investigate what is special about these 228 genes whose expression recovery is delayed. To answer this question, we compared the properties of these 228 genes with the 112 genes whose 5’ H2A.Z drop reduction is activation-dependent and the 677 genes whose 5’ H2A.Z drops are not reduced in mutants (Fig 5.16 A). At 0 minute, all 3 groups of genes are expressed at similar levels in all strains (Fig 5.16 B). Upon caffeine treatment, genes with reduced H2A.Z drop (both activation-dependent and independent) are induced to higher levels than genes whose 5’ H2A.Z drops are not reduced (Fig 5.16 B, C, blue and grey against black). Surprisingly, the activation-dependent genes are slightly more up-regulated than activation-independent ones at 30 minutes (Fig 5.16 C). From the previous chapter, we knew that TATA genes are generally up-regulated in caffeine response, we next checked the proportion of TATA genes in these three groups. Using the 1027 genes with decreased 30-minute 5’ H2A.Z as background, we observed underrepresented TATA genes with genes with no reduced H2A.Z drop (Fig 5.16 D, light grey bar). Between the two groups with reduced H2A.Z drop, activation-dependent ones are highly enriched in TATA genes whereas activation-independent ones are not enriched (Fig 5.16 D, dark grey and blue bars). These observations indicate that activation-independent reduction of 5’ H2A.Z drop in mutants tends to affect most highly induced genes and does not particularly favour TATA genes.
CHAPTER 5. DELAYED GENE RECOVERY IN H2A.Z MUTANTS

Figure 5.16: Highly induced genes tend to have reduced drop of 5’ H2A.Z. (A) Numbers of genes with activation-dependent, activation-independent reduced H2A.Z drop in 4KR and OE than in WT and genes without reduced H2A.Z drop. (B) Expression of these three categories in WT, 4KR and OE at 0, 30 and 120 minutes during caffeine response. (C) Expression of activation-independent and dependent reduced H2A.Z drop are more highly induced compared to genes whose 5’ H2A.Z are unchanged to WT. (D) Activation-dependent genes are enriched for TATA genes (Fisher’s Exact test, \( p-value = 4.9 \times 10^{-7} \)) whereas genes with no reduced H2A.Z drop are depleted for TATA genes (Fisher’s Exact test, \( p-value = 2.1 \times 10^{-6} \)). Activation-independent genes are not enriched for TATA genes.
5.6 Discussion

In this study, we first addressed the question why the cellular level of H2A.Z (i.e. the sum of chromatin-bound and cytoplasmic) is decreased during caffeine time course. From transcriptomic data, we observed that the expression of HTZ1 is not decreased but in fact slightly increased over the time course (Fig S5.1), which rules out the possibility of lower expression. To probe for H2A.Z levels in chromatin, we conducted ChIP-seq experiments for H2A.Z on *S. cerevisiae* cells treated by caffeine. Since we face a situation where the total level of ChIP signal is likely different between samples, which implies that conventional normalisation methods such as depth and quantile normalisation would not be suitable, we developed our own normalisation that is based on constant genomic backgrounds across all samples. We applied this normalisation on WT and 4KR samples and observed decreasing levels of chromatin-bound H2A.Z during caffeine time course, which is consistent with our western blot data of the cellular total.

5.6.1 Limitations and improvements for normalisation

The normalisation for OE samples was different because raw ChIP-seq data show less distinctive differences between the 5’ ends (high signals) and CDSs/IGRs (low signals) compared to WT or 4KR, indicating different background levels to WT or 4KR (Fig 5.4 B). We assumed this is due to higher levels of H2A.Z being deposited into CDSs rather than lower levels being deposited to 5’ ends. However, ChIP-seq alone cannot rule out the latter possibility. To further prove which one is the case, methods such as quantitative PCR would be required. However, our analyses made no direct comparison between OE and WT or 4KR at any time points. Instead, we focused on the change of H2A.Z levels within each individual mutant. Therefore, the differences in normalisation methods are unlikely to affect our conclusions.

A few recent studies have developed a new type of spike-in normalisation in both human (Bonhoure et al., 2014; Orlando et al., 2014; Grzybowski et al., 2015) and *S. cerevisiae* (Hu et al., 2015). In these methods, fragments of reference DNA are added to each sample in proportion to the number of cells. The reference DNA acts as a benchmark for normalisation such that the normalised data would indicate the average ChIP signal per cell and hence allows for direct comparisons between samples. Therefore, spike-in methods would improve the quality
of normalisation of our H2A.Z ChIP-seq especial for the OE samples.

5.6.2 Majority of 5’ H2A.Z drop can be explained by gene activation

Nevertheless, our normalised ChIP-seq agrees well with western blots, both of which show decreasing levels of chromatin-bound H2A.Z during the time course of caffeine response. By examining H2A.Z occupancies at the 5’ ends and inside CDSs of genes separately, we concluded that the drop is primarily contributed by the 5’ ends rather than in the CDSs of genes. The drop of H2A.Z levels at gene promoters could be the consequence of either nucleosome eviction or nucleosome-independent H2A.Z removal. Eviction of promoter nucleosome has been known to happen during gene activation in response to heat shock (Shivaswamy and Iyer, 2008), indicating that the drop of 5’ H2A.Z occupancy we observed contains contribution from nucleosome eviction. In fact, we estimated that over 90% of H2A.Z drop occurred at 30 minutes of caffeine response are found amongst activated genes. However, this is likely an overestimation since nucleosome-independent H2A.Z eviction also occurs at the promoter of activated genes.

5.6.3 Evidence for nucleosome-independent H2A.Z eviction

Our results also support the existence of nucleosome-independent H2A.Z eviction. We observed H2A.Z drop at promoters of genes whose expression is not changed during caffeine response and therefore whose promoter nucleosome is presumably not displaced (Fig 5.8 B). In consistence with this, we observed decreasing expression of Swr1 and increasing expression of Ino80 over the time course (Fig S5.1). These findings imply that H2A.Z may be actively removed from chromatin in a manner that is independent of nucleosome eviction from promoters. The nucleosome-independent H2A.Z eviction is likely also contributing towards the overall H2A.Z drop at gene promoters. However, we could not quantify the extent of nucleosome-independent H2A.Z eviction since its interlaced nature with nucleosome eviction. To further elucidate this, ChIP experiments on nucleosome occupancy and Swr1 or Ino80 binding over the caffeine time course would be helpful in understanding the extent of nucleosome-independent H2A.Z eviction.
5.6.4 Reduced 5’ H2A.Z drop is accompanied by delayed expression recovery

Next, we revealed a reduced drop of the 5’ occupancy of H2A.Z in both mutant strains compared to WT during caffeine response. The reduced drop could either be a consequence of less nucleosome eviction (i.e. nucleosome-dependent) or due to Swr1- or Ino80-related H2A.Z deposition or removal (i.e. nucleosome-independent). From this study alone, we could not distinguish the two possibilities. Nevertheless, we observed that most of the genes with reduced H2A.Z drop in mutant strains are independent of differences in gene activation (Fig 5.14 A), which are likely be nucleosome-independent. Interestingly, genes with activation-independent reduced H2A.Z drop do not show deficiencies in gene induction at 30 minutes but display slower recovery by 120 minutes. The fact that the initial induction is hardly affected implies that the RSC-mediated nucleosome displacement and the release of Pol II may not be affected by H2A.Z mutants, whereas the resetting of chromatin to ground state may be impeded.

It is notable that both 4KR and OE mutants share a common subset of genes whose 5’ H2A.Z decrease is reduced (Fig 5.12 B), despite the fact that their 0-minute H2A.Z abundances differ massively (Fig 5.1 B). However, one similarity between the two mutants is that their percentage H2A.Z drop relative to time-0 are both reduced compared to WT. The rate of H2A.Z drop may be a reflection of the combination of Swr1 and Ino80 activities at gene promoters. It would be interesting to investigate whether the Swr1 or Ino80 complex plays a role in regulating the restoration of chromatin organisation after gene induction. We further demonstrated that these genes are amongst the most highly induced gene during the first 30 minutes of caffeine response. In contrast to the set of all highly induced genes, these genes are not enriched for TATA-containing genes, which indicates that H2A.Z could be affecting TFIID-regulated rather than SAGA-regulated genes.

5.6.5 Similarity to \textit{ino80}Δ mutant

The build-up of H2A.Z at gene promoters can cause genomic instability (Papamichos-Chronakis et al., 2011). The \textit{ino80}Δ mutant results in accumulated chromatin-bound H2A.Z and prolonged activation of several stress-response genes (Klopf et al., 2009; Papamichos-Chronakis et al., 2011), which is similar to our results.
Klopf et al. (2009) further demonstrated that the prolonged activation is independent of nucleosome displacement but due to delayed reassembly of promoter nucleosomes after activation. Moreover, Papamichos-Chronakis et al. (2011) also demonstrated that having low acetylatability of H2A.Z is essential for \( \text{ino80}\Delta \) phenotype since \( \text{ino80}\Delta \) can be recovered by deleting the histone deacetylase Hda1 or introducing the lysine-to-glutamine mutant of H2A.Z to mimic the acetylated form. Since the acetylation status during caffeine response was still unknown, we could not be certain whether our observation is driven by the same mechanism of \( \text{ino80}\Delta \). Additional data on the acetylation status would be useful in determining whether there is a larger proportion of unacetylated H2A.Z accumulated in mutant strains compared to WT.

5.7 Summary

In this chapter, we first developed a normalisation method that is suitable for our data and revealed a global drop of chromatin-bound H2A.Z during caffeine response. The drop is primarily contributed by the decrease of promoter H2A.Z amongst activated genes. The H2A.Z drop is reduced or delayed in both 4KR and OE mutants, which results in higher promoter H2A.Z levels in mutants than in wild type. Interestingly, the increased promoter H2A.Z does not seem to affect initial gene induction but is accompanied by slower expression recovery. These observations share certain degrees of similarity to the \( \text{ino80}\Delta \) mutant, indicating the possibility that H2A.Z may be important for the nucleosome reassembly after gene activation. However, it remains to be established whether our observation and \( \text{ino80}\Delta \) are driven by the same mechanism.
5.8 Supplementary Figure

Figure S5.1: Gene expression of Swr1, Ino80 and Htz1 in caffeine time course.
Figure S5.2: Changes of H2A.Z in CDS are uncorrelated with changes in sense or AS transcripts.
Chapter 6
Discussion and Outlook

The importance of H2A.Z in transcriptional regulation has been long established. H2A.Z is preferably found at promoters of protein-coding genes and the depletion of H2A.Z affects gene activation in both S. cerevisiae and higher eukaryotes. However, previous studies have failed to characterise the molecular mechanisms of how H2A.Z contributes towards transcriptional regulation (Mizuguchi et al., 2004). Previous studies of H2A.Z tend to be focusing on protein-coding genes because non-coding transcription were originally considered as transcriptional noises. In the past decade, it has become clear that non-coding transcripts in fact make up a significant proportion of most eukaryotic transcriptomes. More importantly, more and more ncRNA transcripts have been shown to regulate the transcription of protein-coding genes, which leads to a whole new view of how genes are regulated in eukaryotic genomes. In our study, we specifically addressed the role of H2A.Z in regulating non-coding transcripts.

6.1 Promoter H2A.Z as a Positive Regulator of Transcription

One of our main questions is whether H2A.Z is a transcriptional activator or repressor? While previous works supported H2A.Z being an activator in higher eukaryotes (Hardy et al., 2009), why there is a lack of correlation between H2A.Z and transcripts in S. cerevisiae remains unexplained. By examining the coding and non-coding transcripts, we established that H2A.Z is a general mark for both sense and AS promoters. We further dissected the direct and indirect effects of
H2A.Z on transcripts and demonstrated that H2A.Z is likely a direct activator of transcription. The H2A.Z that activates AS transcripts is also a repressor of the cognate sense transcripts. Therefore, the overall effect of H2A.Z depends on its distribution at the 5’ and 3’ ends of the gene and whether there is an involved AS transcript.

Here, an outstanding question is: at what stage does H2A.Z contribute towards the transcriptional activation? During transcriptional activation, nucleosomes are displaced from the promoter region by the RSC complex, forming the NDR. During the same process, H2A.Z and acetylation are added to the +1 and -1 nucleosomes around the NDR. One possibility is that H2A.Z is required for the chromatin reorganisation during transcriptional activation. While the exact mechanism is still unknown, this idea is consistent with the previous observation that depletion of H2A.Z affects the induction stress-response genes in *S. cerevisiae* (Santisteban et al., 2000; Adam et al., 2001) and causes developmental defects in higher eukaryotes (Creyghton et al., 2008; Subramanian et al., 2013; Hu et al., 2013). An alternative explanation is that H2A.Z is required to maintain the steady-state level of gene expression. Since in vitro studies have reported destabilising roles of H2A.Z on nucleosome, incorporation of H2A.Z may contribute towards maintaining the loosely packed promoter structure for the access of other factors (Suto et al., 2000). Moreover, the +1 nucleosome downstream of the NDR acts as a high-energy barrier to the transcription machinery (Rhee and Pugh, 2012; Nock et al., 2012). In Drosophila, the H2A.Z located at the +1 nucleosome shows anti-correlation with Pol II stalling, indicating that H2A.Z may reduce the energy required to overcome the nucleosome barrier preventing Pol II from entering transcription elongation (Weber et al., 2014). Further studies on the mechanism of how H2A.Z contribute towards gene activation would be helpful in understanding the transcriptional regulation in eukaryotic genomes.

### 6.2 Asymmetry of Bidirectional Promoters

We also observed that H2A.Z is located at bidirectional promoters where pairs of sense and AS transcripts are produced in opposite directions. An interesting question here is: what controls the directionality of transcription? On post-transcriptional level, it has been established that transcripts produced in the AS direction are actively degraded by the poly-adenylation site facilitated early
termination in higher eukaryotes (Ntini et al., 2013) and NNS-mediated termination in *S. cerevisiae* (Schulz et al., 2013). However, it remains an open question whether promoter directionality is controlled at chromatin level. In recent years, more evidence are pointing towards an asymmetrical architecture of bidirectional promoter. Two separate PICs, one in each direction, were shown to be assembled at bidirectional promoters (Murray et al., 2012; Rhee and Pugh, 2012) but it was unclear whether the two PICs are independently regulated. Ramachandran et al. (2015) revealed an asymmetrical distribution of nucleosomes at nucleosome dyads in *S. cerevisiae* and asymmetrical nucleosome dyads are preferably associated with the RSC complex. Similar results were also reported by histone ChIP (Rhee et al., 2014). However, neither Rhee et al. (2014) or Ramachandran et al. (2015) found correlations between the direction of asymmetry and the direction of transcription, indicating that other factors may be involved.

In our analyses, we observed that depletion of H2A.Z has much stronger effects on the AS transcript than on the sense transcript produced from the same bidirectional promoter. We also observed a stronger correlation between H2A.Z and AS transcripts than with sense transcripts. These results are consistent with the asymmetric model of promoter architecture. Further studies on the distribution of other histone marks at bidirectional promoters would be helpful in understanding whether promoter asymmetry is linked with transcriptional directionality.

### 6.3 Non-coding Transcript as a Regulatory Signal for Distal Targets

During stress response, we observed increased levels of lncRNA produced from promoters of induced genes in the AS direction. These promoter-associated AS transcripts represent a large class of lncRNA in *S. cerevisiae* but their functions are largely unknown. In this study, we identified 20 clusters of sense:AS:sense transcripts (i.e. AS transcript between a tandem gene pair) where all 3 transcripts are coregulated. The coregulation seems to be dependent on the AS transcript since tandem gene pairs without intergenic AS transcripts are not coregulated. It was known that certain adjacent gene pairs in *S. cerevisiae* are co-expressed and have functional similarities (Cohen et al., 2000; Ebisuya et al., 2008). Our results support the claim that promoter-associated lncRNAs may act as a medium to
spread regulatory signals to nearby genes (Xu et al., 2011). The exact mechanism of how lncRNA regulates the upstream gene remains to be established. Based on previously characterised mechanisms, we speculate that lncRNAs may interact with other chromatin remodelers to alter the chromatin organisation of their target gene. To further test for this hypothesis, histone modification profiles (especially acetylations and H3K36me3) would be useful for elucidating the mechanism. Moreover, it is reasonable speculate whether the lncRNA-mediated regulation of distal gene is related to distal regulatory elements such as enhancers in higher eukaryotes. Further studies would benefit from comparative genomics for these promoter sequences and coregulated genes across different species.

6.4 H2A.Z May Affect Chromatin Reset after Induction

We also established that the decrease of chromatin-bound H2A.Z abundance during stress response is predominantly caused by gene activation, either due to nucleosome loss or nucleosome-independent H2A.Z eviction. Interestingly, both 4KR and OE mutants are showing less degrees of H2A.Z drop compared to WT. The reduced drop of H2A.Z is largely independent of differences in gene activation in mutants. Surprisingly, we found that the reduced H2A.Z drop does not affect the initial activation but affects the recovery of gene expression from activation. Delayed gene recovery has also been observed in the ino80Δ mutant, where H2A.Z is also accumulated at promoters (Klopf et al., 2009; Papamichos-Chronakis et al., 2011). In ino80Δ, delayed gene recovery is related to delayed nucleosome reassembly after gene activation (Klopf et al., 2009). Here, an important question is whether the effect of 4KR and OE mutants is also related to delayed nucleosome reassembly. To answer this question, genome-wide nucleosome occupancy data over stress response time course would be helpful. Furthermore, Papamichos-Chronakis et al. (2011) demonstrated that the phenotypes of ino80Δ are associated with the deacetylated form of H2A.Z and can be recovered by hda1Δ or 4KQ mutants. It would be interesting to further characterise how H2A.Z acetylation contribute towards the reset of promoter nucleosome after gene induction.
6.5 Concluding Remarks

Histone H2A.Z is an important transcriptional regulator with debatable activating or repressing functions. Our results supported H2A.Z being a general activator of both sense and AS transcripts. Its activating effect on AS transcripts would negatively affect the sense cognate, which explains some of the previously observed repressive effects. We also confirmed that promoters are intrinsically bidirectional and promoter-associated AS transcription is a wide-spread phenomenon in eukaryotic genomes. The AS transcripts could serve as a means to spread the regulatory signal to adjacent or distal genes. Taken together, our results highlighted the interleaved nature of eukaryotic gene regulatory systems. Therefore, it is essential to take holistic approaches (e.g. taking into account both coding and non-coding transcripts) in studying the transcriptional regulation in eukaryotic genomes.


Appendix: Published Paper
H2A.Z marks antisense promoters and has positive effects on antisense transcript levels in budding yeast

Muxin Gu, Yanin Naiyachit, Thomas J Wood and Catherine B Millar*

Abstract

Background: The histone variant H2A.Z, which has been reported to have both activating and repressive effects on gene expression, is known to occupy nucleosomes at the 5' ends of protein-coding genes.

Results: We now find that H2A.Z is also significantly enriched in gene coding regions and at the 3' ends of genes in budding yeast, where it co-localises with histone marks associated with active promoters. By comparing H2A.Z binding to global gene expression in budding yeast strains engineered so that normally unstable transcripts are abundant, we show that H2A.Z is required for normal levels of antisense transcripts as well as sense ones. High levels of H2A.Z at antisense promoters are associated with decreased antisense transcript levels when H2A.Z is deleted, indicating that H2A.Z has an activating effect on antisense transcripts. Decreases in antisense transcripts affected by H2A.Z are accompanied by increased levels of paired sense transcripts.

Conclusions: The effect of H2A.Z on protein coding gene expression is a reflection of its importance for normal levels of both sense and antisense transcripts.

Keywords: Antisense transcript, Chromatin, H2A.Z, Histone variant, Htz1

Background

Chromatin components are key regulators of global gene expression. H2A.Z is a highly conserved histone variant that replaces H2A in a subset of nucleosomes, most prominently those that flank the transcription start sites (TSSs) of protein-coding genes (reviewed in [1]). TSS-adjacent H2A.Z localization is suggestive of a role in transcriptional regulation and indeed H2A.Z has been implicated in gene regulation in multiple organisms [2-5]. As H2A.Z is essential for normal development [6-8] and its over-expression is associated with poor patient prognosis in human cancers [9,10], it is important to understand how H2A.Z contributes to the regulation of gene expression.

Previous studies linking H2A.Z to transcription have primarily focused on protein-coding genes. It has recently become clear that eukaryotic transcriptomes are a complex mixture of coding and non-coding transcripts, with many transcripts being rapidly turned over by RNA-processing machinery such as the exosome (reviewed in [11,12]). In S. cerevisiae, such cryptic unstable transcripts often originate from the 5' nucleosome-depleted-region (NDR) of a downstream tandemly arranged gene and can be detected in the absence of the exosome catalytic component Rrp6 [13,14]. In this study, we used yeast strains lacking Rrp6 to compare H2A.Z occupancy to global transcription and found that H2A.Z occupies the 3' ends of protein-coding genes in addition to its well-known enrichment at their 5' ends. H2A.Z is co-localised with other active histone modifications at the 3' end of genes, at sites that match the start sites of non-coding transcripts transcribed in the antisense orientation relative to the sense protein-coding genes. The deletion of H2A.Z results in the down-regulation of antisense transcripts that normally have H2A.Z in their promoters. This novel association between H2A.Z and antisense transcripts differs fundamentally from the previously described role of H2A.Z in suppressing antisense transcripts in fission yeast [15]. Our findings indicate that H2A.Z is a general marker of TSSs and suggest that some apparently indirect effects of H2A.Z deletion on the expression of protein-coding genes whose
promoters do not contain H2A.Z are mediated through effects on an antisense transcript.

Results and discussion

H2A.Z is significantly enriched at the 3′ ends of genes

Numerous ChIP-chip and ChIP-seq studies, in various organisms and cell types, have revealed high enrichment of H2A.Z at the promoters of protein-coding genes. In this study we used highly specific antibodies ([16]; Additional file 1: Figure S1A,B) to endogenous untagged budding yeast H2A.Z (Htz1) in ChIP-seq experiments to assess whether there were other, less obvious, patterns of Htz1 enrichment that had hitherto gone un-noticed. We found that besides the high enrichment of Htz1 at the 5′ ends of protein coding genes, Htz1 is also enriched at the 3′ ends and inside coding sequences (CDSs; Figure 1A, B). ChIP-seq analysis in an htz1Δ strain confirmed that these signals are all specific to Htz1 (Additional file 1: Figure

---

**Figure 1** About one-third of Htz1 is localised outside of TSSs. A. Average profile of Htz1 at protein-coding genes, with an average transcript depicted by a light blue box. Genes were aligned according to their TSSs or transcription end sites (TESs) and the average levels of Htz1 (reads per million sequenced reads per gene) were calculated for each base pair. For relative alignments of coding regions (CDS; region >300 bp down/up-stream of TSS/TES), all transcripts were stretched or compressed to a constant length and then the average Htz1 level was found at each relative position. B. The percentages of Htz1 signal located at the 5′ end (−300 to +300 relative to TSS), CDS, and 3′ end (−300 to +300 relative to TES) of protein-coding genes are shown, along with enrichment in other intergenic regions (IGR; > 300 bp away from a protein coding gene). The darkness of the colour for each category reflects the length-normalised density of Htz1 signal in reads per kilobase per million mapped reads (RPKM). C. Hierarchical clustering of Htz1 binding profiles, highlighting the clusters of genes with Htz1 enrichment at TSSs, CDSs and TESs. A 25-dimensional vector was generated for each transcript, comprising nine 50-bp windows corresponding to the −2, −1 and +1 nucleosomes around the TSS (5′ −2, 5′ −1 & 5′ +1), 10 windows for the coding sequence (CDS) and six 50-bp windows for the nucleosomes flanking the TES (3′ −1; 3′ +1). Clustering was done using Euclidean distances between the vectors. Data in this figure were generated from one biological replicate, but are essentially identical to a second wild-type biological replicate (Additional file 5: Table S1).
S1C, D). Quantification of ChIP-seq signals revealed that about one-third of the Htz1 signal is found outside of the 5′ ends of protein-coding genes, at 3′ ends, intergenic and intragenic regions (Figure 1B). Notably, the 3′ ends of genes harbour nearly 15% of the Htz1 signal and the density of Htz1 occupancy (i.e. length-normalised occupancy) is higher at 3′ ends than at intra- and inter-genic regions (Figure 1B). Out of 5143 protein coding genes, 1025 have a high level of Htz1 at their 3′ end, while 2506 have high Htz1 levels at their 5′ ends. Clustering of the Htz1 distributions along genes revealed that some have Htz1 only at their 5′ or 3′ ends while others have Htz1 at both 5′ and 3′ ends (Figure 1C). Although prior studies have primarily focused on the TSS-adjacent occupancy of H2A.Z, Coleman-Derr et al. have observed enrichment of H2A.Z in gene bodies and 3′ ends in Arabidopsis thaliana [17], in agreement with our findings in S. cerevisiae.

Peaks of H2A.Z at the 3′ ends of genes correlate with other active histone marks

As antisense transcription is widespread in budding yeast [13, 14], we hypothesised that the prominent peaks of Htz1 at the 3′ ends of genes could reflect the TSSs of antisense transcripts initiating in these regions. If this were the case, the 3′ peaks of Htz1 would be expected to co-localise with histone modifications that are normally present at actively transcribed promoters. We examined levels of histone modifications that are known to be enriched in the promoter regions of active genes [18] and observed that the 3′ regions with high Htz1 occupancy are significantly enriched for active histone marks, including H3K4me3, H3K18ac, H3K9ac and H4K12ac (Figure 2). The correlations we observe between 3′ Htz1 occupancy and other marks of active promoters are consistent with these 3′ peaks of Htz1 potentially being promoters of antisense transcripts.

H2A.Z at the 3′ end of genes marks the start of antisense transcripts

To directly test whether 3′ peaks of Htz1 correspond to the start sites of antisense transcripts, we next compared Htz1 occupancy to strand-specific transcript data generated from a strain lacking the exosomal RNA-degrading enzyme, Rrp6. We focused on the Htz1 signal located upstream rather than downstream of the transcription end site (TES) in order to avoid Htz1 signal that is potentially derived from the 5′ promoter of the adjacent gene in the compact S. cerevisiae genome. We observed a total number of 1076 antisense transcripts in the rrp6Δ strain and the Htz1 peaks at the 3′ ends of genes frequently correspond to the 5′ ends of antisense transcripts (see Figure 3A for examples). Over 50% of genes with Htz1 occupancy in their 3′ end have a detectable antisense transcript in the rrp6Δ strain, while only 5% of genes without 3′ Htz1 have an antisense transcript (Figure 3B). The frequency of antisense transcripts in 3′ regions that are enriched for Htz1 is significantly higher than expected (p = 1.7 × 10^{-187}) and, conversely, the frequency is significantly lower than expected (p = 3.6 × 10^{-184}) in 3′ regions lacking Htz1 (Figure 3C). Randomisation of the Htz1 peaks at the 3′ ends of all genes generated a distribution of peaks coinciding with antisense transcripts that is markedly lower than the actual number we observe, with a false discovery rate close to zero (Figure 3D). We have focused on peaks at the 3′ end of genes because of this striking overlap but we also observed that a lower percentage of coding region Htz1 peaks overlap with the start sites of antisense transcripts (Additional file 2: Figure S2B). Our findings implicate Htz1 as a general marker of TSSs, for both sense and antisense transcripts. As at the 5′ ends of protein-coding genes, where Htz1 is widespread but not at every gene promoter (Figure 3A; [16, 21-24]), Htz1 is not a ubiquitous feature of the 5′ ends of antisense transcripts. This is comparable to other chromatin components that have been implicated in regulating antisense transcript levels, including the ATP-dependent remodeler Isw2 [25], the histone deacetylase complexes Rpd3S [26] and Set3C [27], and methylation of H3K4 [28], which also mark or regulate a sub-set of antisense transcripts.

H2A.Z is important for antisense transcript levels

To test whether Htz1 binding at the 3′ ends of genes contributes to regulation of antisense transcripts, we examined the antisense transcriptomes in rrp6Δ and rrp6Δhtz1Δ cells. To establish which transcripts are likely to be regulated directly by Htz1, we compared changes in transcript levels in the absence of Htz1 to Htz1 occupancy (Figure 4A). This analysis revealed that the majority of affected antisense transcripts are enriched for Htz1, consistent with a direct effect for Htz1 in regulating these transcripts although indirect effects are also formally possible. Notably, down-regulated antisense transcripts have significantly higher Htz1 occupancy at their 5′ ends (i.e. the 3′ ends of genes) while promoters of up-regulated antisense transcripts are less enriched for Htz1 (Figure 4B). Quantification of up- and down-regulated antisense transcripts revealed that the down-regulation of antisense transcripts occurs significantly more often than expected for genes with 3′ Htz1 occupancy (p = 1.0 × 10^{-31}; Figure 4C). In addition, the occupancy of Htz1 at the 3′ ends of genes is positively correlated with antisense transcript levels in the wild-type strain (Figure 4D). Together, these data are consistent with a predominantly activating effect for Htz1 at the promoters of antisense transcripts. While we also observe a predominantly activating effect for Htz1 at the 5′ ends of genes on sense transcripts, the effect is weaker than for antisense transcripts, as the p-values are markedly higher than at the 3′ ends (Additional file 3: Figure S3A, B, C).
Histone modifications known to occupy active promoters include H3K4me3 [19], H3K18ac [20], H3K9ac and H4K12ac [18]. H4K4me3 and H3K18ac data are from the whole genome; H3K9ac and H4K12ac are restricted to chromosome 3.

A, C, E, G. Profiles of histone modifications at the 3' ends of genes with 3' Htz1 enrichment (black lines) or without Htz1 enrichment (grey lines). All four active histone marks are enriched upstream of the TESs of genes that have high 3' Htz1 occupancy.

B, D, F, H. Boxplots of the distributions of histone modification levels in genes with high 3' Htz1 enrichment (3' Htz1; black boxes), without Htz1 enrichment (no 3' Htz1; light grey boxes) or with intermediate levels of 3' Htz1 (Int.; medium grey boxes). The number of genes in each category is indicated. H3K4me3 ($p = 4.0 \times 10^{-29}$), H3K18ac ($p = 8.0 \times 10^{-57}$), H3K9ac ($p = 6.4 \times 10^{-3}$) and H4K12ac ($p = 1.2 \times 10^{-2}$) are significantly higher on genes with high 3' Htz1. *$p \leq 0.05$, ****$p \leq 0.0001$. The $p$-values were obtained using two-tailed $t$-tests.

Figure 2 Htz1 at 3' ends of genes co-localises with active histone modifications. Histone modifications known to occupy active promoters include H3K4me3 [19], H3K18ac [20], H3K9ac and H4K12ac [18]. H4K4me3 and H3K18ac data are from the whole genome; H3K9ac and H4K12ac are restricted to chromosome 3. A, C, E, G. Profiles of histone modifications at the 3' ends of genes with 3' Htz1 enrichment (black lines) or without Htz1 enrichment (grey lines). All four active histone marks are enriched upstream of the TESs of genes that have high 3' Htz1 occupancy. B, D, F, H. Boxplots of the distributions of histone modification levels in genes with high 3' Htz1 enrichment (3' Htz1; black boxes), without Htz1 enrichment (no 3' Htz1; light grey boxes) or with intermediate levels of 3' Htz1 (Int.; medium grey boxes). The number of genes in each category is indicated. H3K4me3 ($p = 4.0 \times 10^{-29}$), H3K18ac ($p = 8.0 \times 10^{-57}$), H3K9ac ($p = 6.4 \times 10^{-3}$) and H4K12ac ($p = 1.2 \times 10^{-2}$) are significantly higher on genes with high 3' Htz1. *$p \leq 0.05$, ****$p \leq 0.0001$. The $p$-values were obtained using two-tailed $t$-tests.
Figure 3 (See legend on next page.)
Antisense transcripts that depend on H2A.Z are primarily transcribed from tandemly arranged genes

H2A.Z has previously been implicated in the regulation of antisense transcript levels in the fission yeast S. pombe, although in that case H2A.Z suppresses antisense transcript levels specifically at convergently transcribed genes by preventing transcriptional read-through [15]. As H2A.Z was not observed at the 3’ ends of genes in that study and had the opposite effect on antisense transcript levels, this previously described role of H2A.Z

Figure 3 Htz1 at 3’ ends marks the start of antisense transcripts. A. Transcripts derived from the + and − strands are shown, along with Htz1 occupancy over the same regions for the protein coding genes YCR059C, YBR128C, YBR096W, YBR019C and YBR020W. Antisense (AS) transcripts are coloured dark blue and sense (S) transcripts light blue. Htz1 ChIP peaks at the 3’ ends of protein-coding transcripts are co-incident with the 5’ ends of antisense transcripts, as indicated by the arrows. B. Fraction of genes with (top) or without (bottom) 3’ Htz1 that have associated antisense transcripts. 537 (52%) out of 1025 genes that have a high level of 3’ Htz1 are associated with antisense transcripts whereas only 155 (9%) out of 3010 genes with no 3’ Htz1 have antisense transcripts. C. Number of 3’ end regions associated with antisense transcripts. Out of the 1025 genes whose 3’ ends are occupied by high levels of Htz1, 537 are associated with antisense transcripts, which is significantly higher than the expected 185 (**p < 0.0001 (1.7 x 10^{-15}), Fisher’s exact test), whereas out of the 3010 genes whose 3’ ends are depleted of Htz1 only 155 are associated with antisense transcripts, which is significantly lower than the expected 542 (**p < 0.0001 (3.6 x 10^{-18}), Fisher’s exact test). D. Comparison of the number of 3’ Htz1 peaks associated with antisense transcripts (green line) to the distribution of random 3’ regions (black bars) that co-localise with antisense transcripts. 987 regions of 150 bp were drawn randomly from the 3’ ends of genes and the number co-localising with antisense transcripts was calculated. This randomisation was performed 1000 times to produce the histogram showing the distribution of random peaks that co-localise with antisense transcripts. The actual association of Htz1 with antisense transcripts is highly significant (p = 0).

Figure 4 Htz1 affects antisense transcript levels. A. Comparison of differential antisense (AS) transcript levels in rp6Δhtz1ΔΔ versus rp6Δ to Htz1 levels at the 3’ ends of genes. Each transcript is shown as an open circle, with its 3’ Htz1 level measured by ChIP-seq being the y-value and its fold change of expression in the rp6Δhtz1ΔΔ strain shown as its x-value. Significantly up- and down-regulated transcripts are coloured in blue and magenta respectively. B. Boxplots of the distributions of 3’ Htz1 levels for down- (n = 255) and up- (n = 169) regulated antisense transcripts show that down-regulated antisense transcripts are significantly enriched for Htz1 (**p < 0.0001 (2.9 x 10^{-21}), two-tailed t-test) compared to transcripts whose expression doesn’t change (n = 3019). C. Actual (solid bars) and expected (hatched bars) number of up-/down-regulated antisense transcripts with and without 3’ Htz1. Down-regulated antisense transcripts with 3’ Htz1 are significantly more numerous than expected (**p < 0.0001 (1.0 x 10^{-21}); Fisher’s exact test), whereas those without 3’ Htz1 are significantly fewer than expected (**p < 0.0001 (1.6 x 10^{-11}); Fisher’s exact test). D. Enrichment of Htz1 at the 3’ end of genes positively correlates with the level of the associated antisense transcript. Genes were classified into bins of seven quantiles according to 3’ Htz1 level. The distribution of antisense transcript levels are plotted for each bin, arranged from low 3’ Htz1 (left) to high 3’ Htz1 (right).
appears mechanistically different from what we have observed in budding yeast. Nonetheless, we checked whether there was a similar tendency of Htz1 to regulate antisense transcripts at convergent genes in budding yeast. We divided all intergenic regions into 4 classes depending on whether their flanking protein-coding genes were tandemly or convergently arranged and on the distance between the protein coding genes (Figure 5A). “Close” genes are separated by <300 bp, while “far” genes have >300 bp between their TSSs/transcription end sites (TESs). We found that most of the 3’ Htz1 signal is found in tandemly arranged genes rather than convergently arranged genes, and is particularly enriched in the “tandem close” category (Figure 5B, C; Additional file 4: Figure S4). To further investigate the relationship between gene organisation and the effect of Htz1 on antisense levels, we examined the number of genes in each category whose antisense transcripts are down-regulated in the absence of Htz1 (Figure 5D). This analysis also demonstrates a significant enrichment in the tandem close category ($p = 3.9 \times 10^{-20}$), indicating that Htz1 primarily affects antisense transcripts that are initiated from the promoters of tandemly arranged genes. The role of H2A.Z at the TSS of antisense transcripts in S. cerevisiae is therefore distinct from its previously described function in suppressing read-through antisense transcripts at convergent genes in S. pombe. This may reflect a difference in transcriptome organisation between S. cerevisiae and S. pombe, as the majority of antisense transcripts in S. pombe are generated by transcriptional read-through at convergent genes [29] while most antisense transcripts are derived from tandem genes in S. cerevisiae [13,14].

**Regulation of antisense transcripts by H2A.Z can affect sense transcript levels**

Htz1 has previously been implicated in regulating sense transcripts and we now show that it affects antisense transcripts in budding yeast. Levels of sense and antisense transcripts at individual genes are generally anti-correlated (reviewed in [30,31]), which raises the question of whether

---

**Figure 5 Htz1 predominantly affects tandemly arranged sense/antisense genes.**

A. Genomic arrangement of the 4 gene categories, with sense transcripts depicted as light blue boxes on the + or – strand and antisense transcripts shown as dark blue boxes. Tandem or convergent refers to the sense transcripts. Close genes have <300 bp between them, while the distance is >300 bp for the “far” genes. The coloured box for each category is the key for B, C & D. B. Htz1 levels at the 3’ ends of genes, aligned by TESs and coloured according to A. The arrow indicates the 3’ peak upstream of the TES that we have focused on. Convergent genes have less Htz1 associated with their 3’ ends, with “tandem close” genes having the highest Htz1 levels. C. Quantification of the 3’ Htz1 signal shows that the majority of 3’ Htz1 is found at tandem genes. D. Htz1 affects a significant number of tandem close antisense transcripts. The number of antisense transcripts down-regulated in rrp6Δhtz1Δ was compared to the number in that category in the genome for each of the 4 gene arrangements. The $p$ values are derived from Fisher’s exact tests.
the regulation of antisense transcripts by Htz1 affects sense transcript levels. We took genes with antisense transcripts and 3’ Htz1 occupancy, and examined the changes in sense transcript levels for genes where the antisense transcript either didn’t change, went up, or went down. Genes whose antisense transcript decreases in the absence of Htz1 show a corresponding increase in sense transcript levels (p = 0.02, Figure 6A). Conversely, the expression levels of sense transcripts corresponding to up-regulated antisense transcripts were decreased (p = 0.05, Figure 6A).

It is challenging to dissect the dependencies of sense/antisense transcript levels due to their inter-connected nature but the fact that Htz1 is more enriched at down-regulated sense and antisense transcripts leads us to propose a model in which the effects of Htz1 on down-regulated transcripts are direct and that up-regulated transcript levels generally result from an effect on the corresponding sense/antisense. To test this, we divided genes into 4 groups based on their Htz1-enrichment patterns (Figure 6B). Group 1 genes have Htz1 at their 5’ ends only and, based on our observations of sense transcript levels (Additional file 3: Figure S3), we would predict that levels of group 1 transcripts should decrease slightly when HTZ1 is deleted. Although the median log (2) expression is lower than 0 in this group, it is not significantly lowered. The behaviour of Group 2 genes is unpredictable as they have Htz1 at both 5’ and 3’ ends, and changes at Group 3 genes (having no Htz1 at either 5’ or 3’ end) can be attributed to indirect effects. However, Group 4 genes (having Htz1 only at 3’ ends) are predicted to have up-regulated sense transcripts and indeed we find that Group 4 genes are significantly up-regulated (p = 7.5 x 10^{-4}; Figure 5C). These findings are consistent with Htz1 having an effect on antisense transcripts that consequently affects sense transcript levels and may indicate that Htz1 is more important for the activation of antisense transcripts than sense transcripts.

**Conclusions**

We find in this study that a significant proportion of Htz1 is located at the 5’ ends of antisense transcripts, and that Htz1 is required for normal levels of these non-coding transcripts in addition to its known role in regulating protein-coding genes. Comparison of strand-specific RNA-seq and ChIP-seq data shows that Htz1 occupancy has a predominantly activating effect on the promoters of antisense and sense transcripts. Previous work has described both activating and repressing roles for H2A.Z at individual genes in both yeast and mammalian cells (reviewed in [32]). While some genes that are up-regulated in htz1Δ have Htz1 at their promoters, these are relatively rare and most promoters of up-regulated transcripts have low Htz1 occupancy, arguing against a direct repressive effect of Htz1 on most transcripts, as previously observed by Li et al. [24] for protein-coding genes. However, up-regulation of protein coding genes that lack Htz1 in their promoters is not due to completely indirect effects, at least in some cases.

**Figure 6** The effect of Htz1 on antisense transcript levels impacts sense transcripts. A. Genes with 3’ Htz1 enrichment and an antisense transcript were divided into three groups based on whether their antisense transcript levels were unchanged (n = 794), up- (n = 28) or down- (n = 159) regulated in r6pΔhtz1Δ. The boxplots show changes in sense (S) transcript levels for these groups. Down-regulated antisense transcripts have a significantly higher fold change in sense expression (*p ≤ 0.05 (0.02), two-tailed t-test) and conversely sense transcript levels are significantly decreased when antisense transcripts are up-regulated (†p ≤ 0.05 (0.02), two-tailed t-test). B. Classification of genes with both sense and antisense transcripts according to 5’ and 3’ enrichment of Htz1 was performed using thresholds for 5’ and 3’ enrichment based on average Htz1 occupancies at the 5’ and 3’ ends. Class 1 (n = 110) have only 5’ enrichment; class 2 (n = 237) have 5’ and 3’ enrichment; class 3 (n = 75) are not enriched for Htz1 at either 5’ or 3’ ends; and class 4 (n = 54) have Htz1 only at the 3’ end. C. Distributions of sense transcript fold changes in the r6pΔhtz1Δ strain for each of the classes of genes illustrated in B. Sense transcripts are significantly up-regulated in group 4 genes that have Htz1 at their 3’ ends only (***p ≤ 0.001 (7.5 x 10^{-4}), two-tailed t-test).
but is mediated by Htz1 affecting levels of antisense transcripts generated from the 3′ ends of these genes. Mechanistically, some changes of sense transcript levels in htz1Δ strains have previously been ascribed to the aberrant activity of the SWR-C when Htz1 is absent [33]. We could not test whether Swr1 is responsible for the changes in AS levels in rrp6Δhtz1Δ because we were unable to generate strains triply mutated for rrp6Δhtz1Δ and swr1Δ.

The importance of H2A.Z for antisense transcription may explain some of its apparently conflicting effects on the expression of protein coding genes and highlights the need to study all transcripts derived from a locus in order to fully understand how genes are regulated.

**Methods**

**Yeast genetics and molecular biology**

Yeast strains were created using standard methods and are described in Additional file 5: Table S2. Cells for RNA extraction and ChiP were in harvested from log-phase cultures growing in SD medium. Total RNA was purified using the Ribopure Purification Kit (Life Technologies) and contaminating genomic DNA was removed by DNaseI digestion. Htz1 ChiP was performed essentially as described previously [34], using affinity purified custom αHtz1 (a660) antibodies [16] in a protocol optimised for maximal Htz1 recovery and extensive chromatin fragmentation by sonication. RNA and purified ChiP DNA were amplified for sequencing using the Illumina TruSeq stranded mRNA sample preparation kit or the Illumina TruSeq ChiP Sample Prep Kit. Libraries were sequenced on either an Illumina GAIIx or HiSeq 2000. ChiP-seq samples had an average of 12 million uniquely mapping reads, and RNA-seq samples had an average of 35 million reads (Additional file 5: Table S3). Data were generated from two independent biological replicates for each strain in each analysis (ChiP-seq/RNA-seq) and correlations between biological replicates were high in all cases (Additional file 5: Table S1 A, B).

**Sequence mapping**

Sequence reads were mapped to the S. cerevisiae genome assembly sacCer1 using Bowtie version 0.12.9 [35], allowing up to 2 mismatches and no ambiguously mapped reads. Genomic coordinates for protein-coding transcripts were obtained from Xu et al. [14].

**ChiP-seq data processing**

Background signal was set to 1.2 standard deviations above the mean of the ChiPinput ratios and then subtracted from the ChiP-seq signal. Where background was higher than the ChiP-seq signal, the value was set to zero. The final ChiP signal at each base pair was normalized to the number of reads per million mapped reads.

**Strand-specific RNA-seq data processing**

RNA-seq mapped reads were segregated into + and − strands. Normalisation was performed such that the total amount of sense-strand RNA was adjusted to 10⁸ arbitrary units and antisense RNA-seq levels were adjusted by the same factor. Transcript levels were then normalized per kilobase of transcript length.

**Data analysis**

Regions that are 150 bp downstream of the TSS and upstream of the TES were used to quantify the 5′ and 3′-end enrichment of Htz1 respectively. Htz1 peaks were associated with a transcript if they had downstream RNA signals >3-fold higher than upstream signals. The nature of the associated transcripts (i.e. sense or antisense) was determined by comparison to ORF-Ts [14]. For differential gene expression, the number of RNA-seq reads were calculated for each sense and antisense transcript [14] and merged in to one file. Differentially expressed genes were identified by DESeq2 [36], using corrected p-value < 0.05. Convergent overlapping transcripts were excluded from the quantification of antisense transcript levels to avoid potential confusion between sense and antisense transcripts in these cases.

**Availability of supporting data**

The data sets supporting the results of this article are available in the GEO repository, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54105.

**Additional files**

*Additional file 1: Figure S1.* 3′ end enrichment of Htz1 is not an artefact of antibody cross-reactivity. The a660 antibodies used in this study are highly specific for Htz1 in ChiP [1] but had not been tested in ChiP-seq experiments. To rule out any possibility that the enrichment of Htz1 in the CDS and 3′ ends of genes was artefactual, we performed ChiP-seq using this antibody with ChiP extracts generated from an htz1Δ strain. The pattern of enrichment in the htz1Δ strain was completely different from the wild-type (WT) Htz1 pattern (Pearson correlation, r = -0.03) and lacked the peaks seen in the WT sample. A, Normalized ChiP signal from htz1Δ (upper) and WT (lower) cells along chromosome 1, with the coordinates displayed at the bottom. Only the coding regions of a few very highly expressed genes had detectable signal in the htz1Δ sample and enrichment at these ‘hyper-ChIPable’ loci is a known artefact in ChiP that is not antibody-related [2]. This confirms that the anti- Htz1 antibody is not precipitating DNA in an Htz1-independent manner. The labeled peaks correspond to: 1, YAL038W/CDC19; 2, pLUG6; 3, YAL003W/ZFP1; 4, YAL001W/CDF; 5, tL(CAA)A; 6, tS(AGA)A. B, Zoomed in views of peaks 1 (YAL038W/CDC19 left) and peaks 5 and 6 (tL(CAA)A, tS(AGA)A right) on chromosome 1. ORF and "SOS Other" annotations on the + and − strands are indicated at the right, and key features are labelled for orientation purposes. C, Htz1 enrichment at three of the genes shown in Figure 3A in the WT and htz1Δ strains. There is no enrichment detectable in the htz1Δ sample, either at the 5′ or 3′ ends of the genes. D, The average signal at the 3′ ends of genes is zero in the htz1Δ sample and is significantly lower than the WT sample (p = 0; t-test).

*Additional file 2: Figure S2.* Coding region Htz1 peaks co-localise with AS transcripts. A, Examples of Htz1 peaks in the mid-coding region overlapping with the start of AS transcripts, as indicated by the arrows.
Colour coding is as for Figure 3A. B. The fraction of Htz1-enriched regions in CDSs associated with AS transcripts is 10% (307 out of 3044). We speculate that other Htz1 peaks in the CDS are also transcription-associated but that these transcripts are not detectable either because they are derived from the sense strand or because they are unstable even in the rnpΔA strain. C. Comparison of the number of CDS Htz1 peaks associated with AS transcripts (green line) to the distribution of random CDS regions (black bars) that co-localise with AS transcripts. 3044 random windows were drawn from a total set of 38599 windows and randomisation was repeated 100 times to generate the histogram. The association of CDS Htz1 peaks with AS transcripts, although lower than the association of 3’ Htz1 peaks with AS transcripts (Figure 3D), is highly significant (p = 1.6 x 10⁻87; Fisher’s exact test).

Additional file 3: Figure S3. Efficacy of Htz1 on sense transcript levels. A. Comparison of differential sense transcript levels in rnpΔA/ΔA versus rnpΔA to Htz1 levels at the 5’ ends of genes. Each gene is shown as an open circle, with its 5’ Htz1 level measured by ChIP-seq being the y-value and its fold change of expression in the rnpΔA/ΔA strain shown as its x-value. Significantly up- and down-regulated transcripts are coloured in light blue and pink respectively. B. Boxplots of the distributions of 5’ Htz1 levels for down- (p = 2.67) and up- (p = 2.55) regulated sense transcripts, show that down-regulated S transcripts are significantly enriched for Htz1 (***p ≤ 0.001; 3.6 x 10⁻5; two-tailed t-test) compared to transcripts whose expression doesn’t change (n = 2921); C. Actual (solid bars) and expected (hatched bars) numbers of up- and down-regulated antisense transcripts with and without 5’ Htz1. Down-regulated sense transcripts with 5’ Htz1 are significantly more numerous than expected (***p ≤ 0.001; 4.2 x 10^-4); Fisher’s exact test). D. There is no obvious correlation between enrichment of Htz1 at the 5’ end of genes and level of the associated sense transcript. Genes were classified into bins of seven quantiles according to 5’ Htz1 level and the distribution of sense transcript levels are plotted for each bin.

Additional file 4: Figure S4. Tandem-clove genes have higher Htz1 at 3’ ends than genes of other arrangements. Distributions of levels of Htz1 at 3’ end for tandem-close, tandem-far, convergent-close and convergent-far genes. The amount of Htz1 within a 150bp window upstream of TEs is displayed. Two-tailed t-tests show that the level of 3’ Htz1 at tandem-close genes is significantly higher relative to levels in the other categories of genes (**p ≤ 0.0001; tandem far p = 2.5 x 10⁻¹⁸; convergent close p = 1.5 x 10⁻¹⁷; convergent far p = 1.8 x 10⁻¹⁸).

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit