Novel Insights into Chromatin Structure and Gene Regulation through Integrative Analysis of High Throughput Genomics Data

Thong T. Nguyen

A thesis submitted to the

School of Mathematics, Statistics and Applied Mathematics
National University of Ireland, Galway

In fulfilment of the requirements for the degree of
Doctor of Philosophy

Under the supervision of
Professor Cathal Seoighe

September 2013

Abstract

Genetics and epigenetics research has evolved dramatically over the last decade, owing to rapid developments in high-throughput genomics techniques. Analysis of the resulting quantities of data requires advanced computational strategies. In this thesis, we use computational and statistical methods to tackle biological questions relating to two major research topics. First, we carried out integrative analysis of next generation sequencing data to answer questions regarding chromatin biology and epigenetics. Second, we performed genome-wide analysis of the effects of genetic variants on gene expression, focussing on two key processes: mRNA decay and mRNA translation.

The first question investigated genome-wide distribution of the histone variant H2AX, a key factor in the DNA damage response pathway. We assessed the genomic landscape of H2AX in human U2OS cells using H2AX ChIP-seq data. Strikingly, we found that H2AX was enriched in heterochromatic regions. Heterochromatin has previously been shown to be refractive to damage signalling through H2AX phosphorylation and, consequently, we hypothesized that the greater abundance of H2AX in heterochromatin helps to ensure sufficient H2AX phosphorylation to signal DNA damage events.

We next turned to characterizing the chromatin organization of the genomic regions that are distal (distal junction – DJ) and proximal (proximal junction – PJ) to human nucleolar organizer regions (NORs). Because they are absent from the reference genome assembly, these regions represent a major gap in our understanding of the epigenetic configuration of the human genome. An integrative analysis of ChIP-seq, RNA-seq, FAIRE-seq and DNase-seq data, generated by the ENCODE consortium, revealed that the DJ resembles euchromatic regions and, surprisingly, harbours transcripts that are transcribed by RNA polymerase II. Laboratory experiments showed that the DJ is localized to the periphery of the nucleolus, where it anchors the ribosomal DNA arrays. This study sheds new light on the role of NORs in nucleolar formation and function, and enables further investigation of the link between nucleoli and human pathologies.

The focus then shifts to studying genetic variation in gene expression. First, we set out to identify *trans*-acting genetic variants that influence RNA stability. We demonstrate that perturbation of RNA stabilization is detectable from mRNA expression data. Using the mRNA expression data generated from 726 HapMap3 samples, we calculated the relative expression of long-lived RNAs versus short-lived RNAs for each sample (referred to as RNA stability score or RS-score). Treating RS-score as a quantitative trait, we applied genome-wide association and identified a SNP, rs6137010, with which it is strongly associated in two Asian populations: Han Chinese from Beijing (CHB) and Japanese from Tokyo (JPT). This SNP is a *cis*-eQTL for *SNRPB* (a core component of the spliceosome) in CHB and JPT. Thus, we propose that the association between this SNP and inter-individual variation in RS-score is likely mediated by changes in *SNRPB* expression levels.

The final question investigated the effects of genetic variants on mRNA translation. We developed a computational pipeline to identify genetic variants that influence allele-specific mRNA translation rate (AST). Analysis of allele-specific events is severely biased by the fact that short read sequences favour mapping to the reference allele. Thus, our pipeline first constructs a haplotype-resolved genome for a given cell-type by making use of high-throughput sequencing data that are publicly available for that cell-type. Both RNA-seq and Ribo-seq data are then mapped to the resulting haplotype-resolved genome in order to identify genes that show evidence of AST. Applying this pipeline for the datasets from HeLa cells, we found 171 protein-coding genes that are associated with AST. Inspection of heterozygous SNPs located in the AST genes revealed two interesting mutations, within the 5'UTR of two genes: *ATP5H* and *SLCO4A1*, that appear to inhibit translation initiation of these genes.

To sum up, this thesis presents novel computational strategies for integrative analysis of large volumes of high-throughput genomics data. By addressing biological questions in the areas of chromatin biology and gene regulation, this thesis yields key insights into the DNA damage response, the role of NORs in nucleolar formation and function, and the effects of genetic variants on mRNA stability and mRNA translation.

Acknowledgments

First and foremost I would like to thank my supervisor Professor Cathal Seoighe for supporting me during these past four years. Cathal is one of the smartest people I know and he has been a truly outstanding mentor. I have been absolutely lucky and privileged to have had the opportunity to work with him, and I would be more than delighted to collaborate with him in the future.

There have been several key individuals who have been involved in the work of this thesis. The group of Dr. Andrew Flaus have contributed to generating the H2AX ChIP-seq data and the other wet-lab work, as presented in Chapter 2. I gratefully appreciate the contribution of Andrew in providing invaluable guidance and ideas for the project. Andrew is also my co-supervisor and I would like to thank him for giving me many good advice. The work presented in Chapter 3 would not have been possible without the contribution from Professor Brian McStay's group and Dr. Austen Ganley's group. I would like to thank the members in these two groups for providing incredible efforts for the project. I also especially thank Brian for giving several months of wet-lab experience.

I thank my fellow bioinformatics PhD students in NUI Galway for showing interest in my PhD research and for exchanging experiences and ideas. I also would like to acknowledge the other postgraduates in the School and Maths, who have always been kind and friendly and provided such a nice working environment.

Lastly, I am deeply grateful for the love and encouragement of my family while I was studying in Ireland. I thank my parents for raising me with a love of science and supporting me in all my pursuits.

Thong T. Nguyen

September 2013

Contents

List of Figures

List of Tables

Chapter 1: Introduction

Recent developments in high-throughput techniques have provided scientists with great opportunities for exploring many aspects of gene regulation and chromatin organization. The mechanisms underlying how gene expression is regulated at individual steps from DNA to protein are becoming better understood. It has also now become more clear how DNA is packaged into chromatin, thereby controlling DNA accessibility. The whole-genome landscape of chromatin structure has been completely revealed in various species and cell types. More importantly, chromatin structure is a major determinant of gene expression variation.

This thesis is organized into six chapters. A review of the relevant biological background in chromatin biology and gene regulation is presented in the following sections of this introduction chapter. Analyses of next generation sequencing data to address questions regarding chromatin biology and epigenetics are carried out in Chapters 2 and 3. Chapter 2 presents a genome-wide analysis of the relative enrichments of the histone variant H2AX between euchromatin and heterochromatin. In Chapter 3, we profile chromatin structure and transcriptome of an rDNA-adjacent region (the distal junction) in order to identify regulatory elements responsible for NOR activity. The focus then turns to genetic variation in gene expression in the next two chapters. Chapter 4 presents an integrative analysis of mRNA expression data and half-life data to pinpoint genetic variants that affect transcriptome-wide RNA stability. In Chapter 5, we develop a novel computational pipeline for the identification of genetic variants that influence the rate of mRNA translation. Finally, the main conclusions of the thesis and suggestions for future investigations are discussed in Chapter 6.

1.1 Gene regulation

Eukaryotic genomes contain many thousands of genes but only a fraction of them are expressed in a specific cell type. Moreover, a given gene can express differently in response to changes in its environment. The expression level of a gene is regulated at multiple steps in the pathway from DNA to mRNA, and mRNA to protein (Figure 1.1). Here we focus on three main steps, *viz* transcription, RNA degradation and translation.

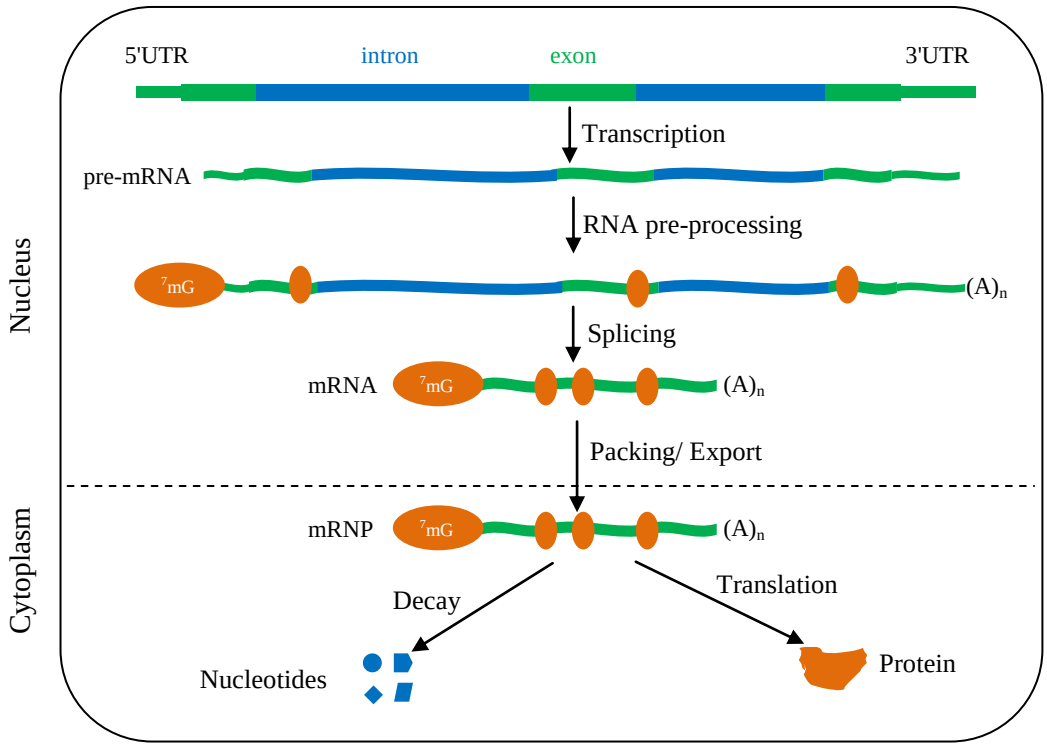


Figure 1.1: Multiple steps of gene regulation. Orange ovals represent RNA-binding proteins. $(A)_n$ represents the poly(A) tail of mRNA. ^{7}mG represents 7-methylguanylate cap.

1.1.1 Transcription

Transcription is the first step of gene expression, in which DNA is copied into RNA by RNA polymerase. There are four types of RNA polymerase in mammals, each is associated with a distinct subset of RNA. For example, RNA polymerase I (Pol I) transcribes ribosomal DNA genes [1], while RNA polymerase II (Pol II) is responsible for synthesis of messenger RNAs (mRNAs) and most small nuclear RNAs (snRNAs) and microRNAs (miRNAs) [2]. Transcription consists of three

main stages: initiation, elongation and termination, but is mainly regulated at the initiation stage [3]. The mechanisms underlying how Pol II is delivered and assembled at the promoter provide insights into understanding transcriptional regulation.

Transcription factors (TFs) have a major impact on gene expression regulation [4]. In humans, they account for around 10% of genes [5]. Signals of TF-binding around the TSS are predictive of gene expression [6]. Moreover, differential binding of TFs is indicative of differential expression of genes [7]. TFs are classified into two groups: general transcription factors (GTFs) and sequence-specific TFs. GTFs bind the promoter of a large fraction of genes and play a key role in the assembly of pre-initiation complexes (PICs) [8]. The sequence-specific TFs, which can act as activators or repressors, target subsets of genes and result in distinct patterns of expression of target genes [9]. Below, we describe basic mechanisms regarding how transcription pre-initiation complexes (PICs) are formed and how transcription is activated and repressed.

PIC assembly by GTFs

The role of transcriptional initiation is to form PICs and thereby recruiting Pol II to the promoter [10]. In eukaryotes, PICs consist of Pol II and at least these general transcription factors (GTFs): TFIIA, TFIIB, TFIID (or TBP), TFIIE, TFIIF, TFIIH and TFIIK [10]. The assembly of PICs requires the ordered recruitment of these factors (Figure 1.2). The TATA-binding protein (TBP) is first recruited to the promoter as part of the TFIID complex, which contains TBP-associated factors (TAFs) [11]. These factors then facilitate the recruitment of Pol II and other GTFs to the core promoter in order to start transcription [12,13]. The PIC is formed within the 5' nucleosome-free region (5'NFR) that bridges two well-positioned nucleosomes around the TSS [14].

When TBP is not part of the TFIID complex, it tends to bind to the TATA-containing promoters through the guidance of the SAGE complex [15]. When TBP is part of TFIID, it often binds to TATA-less promoters [16]. In yeast, the majority of genes have TATA-less promoters. Also the mechanism of PIC assembly differs

considerably between the two class of promoters (TATA-less and TATA-containing) [16,17]. Proper assembly of PICs at the promoter requires interaction between the GTFs (Table 1.1).

A recent study found that the organization of the PIC is conserved between yeast and humans [18]. This study established a consolidated view of transcription initiation in humans [18]. 160,000 transcription initiation complexes were found in the human genome but 95% of these associate with non-coding and non-polyadenylated RNAs [18].

Table 1.1: Roles of general transcription factors

Name	Roles	REFs
TFIID-TBP	recognizes TATA box in the promoter.	[16]
TFIID-TAFs	recognizes other DNA sequences around the TSS and regulates DNA-binding of TBP.	[15]
TFIIB	clamps TBP to DNA and is a linchpin between TBP and Pol II.	[19]
TFIIF	promotes the interaction of Pol II with TFIIB, assists recruiting TFIIE, and promotes downstream elongation events.	[20]
TFIIE	enhances DNA strand separation by Pol II at the TSS, and promotes the activity of TFIIH.	[21]
TFIIH	unwinds DNA at the TSS, phosphorylates Ser5 of the C-terminal domain (CTD) of Pol II; release Pol II from the promoter.	[22]

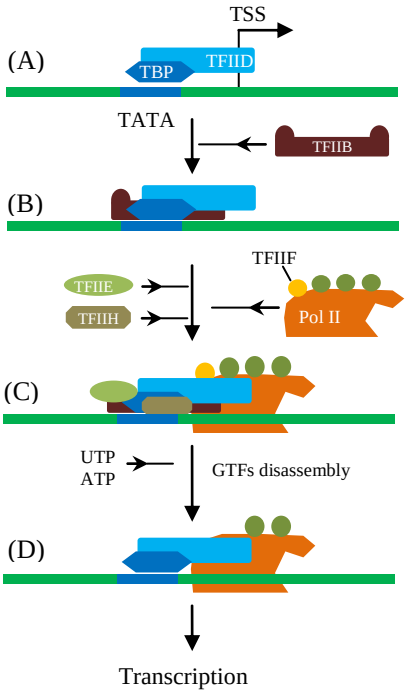


Figure 1.2: Transcription initiation by RNA polymerase II and general transcription factors (GTFs). (A) The TATA box, located ~25 bp away from the transcription start site, is the binding target of TBP/TFIID, which then enables the adjacent binding of TFIIB (B). The rest of GTFs and Pol II are assembled at the promoter (C). (D) TFIIH next uses ATP to pry apart the DNA double helix and locally exposes the template strand. After Pol II is properly assembled and the promoter is at the ready state, GTFs are released and Pol II starts scanning the template strand.

Transcriptional activation

DNA in eukaryotic cells is packaged into chromatin and therefore the transcription initiation requires more proteins than it does on purified DNA [3] (Figure 1.3). First, transcriptional activators bind to enhancer regions and promote the assembly of Pol II and GTFs at the promoter [3]. A typical activator protein consists of two distinct domains [3]. One domain usually contains a structural motif that can recognize a specific DNA pattern. The other domain promotes transcription initiation. Some activators bind directly to GTFs, thereby stimulating their assembly at the promoter [23]. Others interact with mediator and attract it to DNA where it appears to participate in the assembly of PIC [23] (Figure 1.3). Because DNA is occupied by nucleosomes, activators can modify local chromatin structure to make the underlying DNA more accessible [7,24]. We will discuss this in more details in the section 1.3.

Transcriptional repression

Gene repressors silence transcription in various ways [3]. Repressors can compete with activators for binding to the same regulatory DNA sequence. Repressors can also bind to activation domains of activators, thus preventing activators from facilitating the assembly of PIC. Repressors can even bind several GTFs and block the assembly of PIC. In addition, repressors can attract chromatin remodelling complexes to make the promoter DNA less accessible, thereby inhibiting transcription.

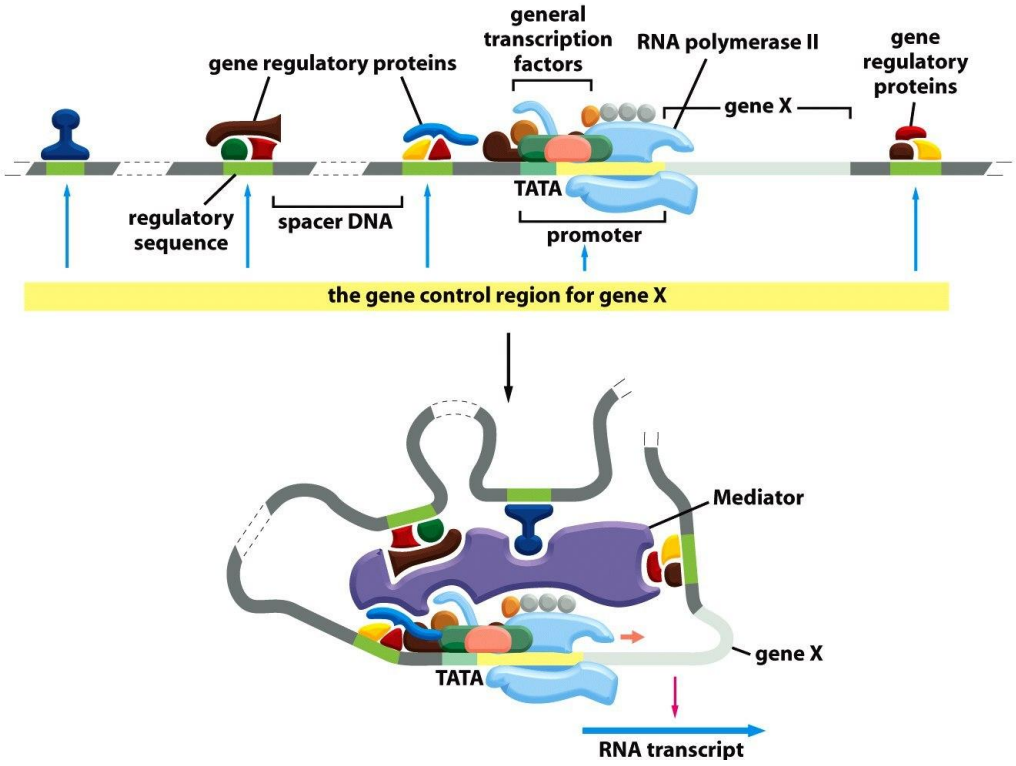


Figure 1.3: Transcription initiation by RNA polymerase II requires activator, mediator and chromatin modifying proteins. Gene regulatory proteins bind to regulatory sequences – in green (e.g. enhancers). Some general transcription factors (GTFs) can recognize and bind TATA-containing elements and then help assemble RNA polymerase II at the promoter. The enhancer can interact with the promoter through a protein complex called mediator. Reproduced with permission of GARLAND SCIENCE: Albert *et al.* [3], copyright 2008.

In conclusion, gene regulatory proteins can switch the transcription of a gene on and off. These proteins usually bind to specific DNA sequences within the promoter or enhancer, which can initiate transcription independent of their distance and orientation with respect to the promoters. When binding to the enhancer, these proteins can contact Pol II by looping out of the intervening DNA. Both activators and repressors can act by mechanisms involving altering chromatin structure and controlling the assembly of the PIC and mediator at the promoter.

1.1.2 RNA degradation

The abundance of protein produced from any given mRNA depends not only on the rate of mRNA translation but also on the rate of mRNA synthesis and decay. Levels of mRNA are regulated at various steps such as: transcription, splicing, post-transcriptional modifications and mRNA export. Interestingly, mRNA decay can be influenced by these same processes. mRNA decay is important for eliminating potentially toxic proteins and for changing protein abundance. Here we focus on two major pathways of mRNA decay including nonsense-mediated mRNA decay (NMD) and AU-rich element (ARE)-mediated mRNA decay.

1.1.2.1 Nonsense-mediated mRNA decay

NMD is triggered when the mRNA contains a premature translation-termination codon (PTC) [25]. NMD is a translation-coupled mechanism, therefore regulators of translation are considered as regulators of NMD as well. NMD usually targets newly synthesized mRNAs at the first round of translation and is known to serve as an mRNA-surveillance mechanism to prevent the synthesis of toxic proteins [26]. A NMD pathway requires many protein complexes, particularly three core *trans*-acting factors: UPF1, UPF2 and UPF3 [27,28]. These core factors are thought to form a trimeric complex that links PTC to mRNA degradation [29]. Additional well-characterized proteins participating in the NMD machinery are SMG1, SMG5, SMG6 and SMG7. The phosphatidylinositol 3-kinase-related kinase SMG1 is responsible for phosphorylating UPF1, thereby allowing UPF1 to interact with mRNA degradation factors [30]. The SMG5-SMG7 complex or SMG6 can bind to the phosphorylated UPF1, leading to degradation of the NMD target [28].

NMD is associated with pre-mRNA splicing, as in many cases the mRNA level is considerably decreased when the PTC is located upstream of an intron [31]. The link between NMD and splicing is mediated by the exon junction complex (EJC) [25]. The EJC is deposited around 24 nt upstream of exon-exon junctions as a consequence of splicing [32]. Four proteins: eIF4AIII, MAGOH, Y14 and MLN51 form the core component of the EJC [32] (Figure 1.4). During the pioneer round of translation, the ribosome scans the mRNA and displaces all EJCs deposited along the mRNA [28]. However the presence of a PTC located at least 50-55 nt upstream of an EJC can stop the ribosome scanning [28]. The NMD factor UPF1 is then

recruited by the EJC to the stalled ribosome to form a surveillance complex that triggers both translation inhibition and mRNA decay [33] (Figure 1.4).

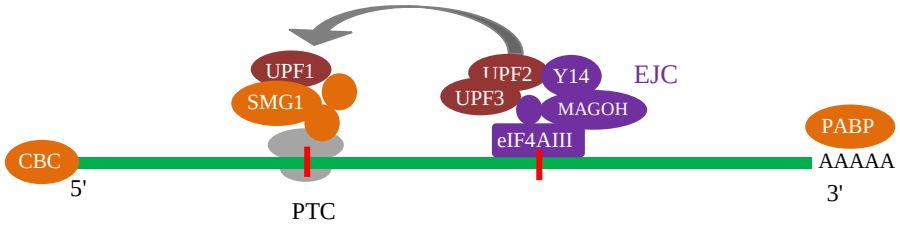


Figure 1.4: Model of nonsense-mediated mRNA decay (NMD) through the exon junction complex (EJC). CBC is the cap-binding complex. Gray complex represents the stalled ribosome at the premature translation-termination codon (PTC). The EJC is represented in violet. NMD factors are represented in dark red ovals.

1.1.2.2 ARE-mediated mRNA decay

AU-rich elements (AREs) are short regulatory sequences located in the 3' untranslated region (3'UTR) of a large proportion of short-lived mRNAs [28]. AREs represent the most common *cis*-acting determinant of RNA stability in mammalian cells. The canonical AREs have one or more copies of the AUUUA pattern that are generally located within an U-rich region. ARE-mediated mRNA decay begins with either synchronous or distributive poly(A) shortening, and is then followed by cleavage of the mRNA body from both ends by exonucleases [34,35].

ARE-binding proteins (ARE-BPs) recognize and bind AREs and participate in rapid deadenylation and degradation of the target mRNAs. ARE-BPs can be involved in both stabilizing and destabilizing of the target mRNAs [28] (Table 1.2). ARE-BPs are regulated by phosphatises, kinases and arginine methyltransferase. Phosphorylation plays a key role in controlling function and binding of ARE-BPs [28]. TTP, one of the best studies ARE-BPs, is phosphorylated at multiple sites, which recruit the 14-3-3 adapter proteins [36]. The 14-3-3 proteins then interfere with the TTP-mediated recruitment of deadenylases to initiate the mRNA decay (Figure 1.5). Another example of ARE-BP, HuR, is also phosphorylated at multiple sites but then functions as a stabilizer of mRNA [37].

Table 1.2: Summary of key ARE-binding proteins

ARE-BP	Function	Modifiers	REFs
TTP (ZFP36 ring finger protein)	Destabilizing	MK2, PP2A, MKP1	[38,39]
BRF1 (RNA polymerase III transcription initiation factor 90 kDa subunit)	Destabilizing	PKB, MK2, PP2A	[40]
BRF2 (RNA polymerase III transcription initiation factor 50 kDa subunit)	Destabilizing	PKB, MK2, PP2A	[41]
AUF1 (AU-rich element RNA binding protein 1)	Destabilizing/ Stabilizing	NPM-ALK	[42]
KSRP (KH-type splicing regulatory protein)	Destabilizing	p38, PI3K	[43]
HuR (Hu antigen R)	Stabilizing	p38, PKCα, PKCδ	[37]

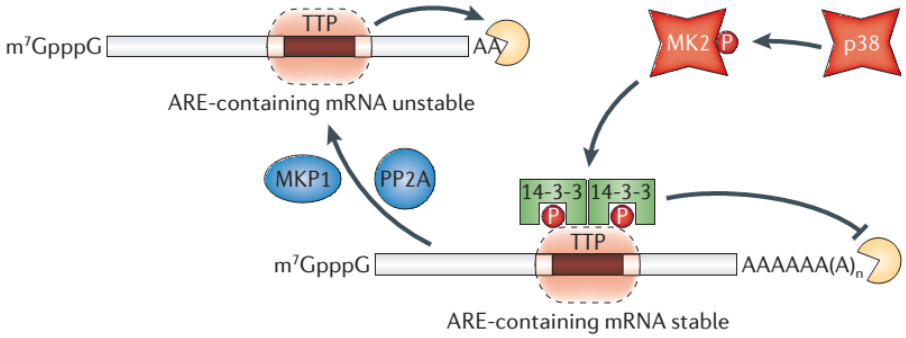


Figure 1.5: Kinase- and phosphatase-mediated regulation of TTP during ARE-mediated mRNA decay. TTP can bind to the ARE of mRNA. In the stable mRNA, TTP is phosphorylated by MK2 and thereby provides binding sites for 13-3-3 proteins that inhibit the interaction of TTP with deadenylases. In the unstable mRNA, phosphates in the TTP are removed by MKP1 or PP2A, leading to the interaction of TTP with deadenylases. Reprinted by permission from Macmillan Publishers Ltd: Schoenberg *et al.* [28], copyright 2011.

To sum up, mRNA decay contributes to steady-stage gene expression level. There are many mRNA decay pathways that involve various factors and processes.

The mRNA decay can be regulated through the interaction of RNA-binding proteins with *cis*-acting elements embedded within the mRNA itself.

1.1.3 Translation

Translation of mRNA is the process in which cellular ribosomes synthesize proteins. Translation consists of four phases: initiation, elongation, translocation and termination, but is mainly regulated at the initiation phase. Therefore, mechanisms of translation initiation provide insights into the regulation of protein synthesis.

The mechanism of translation initiation is divided into several steps [44] (Figure 1.6) and requires many factors – some core factors are described in Table 1.3. Briefly, translation initiation is responsible for forming the 80S ribosomes, in which the start codon is base-paired with the anticodon of the initiator tRNA Met-tRNA$^{Met}_{i}$ (which helps initiate protein synthesis). The initiation begins with the formation of 48S initiation complexes that are then joined with 60S subunits to form 80S ribosomes. The 48S complex is formed by a scanning mechanism. The 43S preinitiation complex, which includes a 40S subunit, eIF2–GTP–Met-tRNA$_i$ complex, eIF1, eIF1A and eIF3, binds the capped 5' proximal region of mRNAs and participates in unwinding the secondary structure of the 5' terminal of the mRNA. Next, the 43S complex scans the 5' UTR to identify the start codon and forms the 48S complex. Then, eIF5 and eIF5B help displace other core initiation factors (eIFs) and join a 60S subunit to the 48S complex at the start codon. Translation is now ready to start.

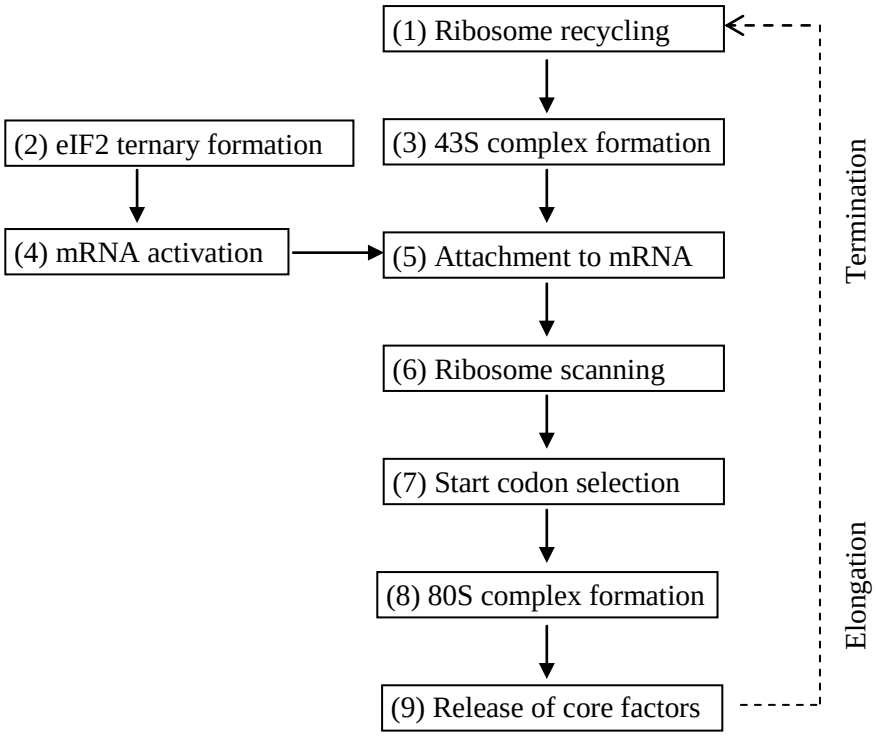


Figure 1.6: The flowchart of steps involved in a canonical pathway of eukaryotic translation initiation. The ribosome recycling step (1) yields separated 60S and 40S subunits, and leads to the formation of the 80S complex, in which Met-tRNAi is based-paired with the start codon in the P-site. The next steps are: eIF2–GTP–Met-$tRNA_i$ complex formation (2); 43S preinitiation complex formation, including a 40S subunit, eIF1, eIF1A, eIF3 and eIF2–GTP–Met-$tRNA_i$ (3); mRNA activation, during which the mRNA cap-proximal region is unwound in an ATP-dependent manner by eIF4F with eIF4B (4); attachment of the 43S complex to this mRNA region (5); scanning of the 43S complex along the 5′ UTR (6); selection of the start (initiation) codon and formation of 48S initiation complex, which results in displacement of eIF1 to enable eIF5-mediated hydrolysis of eIF2-bound GTP (7); joining of 60S subunits to 48S complexes and displacement of other core initiation factors (8); and finally hydrolysis by eIF5B-bound GTP and release of eIF5B and eIF1A (9). This flowchart is based on [44].

Table 1.3: Summary of eukaryotic core initiation factors

Factor	Function	REFs
eIF1	guarantees the fidelity of start codon selection; stimulates scanning of ribosomes and binding of eIF2–GTP–Met-tRNAi to 40S subunits; and prevents premature eIF5-induced hydrolysis of eIF2-bound GTP and Pi release.	[45]
eIF2	forms an eIF2–GTP–Met-tRNAi complex that mediates ribosomal recruitment of Met-tRNAi.	[46]
eIF2B	stimulates GDP–GTP exchange on eIF2.	[44]
eIF3	binds 40S subunits, eIF1, eIF4G and eIF5; promotes binding of eIF2–GTP–Met-tRNAi to 40S subunits; helps attach 43S complexes to mRNA; and possesses ribosome dissociation and anti-association activities.	[47]
eIF1A	enhances binding of eIF2–GTP–Met-tRNAi to 40S subunits; stimulates ribosomal scanning at 5'UTR.	[48]
eIF4E	binds the m7GpppG 5′ terminal.	[49]
eIF4A	serves as DEAD-box RNA helicase to unwind RNA.	[50,51]
eIF4G	binds eIF4E, eIF4A, eIF3, PABP, SLIP1 and mRNA and participates in helicase activity.	[50,51]
eIF4F	is a cap-binding complex that includes eIF4E, eIF4A and eIF4G.	[52]
eIF4B	enhances the helicase activity.	[50,51]
eIF4H	enhances the helicase activity and is homologous to a part of eIF4B.	[50,51]
eIF5	is a GTPase-activating protein that helps the large ribosomal subunit associate with the small subunit.	[44]
eIF5B	participates in assembly of the full ribosome.	[44]

1.2 Chromatin organization

Genetic information of all organisms is stored within DNA. In humans, the DNA content of a single copy of the genome is more than two metres long, in total, and consists of about 3 billion bases. To fit into the small volume of the cell, genomic DNA is tightly packed. Chromatin is the form in which DNA is packaged within the nucleus of the cell [3,53]. Chromatin is organized at multiple levels. The basic unit of chromatin is the nucleosome, which is generally composed of an octamer of four core histones (H3, H4, H2A, H2B) around which 147 to 150 base pairs of DNA are wrapped. Chromatin appears in the "Beads-on-a-String" conformation, where nucleosomes are the beads, and the DNA is the string. Nucleosome positioning and chromatin modifications are major determinants of chromatin organization and functions.

1.2.1 Nucleosome positioning

The nucleosome contains a histone core around which DNA is wrapped. Each histone core consists of two copies of each of the histones H2A, H2B, H3 and H4 (Figure 1.7). Genes that are transcriptionally active or poised normally contain a different version of the histone core in which canonical histones H2A and H3 are replaced with their variants H2A.Z and H3.3, respectively [54,55]. About 147 bp of DNA coils 1.65 times around the histone core. The polypeptide chains of the histone tails are subject to covalent modifications, which strongly influence the global architecture of chromatin. Nucleosomes are arranged as a linear array along the DNA sequence. Beyond the nucleosome core is the linker histone H1 that compacts the nucleosomes into higher-order structure (Figure 1.7).

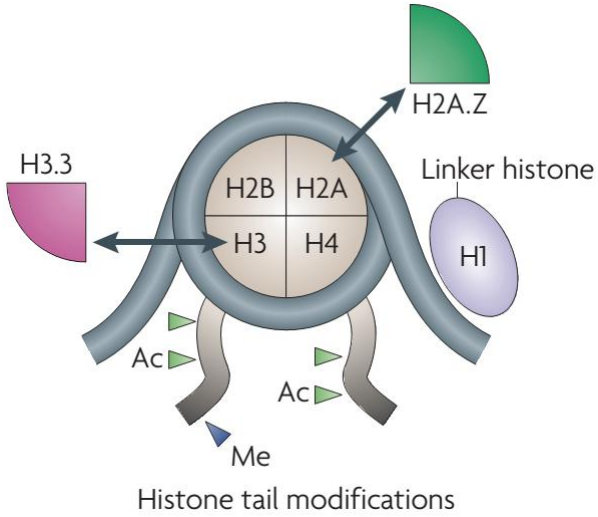


Figure 1.7: A schematic of DNA wrapped around a nucleosome. A nucleosome contains four canonical histones: H2A, H2B, H3 and H4. H3.3 and H2A.Z are variants of H3 and H2A, respectively. H3 and H4 tails can be subject to acetylation (Ac) and methylation (Me). Reprinted by permission from Macmillan Publishers Ltd: Jiang *et al.* [14], copyright 2009.

Over the last decade, many studies have investigated the genomic organization of nucleosomes for various species. The first genome-wide map of nucleosomes was created in yeast [56]. This map showed that the genic nucleosome landscape is conserved among most of the yeast genes (Figure 1.8). Specifically, two well-positioned nucleosomes (-1 and +1 nucleosomes) around the transcription start site (TSS) are bridged by a nucleosome free region (NFR). The gene body is packaged by an array of nucleosomes. This pattern is also apparent in metazoans [57,58]. In humans, nucleosome positioning is affected by transcriptional activity and differs considerably among cell types [59]. Furthermore, sequence context only has a moderate impact on nucleosome positioning *in vivo* [59]. We will discuss in more detail the role of nucleosome positioning in transcriptional regulation in section 1.3.

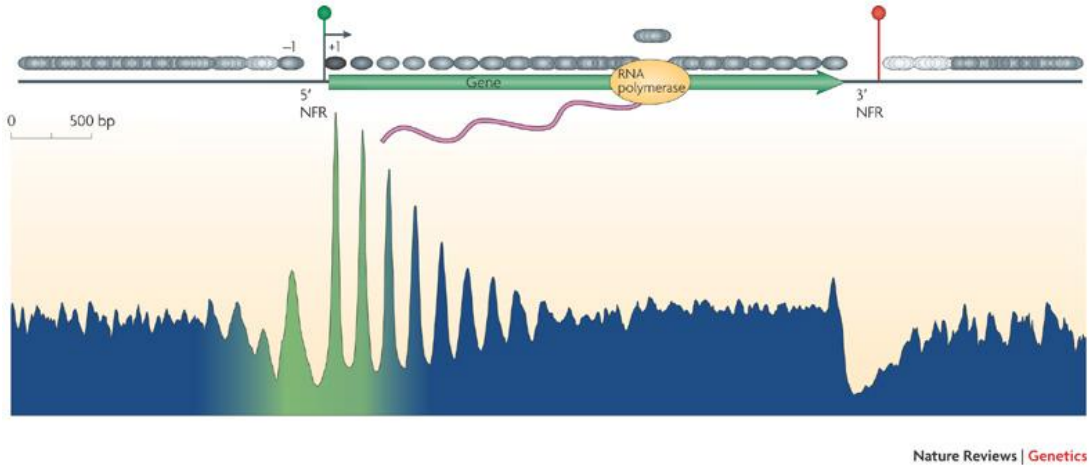


Figure 1.8: Nucleosome landscape of yeast genes. Top plot shows the consensus distribution of nucleosomes (ovals) around all genes. The middle plot shows the gene structure where green and red circles represent the transcription start site (TSS) and transcriptional termination site (TTS), respectively. The bottom plot shows the average nucleosome occupancy level across the gene. The green in this plot indicates high levels of nucleosome occupancy while the blue indicates lower levels of nucleosome occupancy. The green peaks correspond to well-positioned nucleosomes, which are represented by dark ovals in the top plot. The two most well-positioned nucleosomes are immediately downstream (+1 nucleosome) and upstream (-1 nucleosome) of the TSS. The region between these two nucleosomes are referred to as 5' nucleosome free region (5' NFR), which corresponds to the green valley at the bottom plot. The NFR that is upstream of the TTS (referred to 3' NFR) corresponds to the blue valley. Reprinted by permission from Macmillan Publishers Ltd: Jiang *et al.* [14], copyright 2009.

Recent advances in genomics techniques have enabled the identification of complete and high-resolution maps of nucleosomes in various eukaryotic genomes (Table 1.4). The ChIP-chip method, which uses chromatin immunoprecipitation (ChIP) followed by microarray (chip) analysis, allows the detection of the location of DNA-binding proteins throughout a genome. ChIP-chip can identify the positions of hundreds of thousands of nucleosomes in eukaryotic genomes [58,60]. The chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) method allows individual nucleosomal DNA molecules to be sequenced, thereby yielding tens of millions of nucleosomes [61,62].

Table 1.4. Summary of studies of *in vivo* nucleosome positioning.

Study	Organism	Key findings
Yuan *et al.* 2005 [56]	Yeast	First genome-wide map of nucleosomes. Pattern of nucleosome positioning along the gene.
Segal *et al.* 2006 [63]	Yeast	Nucleosome interacts with DNA in a sequence-specific manner. The finding of nucleosome code.
Lee *et al.* 2007 [60]	Yeast	Nucleosome occupancy signatures are correlated with transcription rate and are indicative of gene functions.
Albert *et al.*2007 [61]	Yeast	There is a tight regulatory relationship between promoter elements and the topology of nucleosome borders.
Mavrich *et al.* 2008 [58]	Drosophila	RNA Pol II engages and pauses at the +1 nucleosome.
Mavrich *et al.* 2008 [62]	Yeast	Promoter nucleosomes bind to DNA in a sequence-specific manner, but adjacent nucleosomes are dictated by packing principles.
Mito *et al.* 2007 [64]	Drosophila	Nucleosomes are replaced at *cis*-regulatory domains in order to maintain accessibility of *cis*-regulatory elements.
Schones *et al.* 2008 [65]	Human	First high-resolution and genome-wide human nucleosome map. Nucleosome phasing relative to the TSS is directly correlated to RNA Pol II binding.
Shivaswamy *et al.* 2008 [66]	Yeast	In response to physiological perturbation, several nucleosomes in the promoter are remodeled to enhance the accessibility of transcription factors that mediate transcriptional changes.
Valouev *et al.* 2008 [67]	C. elegans	C. elegans chromatin lacks of universal sequence-dictated nucleosome positioning.
Rhee *et al.* 2012 [68]	Yeast	Role of +1 nucleosome in assembling transcription pre-initiation complexes around TSS.
Lickwar *et al.* 2012 [69]	Yeast	Continual turnover of transcription factors and nucleosomes can poise a site for transcriptional activation.
Valouev *et al.* 2011 [59]	Human	Nucleosome positioning is dynamic among different human cell types and among distinct epigenetic domains in the same cell type.

Kundaje *et al.* 2012 [70]	Human	Genome-wide patterns of histone modifications, nucleosome positioning and sequence composition around transcription factor binding sites.

In addition, the strategy of using microccocal nuclease (MNase) digestion followed by massively parallel sequencing, MNase-seq, provides ultra-high resolution maps of nucleosomes as the MNase digestion can yield mono-nucleosomes by cutting the linker region of nucleosomes [65,70]. The latest method, chromatin immunoprecipitation with lambda exonuclease digestion followed by high-throughput sequencing (ChIP-exo) [71], is considered to yield highest resolution map of the nucleosomes so far. This method enables viewing the interaction between nucleosomes and transcription factors at single nucleotide resolution [68].

These nucleosome maps allow studying the genomic properties of chromatin. The majority of nucleosomes are distributed randomly and continuously in an array (referred to as fuzzy), while a proportion can be positioned within a narrow genomic region (referred to as phased). A nucleosome has two fundamental relationships with its DNA. A translational setting defines a nucleosomal midpoint relative to a given DNA locus. A rotational setting defines the orientation of DNA helix on the histone surface. Fuzzy nucleosomes have preferred positions that tend to be about 10 bp apart [61].

Phased nucleosomes are likely spaced and bridged by short sequences of linker DNA, which usually has a fixed length. Higher eukaryotes tend to have longer linker DNA. The length of linker DNA is ~18bp in yeast [60,62], ~28bp in drosophila and *C. elegans* [58,67], and ~38bp in human [57,65]. The nucleosome spacing can be determined by protein complexes such as the chromatin accessibility complex (CHRAC) and spacing complexes of the imitation switch (ISWI) [72,73,74]. These protein complexes bind to nucleosomes and a portion of linker DNA and then shift nucleosomes towards each other, thereby shortening the linker DNA. This process continues until the space between nucleosomes reaches a certain limit that does not allow ISWI binding [74,75]. The space is further determined by the linker histone H1 [76]. Different lengths of linker DNA in different eukaryotes

may reflect evolutionarily divergence of ISWI subunits or histone H1 that imply species-specific linker length. The length of linker DNA can affect the binding of other DNA-binding proteins (e.g. transcription factors) and therefore contributes to transcriptional regulation. Long linkers (NFRs) are one of the key elements that link nucleosome organization to gene regulation [14].

1.2.2 Chromatin modifications

The surface of the nucleosome is subject to various modifications that have profound influences on many cellular processes such as transcription, DNA repair and DNA replication. The canonical histones appear in a specific structure that facilitates the modifications (Figure 1.9A) [3]. Each histone contains two parts: the histone fold and the N-terminal tail. The histone fold is likely unmodified, but the N-terminal tail is long and can easily be modified at many residues (Figure 1.9B). There are at least seven distinct types of histone modifications (Table 1.5), with four popular types illustrated in Figure 1.9B.

Table 1.5. Different types of histone modifications

Modifications	Amino Acids[1]	Processes[2]
Acetylation	K	Transcription, Repair, Replication, Condensation
Methylation	K, R	Transcription, Repair
Phosphorylation	S, T	Transcription, Repair, Condensation
Ubiquitylation	K	Transcription, Repair
Sumoylation	K	Transcription
ADP ribosylation	E	Transcription
Deimination	R	Transcription

[1]K: lysine, R: Arginine, S: Serine, T: Threonine, E: Glutamic acid. [2] Repair: DNA repair, Replication: DNA replication, Condensation: Chromosome condensation.

There are two major functions of histone modifications. First, they can play a role in the formation of global chromatin architecture. Modifications help partition the genome into two distinct forms of chromatin: heterochromatin and euchromatin. Heterochromatin is highly condensed and represses many cellular activities, but euchromatin is more open and accessible. Second, histone modifications have a

major impact on the orchestration of DNA-based biological processes. To promote the execution of a DNA-based task, some histone modifications are formed and act together. Each task requires a specific set of modifications that can recruit the machinery to access the DNA and execute the task.

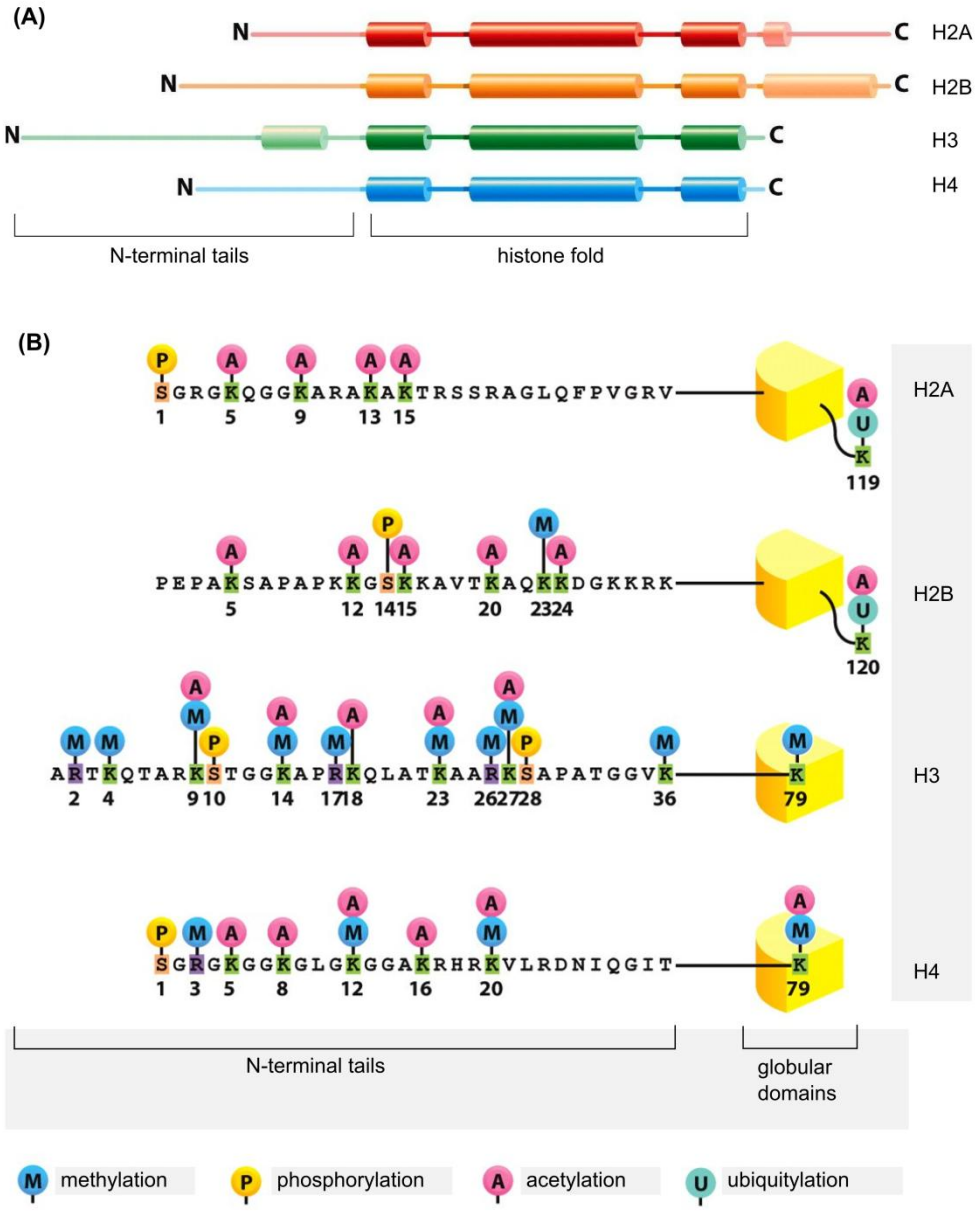


Figure 1.9: Structures and modifications of core histones. (A) Each histone contains a histone fold region and an N-terminal tail. (B) Four main classes of modifications found in the N-terminal tails, including methylation (M), phosphorylation (P), acetylation (A) and ubiquitylation (U). Reproduced with permission of GARLAND SCIENCE: Albert *et al.* [3], copyright 2008.

Formation of global chromatin architecture

The chromatin exists in two distinct forms: active euchromatin and silent heterochromatin. Each form is associated with a distinct set of histone medications. The separation of euchromatin and heterochromatin along the genome is performed by boundary elements (e.g. the CTCF transcription factor), which prevent heterochromatin from spreading into adjacent euchromatic regions [77,78,79]. In yeasts, boundaries between euchromatic and heterochromatic regions are maintained by the occurrence of methylation at H3K4 and H3K9 [80]. Therefore one crucial role of histone medications is to mark and preserve different chromatin states along the genome.

Euchromatic regions cover a large proportion of the genome. Euchromatin DNA is accessible, thereby enhancing DNA repair and replication [80]. Genes within these regions can be actively transcribed. Therefore the pattern of histone modification in euchromatin functions to mark this transcriptionally active state. Interestingly, levels of euchromatin modification are correlated with levels of transcription. Active genes are associated with high levels of acetylation and are tri-methylated at H3K4, H3K36 and H3K79 [57,81].

Heterochromatin is a crucial structure that mediates gene-silencing and ensures chromosome segregation and genomic integrity [80]. Heterochromatin exists in two forms: constitutive and facultative heterochromatin. Constitutive heterochromatin is mainly found at the centromeric and telomeric regions, which contains high density of repetitive DNA elements such as transposable elements and satellite sequences. Facultative heterochromatin is often located in developmentally regulated regions, where the chromatin state can change in response to gene activities and cellular signals.

Histone modifications play an important role in the formation of heterochromatin [82]. Heterochromatin is highly associated with hypoacetylation and methylation at H3K9 [82]. Figure 1.10 illustrates the canonical model of heterochromatin assembly. The assembly is generally initiated by the recruitment of histone deacetylases (HDACs) in order to ‘erase’ the acetylation at lysinses 9 and 14 of histone H3. The methyltransferases (HMTs) arrive to set the methylation

marks, which provide binding sites for the heterochromatin protein Swi6/HP1. Finally, self-activation of the heterochromatin protein establishes and spreads the inactive chromatin [83].

Bivalent domains are found to possess both activating and repressive modifications. The bivalent chromatin structure is normally represented by large regions of H3K27me3 harbouring smaller regions of H3K4me3 [84,85]. In embryonic stem (ES) cells, this pattern marks the promoter of more than 2000 genes, a portion of which are key developmental genes [84,86]. When ES cells were made to differentiate, bivalent domains only retained either the activating mark H3K4me3 or the repressive mark H3K27me3 [84,85]. Thus genes marked with bivalent domains are poised in ES cells, but can be activated or silenced upon differentiation.

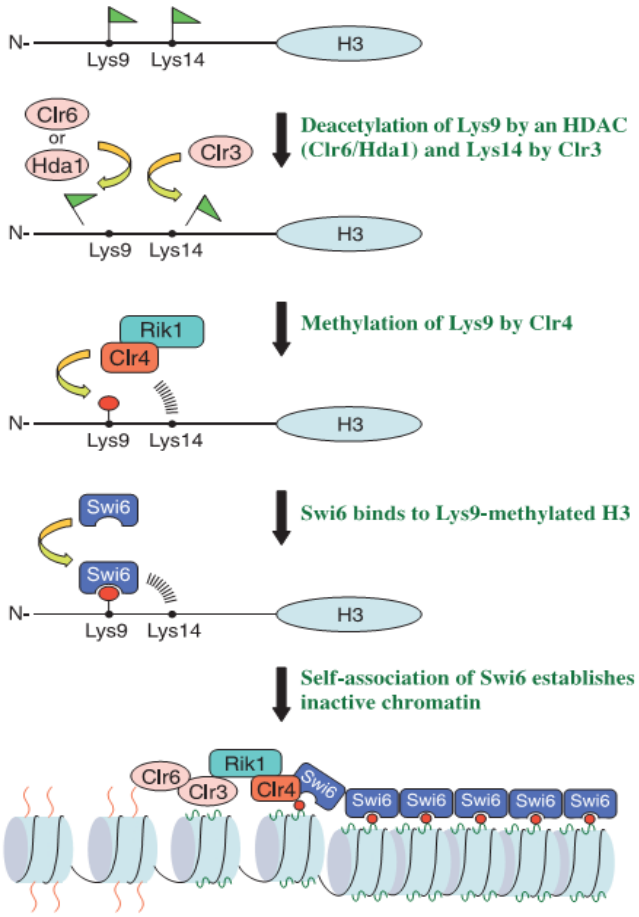


Figure 1.10: Model of heterochromatin formation and spreading. Green flags and red lollipops represents acetylation and methylation, respectively. Orange and green

protrusions represents N-terminal tails with and without acetylation, respectively. Sourced from [83]. Reprinted with permission from AAAS.

Orchestration of DNA-based biological processes

Histone modifications contribute to various processes that require DNA access such as transcription, the DNA damage response (DDR) and DNA replication [87]. The impact of histone modifications on translational regulation has been studied in detail over the last decade – we will describe this in section 1.3. Below, we discuss the role of histone modifications in DDR and DNA replication.

H2AX, an evolutionarily conserved variant of histone H2A, is an important chromatin factor. The phosphorylated form of H2AX, referred to as γH2AX, serves as a key regulator of the DDR. DNA double strand breaks (DSBs) have been known as the most common type of DNA damage [88]. The role of γH2AX in response to DSBs is similar to the role of H3K9me3 in the assembly of heterochromatin, mentioned above. γH2AX acts to orchestrate the formation of the DDR complex around the damage sites. Indeed, once the DSB happens, H2AX is rapidly phosphorylated to form γH2AX foci that trigger an ordered recruitment of DDR proteins to repair the damage [88].

In mammals, H2AX constitutes around 10% of the total histone H2A [89]. The mechanism of γH2AX formation and spreading is well established (Figure 1.11). Once a DSB is induced, the MRN complex is recruited and binds to the ends of the DSB. MRN then mediates the recruitment of ATM, one of the phosphoinositide 3-kinase related protein kinase (PIKK) that is responsible for phosphorylating H2AX at the DSB to form γH2AX. The phosphorylation of H2AX occurs at the conserved serine residue 139 (Ser139) of its C terminus. The resulting γH2AX recruits and provides docking sites for the mediator of DNA damage checkpoint 1 (MDC1), which stimulates further MRN-ATM complex recruitment and spreading of γH2AX to adjacent nucleosomes. This positive feedback loop can extend the γH2AX foci to more than 2 Mb of chromatin around the DSB [90].

Chromatin organization has a major impact on DNA replication. Genomic regions near the telomere are often targeted by HDACs, and are usually late-replicating due to the local chromatin structure [91]. Genome-wide analysis of DNA

replication timing has shown that DNA replicates earlier in euchromatin than heterochromatin [92,93,94,95]. A recent report also showed that euchromatin marks (e.g. H3K4me2, H3K4me3, acetylations of histones H3 and H4) are enriched in sequences replicated in early S phase from HeLa cells [96]. Perturbation of histone acetylation strongly affects DNA replication. Knocking down of Rpd3 (an HDAC) shortens the length of S phase, because of the earlier activation of a subset of replication origins [97].

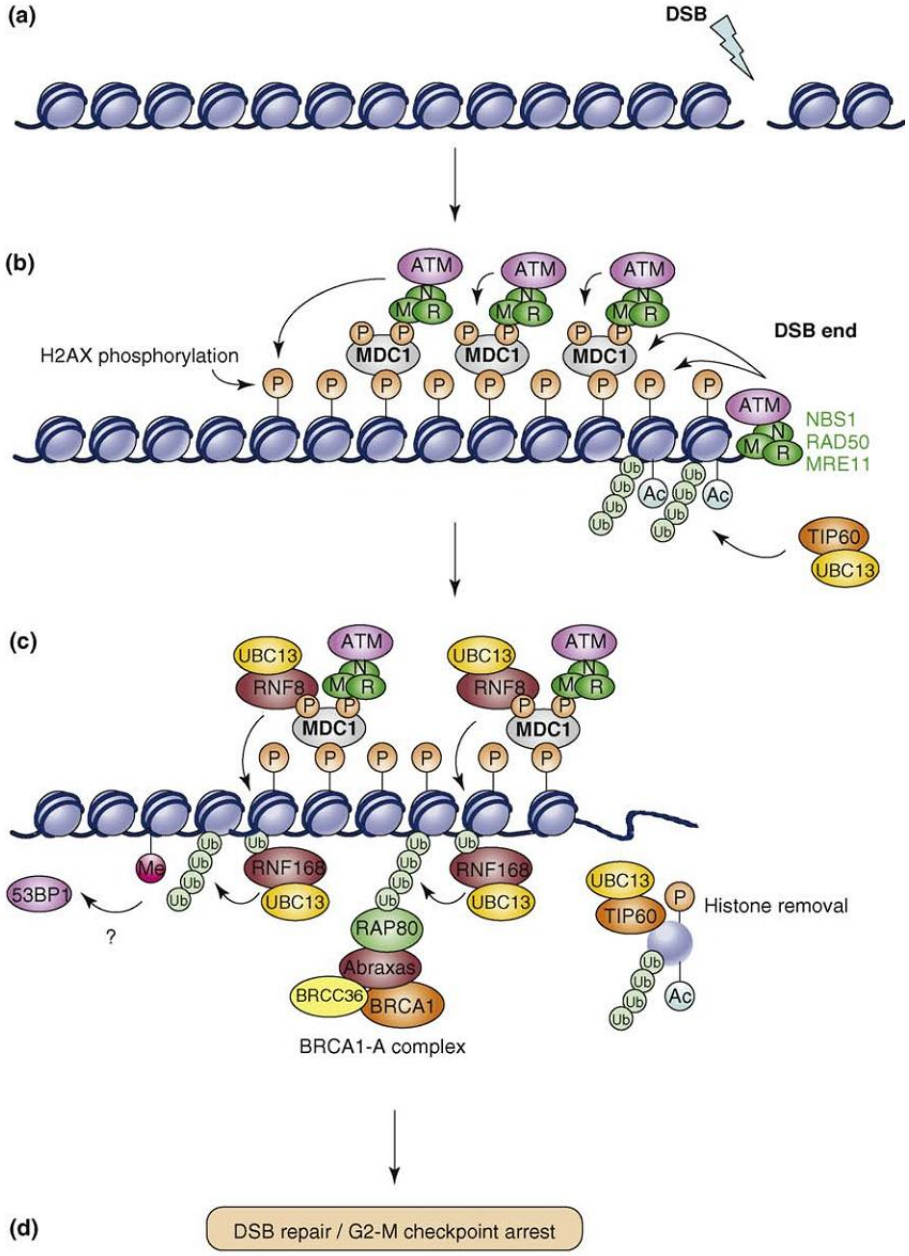


Figure 1.11: The role of γH2AX in the DSB repair. (a) DSB is induced by ionizing radiation. (b) The ends of the DSB are targeted by the MRN complex (Mre11,

Rad50 and Nbs1) that delivers ATM. The ATM then phosphorylates (labeled as P) H2AX to form γH2AX, which in turn recruits MDC1. The MRN-ATM complex is further recruited by the phosphorylation of MDC1 and thereby forming more γH2AX at nearby nucleosomes. This cycle is repeated and leads to the formation of around 2 Mb γH2AX foci surrounding the DSB. Next, TIP60 acetylates (Ac) γH2AX and then cooperates with the E2 ubiquitin-conjugating enzyme UBC13 to regulate polyubiquitylation (Ub) of acetylated γH2AX. (c) The histone at the ends of the DSB that contains the acetylated and polyubiquitylated γH2AX is evicted, likely to make way for repair proteins participating in subsequent steps. The RNF8-UBC13, which is recruited by phosphorylated MDC1, can bind to ubiquitylated histones and stimulates the formation of ubiquitin conjugates. 53BP1 and BRCA1-A complexes are next delivered by the polyubiquitylated histones and start the DSB repair and/or checkpoint arrest (d). Reprinted from [88], Copyright 2009, with permission from Elsevier.

1.3 The role of chromatin organization in gene regulation

1.3.1 Nucleosome positioning and gene regulation

Global nucleosome maps have provided insights into the organization of nucleosomes around genes and how this organization influences transcription. As described above, there is a distinctive pattern of nucleosome positioning around genes, particularly at the TSS (Figure 1.8). The -1 nucleosome, which is located upstream of the TSS, occupies a region from -300 to -150 bp relative to the TSS [14]. Importantly, it has an impact on the accessibility of promoter regulatory elements. Dynamic remodeling of the -1 nucleosome is associated with the formation of transcription machinery. The -1 nucleosome is evicted upon the recruitment of RNA polymerase II (Pol II), but after the transcription pre-initiation complex (PIC) has partly assembled [98].

Following the -1 nucleosome is a nucleosome free region (referred to as the 5' NFR), which is immediately upstream of the TSS. The discovery of NFRs provided key insights into understanding the contribution of nucleosome organization to transcriptional regulation. The 5' NFR is likely the site for the assembly of the transcription machinery [14,68]. Whereas the 3' NFR, adjacent to the transcription

termination site, is likely the sites for the disassembly of the transcription machinery. However, the 3' NFR of one gene may severe as the 5' NFR of the next downstream gene [14,68]. Interestingly, formation of the 5' NFR facilitates the deposition of histone variant H2A.Z from the two 5' NFR-flanking nucleosomes (-1 and +1 nucleosomes) [99]. This suggests the role of 5' NFRs in the assembly of promoter chromatin architecture.

Downstream of the 5' NFR is the TSS, which is followed by the +1 nucleosome. Among all nucleosomes around the gene, the +1 nucleosome appears to be the most well-positioned one. Transcription requires the assembly of PIC and the recruitment of Poll II to the promoter. In active genes, the +1 nucleosome normally contains histone variants H3.3 and H2A.Z and histone tail modifications, all of which can be involved in nucleosome remodeling and PIC assembly [80,100]. During each cycle of transcription, +1 nucleosomes can be evicted, but they may rapidly return to their original locations after Pol II has passed [61]. A recent study in yeast has revealed a clear picture of how the +1 nucleosome contributes to PIC organization and transcription dynamics (Figure 1.12) [68]. In TATA-containing promoters, +1 nucleosomes and PICs can competitively assemble. This may result in the loss of +1 nucleosomes and, thereby, releases the Pol II and allows higher level of transcription [68]. Whereas, in the TATA-less promoters, +1 nucleosomes cooperatively assemble with PICs and, thereby, can assist the TSS selection by pausing Pol II (Figure 1.12) [68].

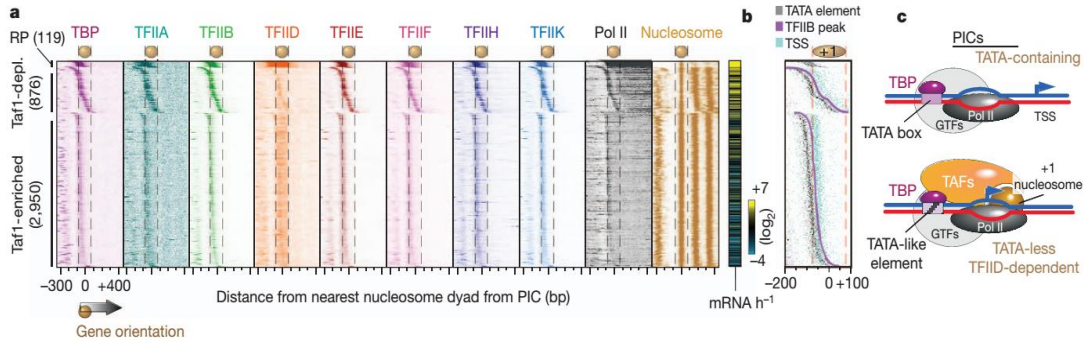


Figure 1.12. The role of +1 nucleosome in PIC assembly. a, Occupancy of general transcription factors (GTFs) around the +1 nucleosome in two groups of promoters: TATA-containing (Taf1-depleted) and TATA-less (Taf1-enriched). The nucleosome borders are denoted by vertical dashed black lines and the right panel

shows transcription frequency. **b**, Same as panel **a**, but showing an overlay of TATA elements, TFIIB and TSS. **c**, Model of PIC organization at TATA-box-containing and TATA-less/TFIID-dependent genes. Reprinted by permission from Macmillan Publishers Ltd: Rhee *et al.* [68], copyright 2012.

Following the +1 nucleosome, nucleosomes display less tight positioning. The nucleosome +2 is immediately downstream of the +1 nucleosome. Compared with the +1 nucleosome, the +2 nucleosome contains less H2A.Z and is less modified. The +3 nucleosome displays less of these properties than the previous upstream one. Looking further downstream, nucleosomes tend to be less positioned. However, nucleosome positioning is distinct between introns and exons. Nucleosomes are well-positioned in exons and preferentially bind to exons rather than introns [101,102]. Nucleosome positioning has been considered as a marker for exons, even among inactive genes [103].

1.3.2 Histone modifications and gene regulation

Much of eukaryotic DNA is occupied by nucleosomes and is packaged within higher-order chromatin structures [100]. Transcription can be initiated by the targeting of pioneer transcription factors (also known as activators) to the promoter [104]. This initial binding can help open up the local chromatin and make it more accessible for other binding factors [104]. Once bound to the promoter, activators can trigger a cascade of ordered recruitment of co-activator complexes such as chromatin-remodeling complexes and histone-modification enzymes (Figure 1.13). These complexes not only promote the binding of activators to DNA but also make nucleosomal DNA more accessible to general transcription factors. In addition, transcriptional repression often requires the recruitment of chromatin-remodeling complexes and histone-modification enzymes to the promoter. This pathway can be involved in the formation of heterochromatin, as we discussed above (Figure 1.6), to condense the chromatin and silence the target gene. Thus chromatin modifications and transcription factors can act together to control gene expression.

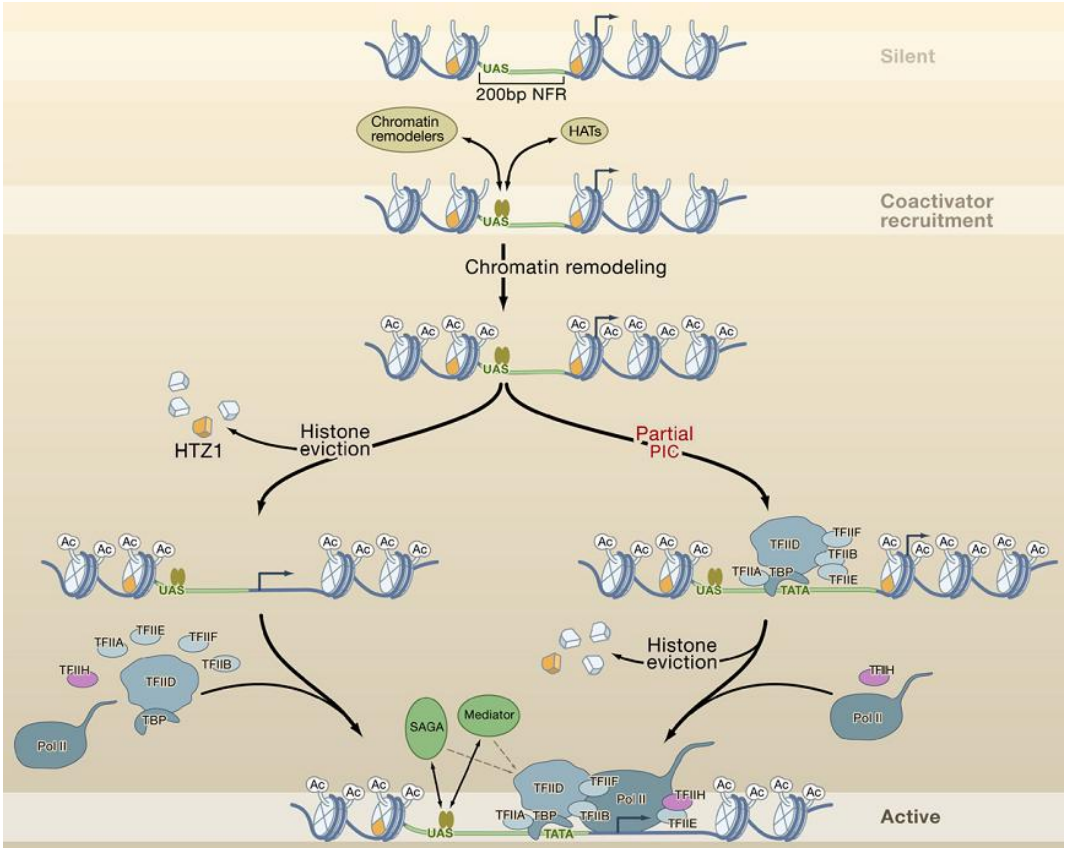


Figure 1.13: Models of the regulation of transcription initiation by chromatin modifications. Chromatin remodelers and enzymes such as histone acetyltransferase complexes (HATs) are delivered to the upstream-activation sequence (UAS) to modify histone tails of nearby nucleosomes. The assembly of PICs at the core promoter requires the eviction of a Htz1-containing nucleosome. Htz1, a variant of histone H2A.Z, is enriched at promoters that are poised for transcriptional activation. Reprinted from [100], Copyright 2007, with permission from Elsevier.

In terms of their roles in the regulation of transcription, histone modifications are classified into those that are associated with activation (referred to as activating marks) and those that are associated with repression (referred to as repressing marks). A genome-wide and high-resolution map of 39 histone modifications in humans was established by Barski *et al.* [57] and Wang *et al.* [81] using ChIP-seq (Table 1.6). Only five modifications were found to be repressing marks (Figure 1.14). Interestingly, levels of histone modifications around the promoter were reported to be well correlated with gene expression [57,81,105]. Also, this relationship holds true across different human cell lines [106].

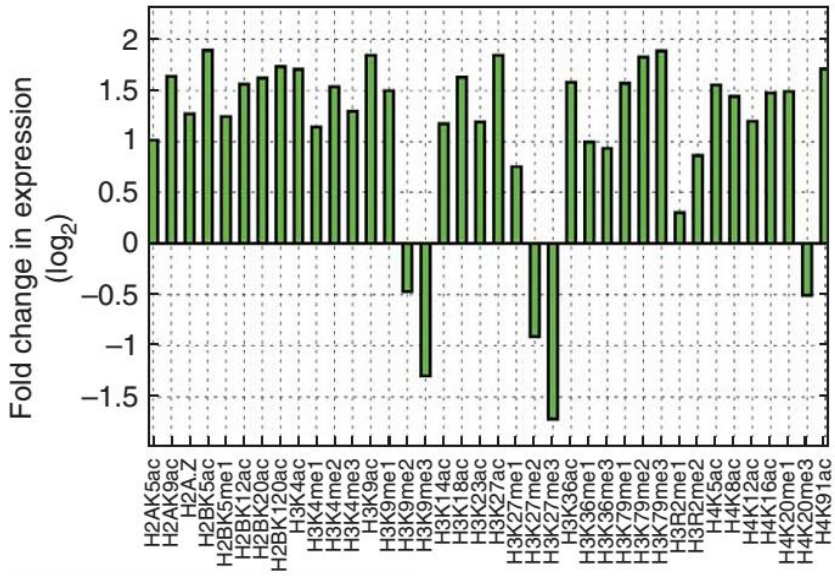


Figure 1.14: Correlation between histone modifications and gene expression. Positive and negative bars correspond to activating and repressive modifications, respectively. Reprinted by permission from Macmillan Publishers Ltd: Wang *et al.* [81], copyright 2008.

The majority of histone modifications that influence transcription are methylation and acetylation of lysines [81,100,107]. Acetylation can loosen chromatin condensation, partly because adding an acetyl group to lysine removes its positive charge, thereby increasing DNA accessibly. Thus acetylation is primarily associated with transcriptional activation. These types of modifications are reversible. Acetyl groups are added to lysines by a set of HATs and can be removed by a set of histone deacetylase complexes (HDACs), which is referred to as deacetylation [80].

Table 1.6: Summary of histone modifications

Modifications[1]	Description[2]	Localizations	REFs
H2A.Z	Variant of H2A	Promoter, Backbone[3]	[57,108]
H2AK5ac, 9ac	H2A lysine 5 acetyl	Euchromatin	[81]
H2AK9ac	H2A lysine 9 acetyl	Euchromatin	[81]
H2BK120ac	H2B lysine 120 acetyl	Promoter, Backbone	[81]
H2BK12ac	H2B lysine 12 acetyl	Promoter, Backbone	[81]
H2BK20ac	H2B lysine 20 acetyl	Promoter, Backbone	[81]
H2BK5ac	H2B lysine 5 acetyl	Promoter, Backbone	[81]
H2BK5me1	H2B lysine 5 mono-methyl	Euchromatin	[81]
H3K14ac	H3 lysine 14 acetyl	Euchromatin	[109]
H3K18ac	H3 lysine 18 acetyl	Backbone	[81]
H3K23ac	H3 lysine 23 acetyl	Euchromatin	[81]
H3K27ac	H3 lysine 27 acetyl	Backbone	[81]
H3K27me1	H3 lysine 27 mono-methyl	Euchromatin	[57]
H3K27me2, me3	H3 lysine 27 di,tri-methyl	Polycomb-repressed	[57,84]
H3K36ac	H3 lysine 36 acetyl	Backbone	[81]
H3K36me1, me3	H3 lysine 36 mono,tri-methyl	Elongation	[110,111]
H3K4ac	H3 lysine 4 acetyl	Promoter, Backbone	[81]
H3K4me1, me2	H3 lysine 4 mono,di-methyl	Enhancer; Backbone	[112]
H3K4me3	H3 lysine 4 tri-methyl	Promoter; Backbone	[113] [81]
H3K79me1, me2, me3	H3 lysine 79 mono,di,tri-methyl	Euchromatin	[57]
H3K9ac	H3 lysine 9 acetyl	Backbone	[81]
H3K9me1	H3 lysine 9 mono-methyl	Backbone	[57,81]
H3K9me2, me3	H3 lysine 9 di,tri-methyl	Heterochromatin	[114]
H3R2me1,me2	H3 arginine 2 mono, di-methyl	Euchromatin	[81]

H4K12ac,16ac	H3 lysine 12, 16 acetyl	Euchromatin	[81]
H4K20me1	H4 lysine 20 mono-methyl	Elongation	[115,116]
H4K20me3	H4 lysine 20 tri-methyl	Heterochromatin	[115,116]
H4K5ac	H4 lysine 5 acetyl	Backbone	[81]
H4K8ac	H4 lysine 8 acetyl	Backbone	[81]
H4K91ac	H4 lysine 91 acetyl	Backbone	[81]

[1]Similar modifications with the same function were shown in the same row, e.g. H3K9me2, me3 corresponds to two modifications: H3K9me2 and H3k9me3. [2]acetyl: acetylation, methyl: methylation. [3]Backbone corresponds to robust modifications acting together at promoter regions [57]. Rows presenting repressive modifications are shaded in grey.

Histone methylations can be associated with activation or repression, depending on their state and position. Histone methyl transferases (HMTs) are responsible for adding the methyl groups at lysines. HMTs normally modify one single lysine on a single histone. Similar to the acetylations, the methylations can be reversed by the histone demethylases. A brief description of many histone marks is shown at Table 1.6. Methylation marks are associated with transcription from different localizations such as promoter, transcribed region and enhancer. Here we focus on some well-studied marks that play key roles in transcriptional regulation.

Methylation of histone H3K4 is strongly associated with gene activation. H3K4me3 is a prominent histone mark that is found within core promoters and early-transcribed regions of active genes [113] (Figure 1.15). H3K4me3 is directly coupled to other chromatin remodeling complexes to facilitate transcriptional activation by enhancing the accessibility of chromatin to the transcriptional machinery [117]. A recent study reported that H3K4me3 regulates the PIC assembly through direct interactions with TAF3 [118]. Although they are found at active promoters, H3K4me1 and H3K4me2 are preferentially associated with enhancer regions [112]. H3K4me1 is now considered as a universal mark of many enhancers. Enhancers are clusters of DNA sequences capable of communicating with promoters by a long-range interaction mechanism (Figure 1.15).

Methylation of H3K36 by the SET2 histone methyltransferase is a key histone mark associated with transcriptional elongation. Both H3K36me2 and H3K36me3

are enriched at the 3' ORF [119]. H3K36me3 can recruit a subunit of the Rpd3S histone deacetylase complex, thereby forming a hypoacetylated structure within ORFs [110,111]. This structure helps prevent spurious intragenic transcription initiation [111]. The chromatin pattern that corresponds to H3K4me3 followed by H3K36me3 (referred to as K4-K36) indicates RNA Pol II-transcribed transcripts [120]. This K4-K36 pattern was used to identify many long non-coding RNAs in mouse cells [120].

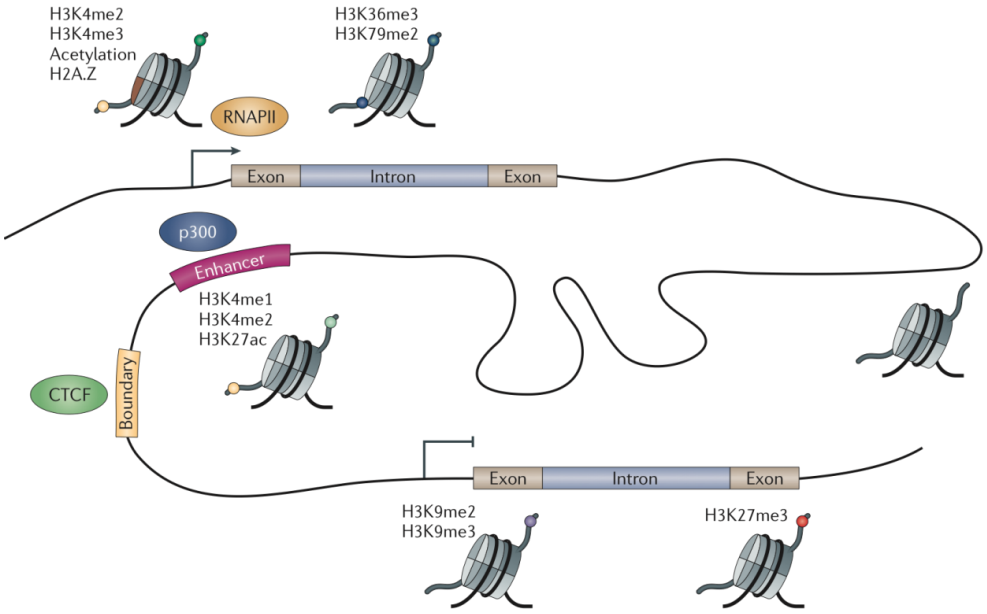


Figure 1.15: Genomic localizations of key histone marks. H3K4me2, H3K4me3, acetylation and H2A.Z are marks of active promoters. H3K36me3 and H3K79me3 are marks associated with gene bodies. H3K4me1 is the key mark of enhancer regions, which can be the target of p300. A chromatin loop can enable interaction between an enhancer and a promoter. H3K9me2, H3K9me3 and H3K27me3 are key marks of inactive gene. Reprinted by permission from Macmillan Publishers Ltd: Zhou *et al.* [85], copyright 2011.

Methylation of H3K9, particularly H3K9me3, is a key factor participating in the pathway of heterochromatin formation and gene silencing at repetitive DNA elements [114]. H3K9me3/2 provides binding sites for heterochromatin proteins (e.g. Swi6) that can recruit the HDACs to reduce the Pol II accessibility, thereby leading to transcriptional silencing [79]. However, H3K9me1 was reported to be associated with transcriptional activation [57].

Tri-methylation of H3K27 is another key repressive mark [121]. H3K27me3 is associated with the gene silencing that is mediated by the Polycomb group (PcG) proteins. The most well-known role of H3K27me3 is to provide a docking site for Polycomb repressive complex 1 (PRC1), a family of PcG. In a typical PcG-mediated silencing pathway, H3K27me3 is initially catalyzed by the histone modifying enzyme EZH2, which is a component of PRC2. H3K7me3 in turn recruits PRC1, which catalyzes H2Aub1 and thereby impedes RNA Pol II elongation [121].

Chromatin can be segmented into a set of discrete states, where each state is associated with one or more histone modifications and indicates a specific biological role [122,123,124,125,126,127,128,129]. More than 100 different histone modifications have been reported [80], allowing genome-wide characterization of chromatin states. Using only nine different chromatin marks, Ernst *et al.* [123] were able to segment the human genome into 15 distinct states. The resulting maps of chromatin sates differ considerably among different cell types, an example is provided as Figure 1.16 [123]. Importantly, patterns of enhancer activity and promoter activity were shown to be strongly correlated with patterns of nearest-gene expression, thereby providing a means of linking enhancers to potential target genes [123].

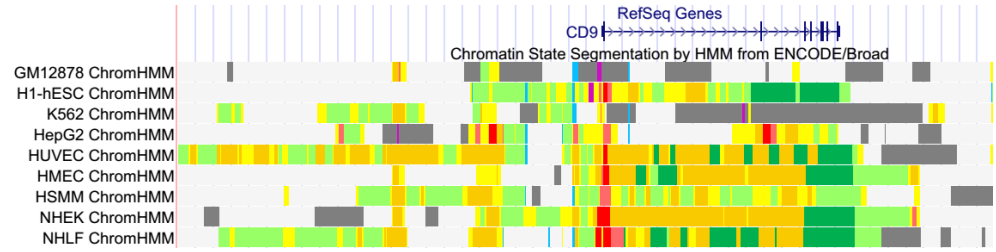


Figure 1.16: Chromatin states across 9 different human cell types surrounding the CD9 gene on chromosome 12. Different colors represent different states: bright red - active promoter; light red - weak promoter; purple - poised promoter; orange - strong enhancer; yellow - weak enhancer; blue - insulator; green - transcription; gray - Polycomb-repressed; and light gray - heterochromatin or low signal. This figure is generated from the UCSC genome browser [130] using the chromatin state segmentation data from Ernst *et al.* [123].

1.4 High-throughput genomic technologies

Recent developments in high-throughput techniques have provided great opportunities for exploring many aspects of genomics research. Microarrays and next generation sequencing are two major technologies that have been applied widely over the last decade. Next generation sequencing is now used more often as it can yield higher resolution of data. This section focused on two applications of next generation sequencing: ChIP-seq and RNA-seq.

1.4.1 ChIP-seq and analysis methods

ChIP-seq overview

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP–seq) is used to detect and characterize DNA-protein binding events, nucleosomes and histone modifications (Figure 1.17; reviewed by Park [131] and Furey [132]). In a chromatin immunoprecipitation (ChIP) experiment for DNA-binding proteins, the protein of interest is cross-linked to DNA by treating cells with a gentle formaldehyde fixation and the chromatin is then fragmented by sonication. This is referred to as X-ChIP (X denotes cross-linking) [133]. The fragments corresponding to the protein of interest are collected by using an antibody that is specific to that protein. Finally, the resulting fragments are purified and the released DNA is sequenced on any sequencing platform. Most of the ChIP-seq data have been generated by Illumina platforms [132]. In a ChIP experiment for nucleosomes or histone modifications, micrococcal nuclease (MNase) treatment without cross-linking can be used to replace sonication because MNase can digest linker DNA more efficiently [65]. This is referred to as N-ChIP (N denotes native) [134].

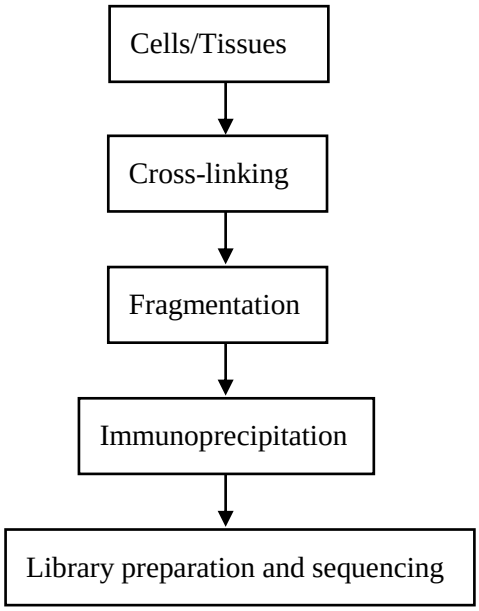


Figure 1.17: Overview of a ChIP-seq workflow. A chromatin immunoprecipitation (ChIP) experiment followed by sequencing (seq) can profile proteins (non-histone ChIP) or histone modifications (histone ChIP) of interest. During the ChIP process, DNA and the protein (or histone modification) of interest can be cross-linked, and the resulting chromatin is sheared into fragments. The antibody that is specific to the protein of interest is then used to select the corresponding DNA fragments. The resulting DNA fragments are purified and can be sequenced on any sequencing platform.

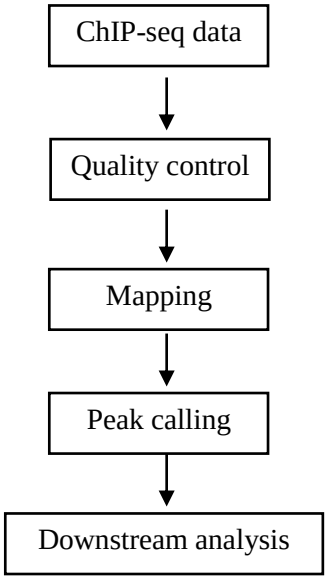


Figure 1.18: Four basic steps in a typical workflow for ChIP-seq data analysis.

Analysis of ChIP-seq data

A typical pipeline for analysing ChIP-seq data contains four main steps: quality control, filtering, mapping, peak calling and downstream analysis (Figure 1.18).

Quality control

The raw sequence data coming from high throughput sequencing pipelines are often in fastq format (http://maq.sourceforge.net/fastq.shtml). The first step in the analysis pipeline is to check the quality of raw sequence data – referred to as quality control. FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), HTSeq (http://www-huber.embl.de/users/anders/HTSeq/) and the ShortRead package in Bioconductor [135] are helpful for this purpose. Various analyses (e.g. sequence quality, GC content; see Figure 1.19) help check whether the raw data have any problems of which users should be aware before carrying out further analysis.

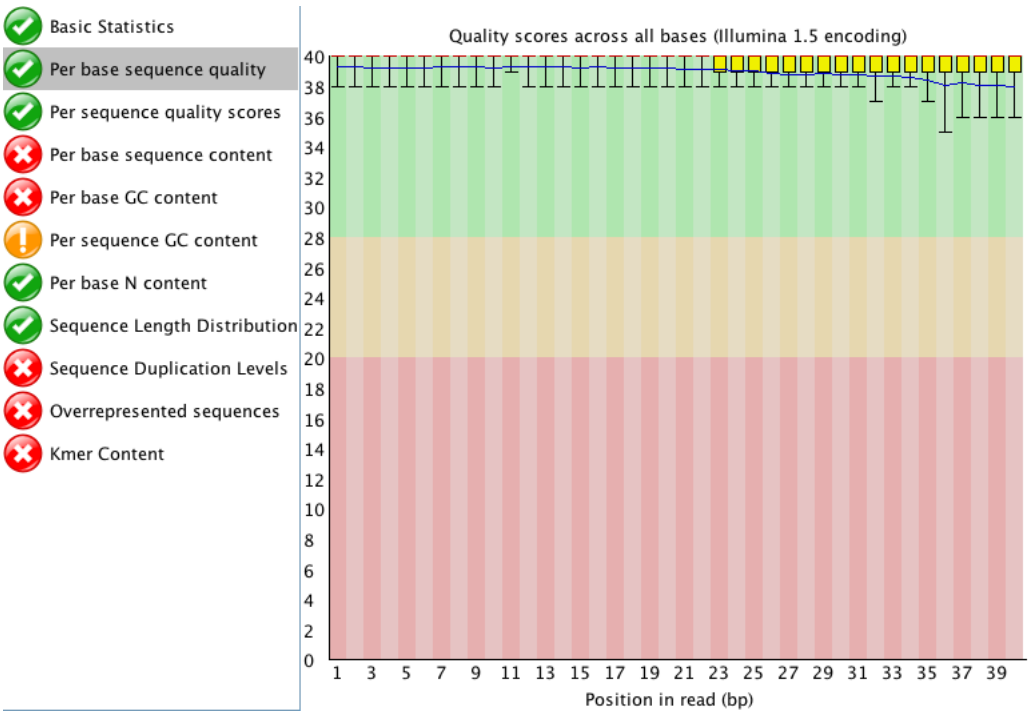


Figure 1.19: Quality control output from FASTQC. The left panel shows 11 different analyses for checking the quality of a fastq file. Green, yellow and red icons represent three statuses of the analysis: pass, warn and fail, respectively. The right panel shows the distribution of sequence quality for each base along the read.

The Y-axis represents the quality score that is calculated as $-10\log_{10}(p)$, where p is the probability that the corresponding base call is incorrect. Three layers in the right panel: green, yellow and red show the good quality, acceptable quality and bad quality ranges, respectively.

Filtering

Filtering is helpful to yield better results when mapping the raw data to the reference genome. Filtering can be used to trim the low quality bases from reads and discard low quality reads. Also, the raw data sometimes contain the barcodes or adapters, which should be removed before mapping. Other manipulations can be done during the filtering steps. Galaxy (https://usegalaxy.org/) and FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) provide various functions to handle this task.

Mapping

The next step is to align the short read sequences to a reference genome. The key question is how to align many short reads to a large reference genome at the lowest computational cost. Over the last decade, many tools were developed to optimise the computational efficiency (reviewed in Li *et al.* 2010 [136]; some commonly used ones are listed in Table 1.7). Tools such as Bowtie, BWA and MAQ use "indexing" methods to accelerate the mapping. For example, Bowtie indexes the genome using the Burrows-Wheeler transform. For the human genome, it allows more than 25 million 35-bp reads to be mapped per CPU hour, with a memory footprint of less than 2 GB [137].

Table 1.7: A selection of short-read aligners

Aligner	Reference	URL
Bowtie	[137]	http://bowtie-bio.sourceforge.net
Bowtie 2	[138]	http://bowtie-bio.sourceforge.net/bowtie2
BWA	[139]	http://bio-bwa.sourceforge.net/
MAQ	[140]	http://maq.sourceforge.net/
SOAP2	[141]	http://soap.genomics.org.cn/soapaligner.html
Mosaik		http://code.google.com/p/mosaik-aligner/

Peak calling

Following mapping, the peak calling step identifies sites of protein:DNA binding by finding regions of increased sequence read tag density relative to the background [142]. Many peak calling software packages (referred to as peak callers) have been developed (a selection is shown in Table 1.8). Peak calling can be divided into five basic components: (1) signal profiling, (2) background estimation, (3) peak identification, (4) signification ranking and (5) artifact removal.

Most ChIP-seq data is from single-end sequencing. To account for all possible bases involved in protein binding, reads are extended towards 3' end by a shift-size that can be determined from the data [142]. The extended read is referred to as tag. There are two simple methods to define the signal profile of ChIP-seq data. The first method is to slide a window of fixed length across the genome then count the number of tags in each window [143,144,145,146]. The second method is to count the number of tags overlapping each position along the genome [147,148].

Table 1.8: A selection of peak calling software

Tool	Reference	URL
MACS	[143]	http://liulab.dfci.harvard.edu/MACS/
SCICER	[149]	http://home.gwu.edu/~wpeng/Software.htm
ZINBA	[150]	http://code.google.com/p/zinba/
cisGenome	[144]	http://biogibbs.stanford.edu/~jihk/CisGenome/index.htm
SiSSRs	[145]	http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/sissrs/
Spp	[146]	http://compbio.med.harvard.edu/Supplements/ChIP-seq
PeakSeq	[151]	http://info.gersteinlab.org/PeakSeq
ERANGE	[152]	http://woldlab.caltech.edu/rnaseq

The background model is helpful to filter out false positive regions that come from DNA shearing biases or sequencing artifacts. When control data (e.g. Input DNA) are available, the treatment signal can be normalized by subtracting the control signal or dividing by the control signal in each window. When the control data are not available, the background tag distribution can be modelled with a random distribution, such as Poisson [143] or negative binomial distribution [144].

A peak is called when the (normalized) signal exceeds a pre-defined threshold. Most peak callers calculate p-value and false discovery rate (FDR) for each peak, thereby providing the significance of the called peaks. Peaks containing only a few reads are assumed to be PCR amplification artifacts and removed. Peaks with a significant imbalance between the numbers of tags on each strand should also be discarded.

Downstream analysis

One of the most popular analyses following the peak calling step is the motif identification. The DNA sequences underlying highly-significant peaks can be used to search for DNA pattern of the protein. DREME [153], a component of the MEME Suite of motif-based sequence analysis tools (http://meme.nbcr.net), was developed for this task. Another useful downstream analysis is peak annotation that helps determine whether the significant peaks are associated with any interesting functions or pathways [154].

1.4.2 RNA-seq and analysis methods

RNA-seq overview

RNA-seq (RNA sequencing) uses high-throughput sequencing technology for transcriptome profiling (reviewed in [155,156]). Figure 1.20 demonstrates how an RNA-seq experiment works. In general, a pool of RNA molecules is converted to a complementary DNA (cDNA) library, which is then cut into small cDNA fragments. Each fragment can be amplified with PCR and is then sequenced by a high-throughput sequencing platform (e.g. Illumina) from one end (single-end sequencing) or from both ends (pair-end sequencing). The resulting reads are normally 30-500 bp, depending on the purpose and the platform used. Next, the reads are mapped to a reference genome or a reference transcriptome to quantify the expression of genes.

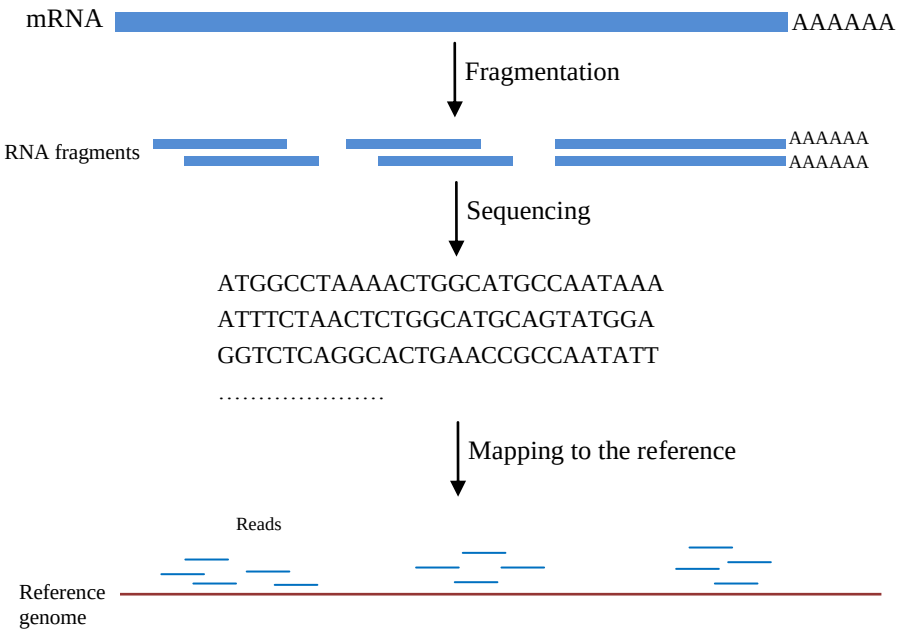


Figure 1.20: Overview of RNA-seq. mRNA molecules with poly (A) tail are selected and sheared into small RNA fragments. These fragments are then sequenced and mapped onto a reference genome or transcriptome.

RNA-seq data analysis

Similar to ChIP-seq data, the RNA-seq data derived from a sequencing platform are checked for quality (quality control) and are pre-processed. The resulting RNA-seq data are then mapped to a reference genome. RNA-seq reads are typically generated from transcribed and processed RNA sequences where intronic and intergenic sequences are excluded. Therefore if a read maps to a splicing junction, it will fail with the alignment methods used for ChIP-seq data. Tophat [157], STAR [158] and QPALMA [159] were developed to handle this issue. Tophat uses Bowtie to map reads to the genome and identify putative exons from clusters of mapped reads. The unmapped reads are then mapped against the sequences flanking the possible junctions.

Many developments in RNA-seq allow various aspects of transcriptomics to be investigated, such as gene/isoform expression quantification, differential gene expression, transcription start site mapping, strand-specific measurements, non-coding RNA detection, gene fusion detection, transcriptome assembly, and detection of alternative splicing events [156].

In RNA-seq studies, transcript levels are measured in reads per kilobase of exon model per million mapped reads (RPKM) [160]. The RPKM indicates the molar concentration of a transcript by normalizing for RNA length and for the total number of mapped reads in the study. Cufflinks [161] quantifies transcript levels in Fragments Per Kilobase of exon per Million fragments mapped (FPKM), which is analogous to the RPKM measure. The FPKM reflects the relative abundances of transcripts in terms of the expected biological objects (fragments). The Tuxedo protocol [162], uses a collection of tools: Bowtie, Tophat, Cufflinks and cummeRbund (http://compbio.mit.edu/cummeRbund/), to provide a means for differential gene and transcript expression analysis.

Chapter 2: Genome-wide Distribution of DNA Damage Dependent Histone H2AX

My contribution: I performed all the data analyses in this chapter.

2.1 Abstract

Histone variant H2AX plays a key role in the DNA damage response. It is present around the sites of double strand breaks (DSBs) and phosphorylation of H2AX acts to signal these events, expediting their repair. However, the global distribution of H2AX in the human genome still remains poorly understood. We have generated high-resolution and genome-wide maps of H2AX in U2OS cells by using Illumina high-throughput sequencing. Strikingly, we found that H2AX appeared to be abundant in heterochromatin regions, marked by trimethylation of lysine 9 in histone H3 (H3K9me3), which are known to replicate later than euchromatin regions during DNA replication process. In support this finding, we observed that H2AX is enriched in late-replicated DNA. Parallel biochemical investigations showed that H2AX expression is up-regulated in the later phases of DNA replication, consistent with preferential incorporation into heterochromatin. Heterochromatin has previously been shown to be refractive to damage signalling through H2AX phosphorylation and, consequently, we hypothesize that the greater abundance of H2AX in heterochromatin helps to ensure sufficient H2AX phosphorylation to signal DNA damage events. However, we discovered several problems with the quality of the ChIP-seq sequence data that was generated for this project, resulting in a need for caution in the interpretation of the results obtained.

2.2 Introduction

Histone variant H2AX is reported to be present in 2–20% of mammalian nucleosomes and serves as a key regulator of the DNA damage response [163,164]. It is involved in the recruitment and accumulation of the DNA repair proteins to the sites of DNA double-stranded breaks (DSBs) [165,166,167]. H2AX has been reported to play roles in preventing chromosomal abnormalities that are associated

with cancer in mammals [89,168], and to act as a tumor suppressor [169,170]. In eukaryotes, upon DSB induction, H2AX becomes rapidly phosphorylated at serine 139 – termed as γH2AX – within several minutes and forms large foci of γH2AX [90,171]. These γH2AX foci spread into large domains (from 0.5 to 2 Mb) around the DSBs to signal the damage and recruit DNA repair factors [172]. However, the global distribution of H2AX itself throughout the genome still remains poorly understood.

γH2AX foci formation is known to associate with changes in chromatin structure. In eukaryotes, chromatin is a complex of DNA and histone proteins (e.g. histones) that forms chromosomes. Chromatin is classified into two major forms: euchromatin that is more open and is generally transcribed, and heterochromatin that is highly condensed and represses many cellular activities. γH2AX foci have been shown to occur preferentially in open chromatin regions in mammalian cells following X-irradiation [173,174]. Correspondingly, γH2AX foci are largely excluded from the heterochromatin regions both in yeast and mammalian cells [173]. But the DNA damage response is not necessarily less efficient if the DSBs happen within heterochromatin. The chromatin regions around heterochromatic DSBs are generally remodeled to a more relaxed state that facilitates the formation of γH2AX and the recruitment of repair factors [175]. It is possible that the abundance of H2AX itself contributes to adequate formation of γH2AX and DNA repair factors in heterochromatic regions. Therefore, the genomic distribution of H2AX may provide insights into DNA damage signalling and the formation of γH2AX foci.

In this study we have determined a comprehensive landscape of H2AX in the human genome for U2OS cells using Chromatin ImmunoPrecipitation (ChIP) following by high-throughput sequencing (seq) (ChIP-seq). The human osteosarcoma U2OS cell line was derived from the bone tissue of a fifteen-year-old human female suffering from osteosarcoma [176]. We have analysed the links between H2AX and other genomic features including DNA replication timing and chromatin structure. Our analyses showed that H2AX enrichment increases monotonically with DNA replication timing. This led us to compare H2AX abundance in euchromatin and heterochromatic regions marked by H3K9me3. In

addition, RT-qPCR experiments show that the expression of H2AX peaks in the later stages of S phase, suggesting that H2AX could be preferentially incorporated in heterochromatin, which is generally late replicated. However, this study has several caveats concerning GC content and sequencing depth of the ChIP-seq data. Consequently, there is a need for caution in the conclusion about the enrichment of H2AX in heterochromatin regions. This study is a collaboration with Dr. Andrew Flaus's lab in the Centre for Chromosome Biology, NUI Galway who generated the ChIP-seq data sets and carried out the RT-qPCR experiments.

2.3 Results

2.3.1 H2AX and H2B ChIP-seq libraries

To indentify the distribution of H2AX across the human genome, we performed ChIP-seq for two histone variants H2AX and H2B. A summary of these data is shown in Table 2.1 and Figure 2.1. The first 60 bp from the 5' end towards 3' end of each read had very high quality scores (Figure 2.1).

Table 2.1: Summary of H2AX and H2B ChIP-seq libraries

Histone	# of reads	Read length	Coverage[1]	# of uniquely mapped reads
H2AX	13,328,281	80bp	36%	7,291,642 (55%)
H2B	9,322,750	80bp	25%	5,857,169 (63%)

[1]Coverage represents percentage of the genome covered by the reads and is calculated as (L x N)/G where L is read length, N is total number of reads, and G is the length of genome.

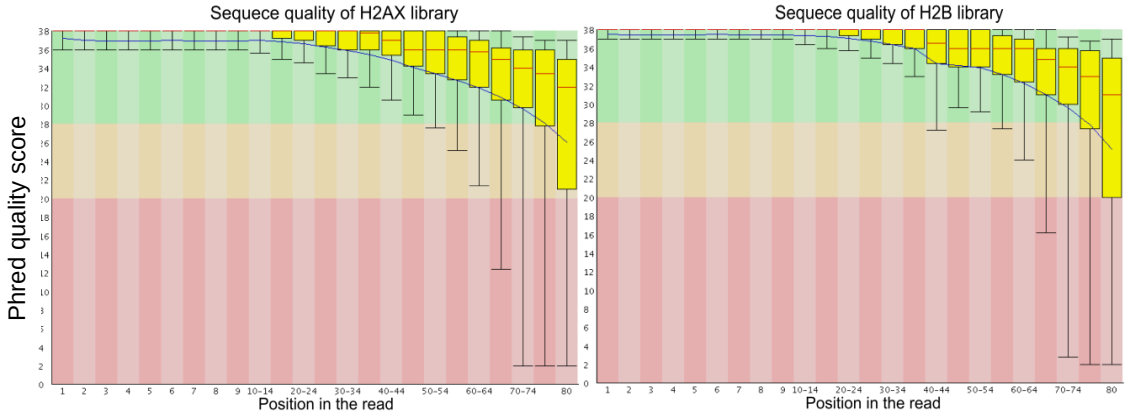


Figure 2.1: Per base sequence quality of H2AX and H2B libraries. These plots were generated using FASTQC[1], where Phred quality score is calculated following Cock *et al.* [177].

However, the GC contents of both H2AX mapped reads and H2B mapped reads, which average 51% and 57%, respectively, are high compared to the GC contents of other ChIP-seq data available in the public domain – that generally from 43% to 49% in average (Figure 2.2). The GC content bias has previously been shown to originate both from library preparation and from amplification before and during sequencing [131,178,179]. In addition, sequence duplication levels of the two ChIP-seq libraries are very high. Using Picard (http://picard.sourceforge.net/) we found sequence duplication levels in H2AX and H2B data sets are 68% and 63%, respectively. We also noted that these ChIP-seq data have low sequencing depth (see Table 2.1), relative to recent ChIP-seq guidelines of ENCODE [180]. Therefore, to determine the genomic distribution of H2AX with greater confidence would require generation of more and better quality ChIP-seq data.

[1] FASTQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

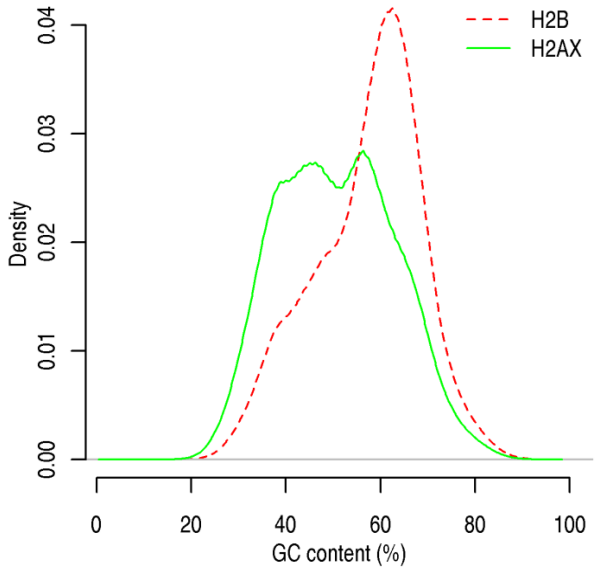


Figure 2.2: Density of GC contents of H2AX reads and H2B reads that are mapped uniquely to the human genome.

2.3.2 H2AX is abundant in heterochromatin regions

We analysed the H2AX and H2B ChIP-seq datasets to identify the global distribution of H2AX throughout the human genome. The H2B abundance was used to control for nucleosome density because H2B is a canonical histone variant and is therefore expected to present in all nucleosomes. We set out to assess whether histone H2AX is preferentially localized in heterochromatic or euchromatic regions. For this purpose, we obtained ChIP-seq data for trimethylation of lysine 9 in histone H3 (H3K9me3) and trimethylation of lysine 4 in histone H3 (H3K4me3) from Barski *et al.* [57] which was generated from $CD4^+$ T cells. We note that the genome-wide distribution of H3K4me3 and H3K9me3 might differ between $CD4^+$ T cells and U2OS cells (used for generating the H2AX and H2B ChIP-seq data). H3K9me3 and H3K4me3 are known to mark heterochromatin and euchromatin, respectively [57]. We used MNase-seq data which was generated from the same cell type ($CD4^+$T) to measure nucleosome genome-wide occupancy to normalize the H3K9me3 and H3K4me3 data [65]. We mapped all the datasets to the human genome version HG18 using Bowtie [137], and kept only reads that mapped uniquely for subsequent analyses (see Methods). Using non-overlapping windows of size 100kb we assessed the relationship between the enrichment of H2AX and the enrichment of H3K9me3 and H3K4me3 across the human genome. In each

window, we normalized the H2AX counts by dividing by the number of mapped reads from the H2B experiment; and normalized H3K9me3 counts and H3K4me3 counts by dividing by the number of mapped reads from the MNase-seq experiment. (see Methods).

Interestingly, we found that the normalized H2AX signal is strongly correlated with the normalized H3K9me3 signal (Spearman *rho* = 0.43; $p < 1 \times 10^{-200}$), indicating that H2AX is enriched in heterochromatin regions. Correspondingly, the normalized H2AX signal is negatively correlated with the normalized H3K4me3 signal (Spearman *rho* = -0.40; $p < 1 \times 10^{-200}$). We classified all the 100 kb windows into five levels increasing from "Lowest" to "Highest" by the normalized signal of the chromatin marks. Consistent with the above observations, the normalized H2AX signal increases monotonically from "Lowest" to "Highest" levels of H3K9me3 and decreases monotonically from "Lowest" to "Highest" levels of H3K4me3 (Figure 2.3).

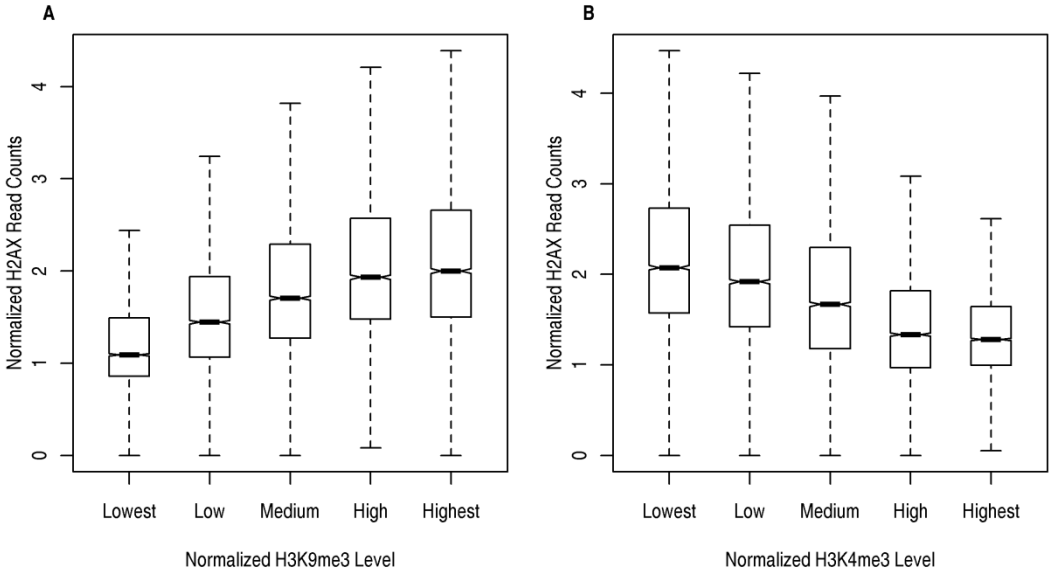


Figure 2.3: Enrichment of H2AX in different levels of H3K9me3 and H3K4me3. (A) Boxplots show the normalized enrichment of H2AX increasing from lowest level to highest level of H3K9me3. (B) Boxplots show the normalized signal of H2AX decreasing from lowest level to highest level of H3K4me3. H2AX was normalized to H2B; H3K9me3 and H3K4me3 were normalized to nucleosome density.

2.3.3 H2AX is enriched in later phases of DNA replication

We downloaded a high-resolution replication timing data from a study of Chen *et al.* [95] which was generated from HeLa cells by massively parallel sequencing of whole-genome nascent BrdU-labeled replicating DNA. We binned the DNA replication time into five phases ranging from earliest (S1) to latest (S5). Interestingly, H2AX is more enriched in the latter phases of DNA replication (Figure 2.4), and is strongly correlated with the DNA replication timing (Spearman *rho* = 0.34; $p < 1 \times 10^{-200}$). This is consistent with our prior observations regarding the enrichment of H2AX in heterochromatin because heterochromatin replicates later during the DNA replication process than euchromatin. Previously, chromatin compaction has been reported to be negatively correlated with DNA replication time both in *D. Melanogaster* [92,93,94] and in human [95].

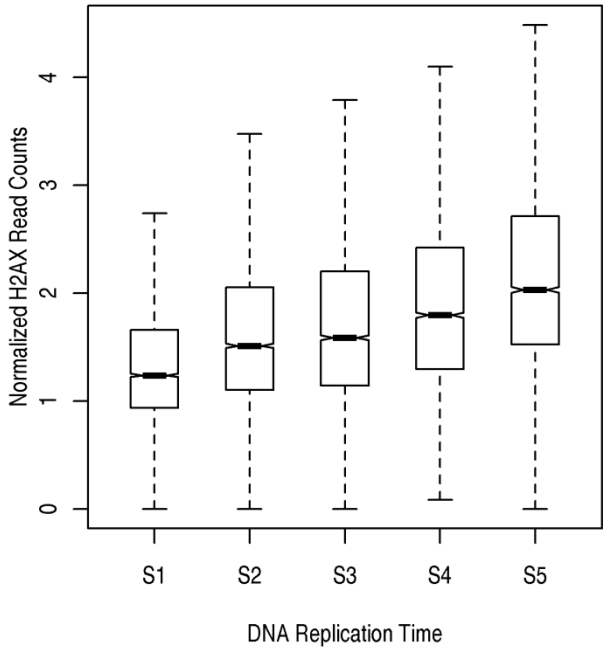


Figure 2.4: Enrichment of H2AX in different phases of DNA replication. DNA replication timing was binned in five equal periods of S-phase from S1 (earliest) to S5 (latest) following Chen *et al.* [95].

RT-qPCR experiments showed that H2AX tends to express late during DNA replication (S phase of the cell cycle) (Figure 2.5). We found that the expression of H2AX peaks later than the expression of H2A in the S phase. This result, together

with the fact that heterochromatin is late replicated, suggests a mechanism by which H2AX is preferentially incorporated into heterochromatin. In addition, as the mutation rate is known to markedly increase in late-replicating regions of the human genome [181], it is possible that H2AX is enriched in the regions having high mutation rate.

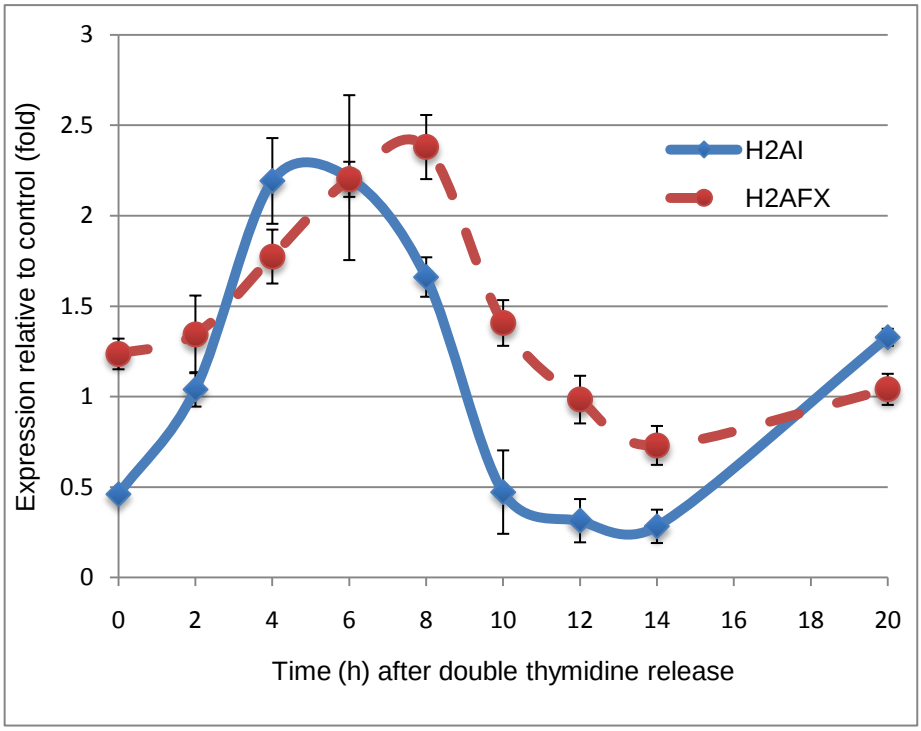


Figure 2.5: Relative expression levels of H2A and H2AX from RT-qPCR experiments. GAPDH is used as the endogenous control. Error bars represent standard error of the mean calculated over three biological replicates (in plate triplicates). H2AI and H2AFX are genes coding for canonical histone H2A and histone variant H2AX, respectively. Time after double thymidine release is an indicator of S phase progression. This figure was produced by collaborators in Dr. Andrew Flaus's lab.

2.3.4 Enrichment of H2AX within different chromatin states

Previously, a study of Ernst *et al.* has identified ten major chromatin states in the human genome and annotated the whole genome with these states in nine human cell types [123]. We obtained this data from UCSC genome browser [182] and measured the enrichment of H2AX relative to H2B within each chromatin state

(see Methods). Strikingly, we found that H2AX is more abundant in the heterochromatin state than the promoter/enhancer states (Figure 2.6). This is consistent with the results we found above that H2AX is also enriched in heterochromatic regions. Surprisingly, H2AX is enriched in transcribed states, suggesting H2AX might be involved in mRNA transcription process.

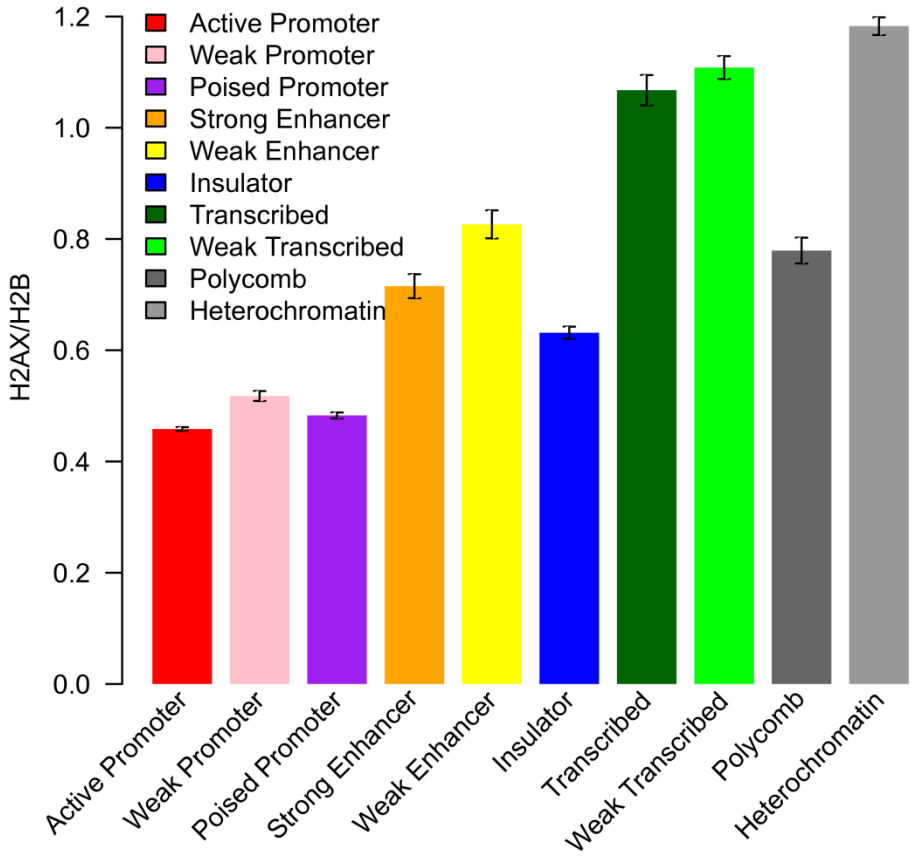


Figure 2.6: Enrichment of H2AX in different chromatin states. H2AX/H2B was calculated as the total number of H2AX reads divide the total number of H2B reads within each of ten chromatin states. The chromatin state data were obtained from nine different cell types [123]. Error bars represent standard error of the mean of the H2AX/H2B ratio.

2.3.5 H2AX is enriched in repetitive regions

Repetitive and repeat-derived DNA elements have been estimated to comprise up to 69% of the human genome [183]. Regions of high density of repetitive DNA elements are the main target of heterochromatin formation [184,185,186].

Heterochromatin is known to protect genome integrity and stability by repressing illegitimate recombination between dispersed repetitive DNA elements [185]. Motivated by the previous observations that H2AX is enriched in heterochromatic regions, we set out to determine whether H2AX is enriched in repetitive regions.

We used RepeatMasker [187], with default parameters, to screen short read sequences of H2AX and H2B for interspersed repeats and low complexity DNA sequences. We separated the short read sequences into three groups based on mapping results: multi-mapped, un-mapped and unique-mapped corresponding to reads mapped to multiple locations, reads failed to map, and reads mapped uniquely, respectively. As expected, we found that H2AX sequences comprise of more repetitive DNA elements than H2B sequences, particularly within the unique-mapped group, indicating H2AX is more abundant in repetitive regions (Figure 2.7). We also observed that H2AX is more enriched for long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) (Figure 2.8).

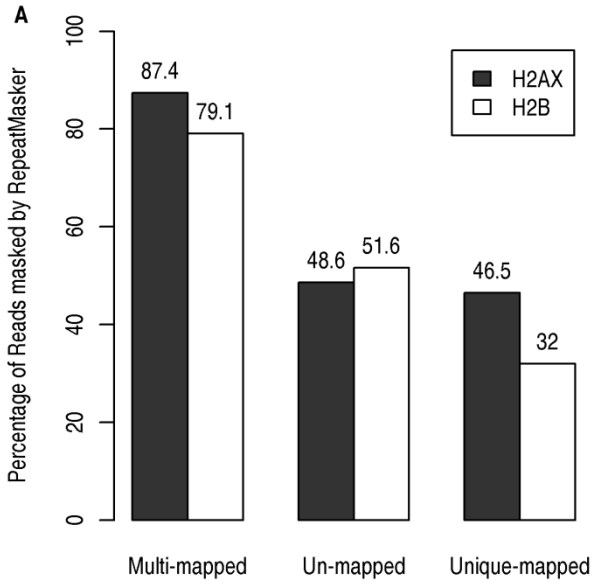


Figure 2.7: Percentage of H2AX and H2B short read sequences that are comprised of repetitive DNA elements. The number above each bar represents the exact percentage corresponding to each category.

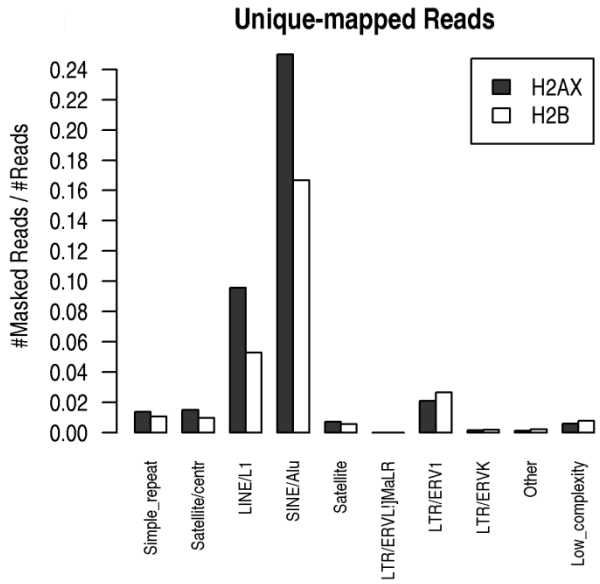


Figure 2.8: Proportion of H2AX and H2B short read sequences masked by RepeatMasker in different types of the repetitive elements.

2.4 Discussion

In this study, we generated H2AX and H2B ChIP-seq libraries and characterized global landscape of these histone variants across the human genome. We found multiple lines of evidence supporting the enrichment of H2AX in heterochromatin regions using our ChIP-seq datasets. Heterochromatin formation has been shown to contribute to maintaining genome stability [188]. Heterochromatic regions are condensed and, therefore, inhibit or slow down many cellular processes [3]. For example, heterochromatic DSBs are repaired with slower kinetics and with less efficiency than euchromatic DSBs [189,190]. Within heterochromatin regions, H2AX is not efficiently phosphorylated in response to DSBs [174,191]. Therefore, we hypothesize that the greater abundance of H2AX in heterochromatin helps to ensure sufficient H2AX phosphorylation to signal heterochromatic DSBs.

However, these findings come with several caveats. The enrichment of H2AX relative to H2B within the heterochromatin state is quite weak (only around 1.2) and that might not be high enough to conclude with confidence that H2AX is preferentially incorporated into heterochromatin (Figure 2.6). The abundance of H2AX in heterochromatin may be more pronounced if the coverage of H2AX and

H2B ChIP-seq data is improved. Previous ChIP-seq studies have shown that higher sequencing coverage normally yields better power of detecting protein binding sites [131,146]. After removing duplicate reads in the H2AX and H2B data sets, we still found high correlation between the normalized H2AX signal and the normalized H3K9me3 signal (Spearman *rho* = 0.4; Table 2.2). Therefore, the read duplicate problem should not be a concern in these data sets but removing duplicate reads reduced the coverage.

To address these problems, we made a second attempt to generate additional data for this study. The additional data consisted of ChIP-seq for H2A, H2AX, H3K9me3 and H3K4me3. We also generated two control data sets: IgG control and Input DNA. The Input DNA was derived from a portion of the DNA sample removed prior to immunoprecipitation (IP) and the IgG control was from non-specific IP. All datasets were sequenced at high coverage depth and at good quality. However, they are unexpectedly well correlated with one another. The genomic distribution of these data across the human genome are very similar to the Input DNA. We suspected that the ChIP experiments have pulled down all DNA fragments, not only those containing the protein of interest. As a consequence, all the data sets generated in the second attempt are similar to the Input DNA. There are many steps in a ChIP experiment and it is difficult to determine precisely the reason for the failure of these experiments. One possibility is that protein and DNA were not cross-linked together during ChIP experiments and consequently all DNA fragments were pulled down for sequencing.

Recently, a study by Seo *et al.*, carried out at a similar time to our study, profiled the global distribution of H2AX in the human Jurkat cells [192]. The Jurkat cell line is an immortalized line of human T lymphocyte cells that was derived from the peripheral blood of a 14 year old boy with T cell leukemia [193]. Surprisingly, they found endogenous H2AX is concentrated on the transcription start site of actively transcribed genes [192]. Also, a more recent study from this group [194] reported that, in activated T cells, H2AX is abundant in early-replicating genomic regions. Using the Jurkat H2AX ChIP-seq data from Seo *et al.* [192], we calculated the enrichment of H2AX in different chromatin states using the Jurkat Input DNA as control – the same procedure we applied for our H2AX/H2B data (see Section

2.3.4). We observed that the Jurkat H2AX is preferentially localized in the poised promoter state and the active promoter state (Figure 2.9) – in sharp contrast to our results above that H2AX is enriched in heterochromatin state (Figure 2.6). We calculated the normalized signal of the Jurkat H2AX in 100 kb non-overlapping windows across the human genome using the Jurkat Input DNA as control – the same method we carried out to measure our normalized H2AX signal (see the section 2.3.2), and compared our U2OS data and Seo's Jurkat data. Spearman correlation coefficients among all factors are shown in Table 2.2. Strikingly, our normalized H2AX signal is negatively correlated with the normalized Jurkat H2AX signal (Spearman *rho* = -0.6). Interestingly, we observed that the H2AX signal (without normalization) is highly correlated between our data and the Seo's Jurkat data (Spearman *rho* = 0.73; $p < 1x10^{-200}$); but the H2B signal shows no significant correlation with the Jurkat Input data (Spearman *rho* = 0.10). Therefore the difference between the two studies likely resulted from the difference between the H2B data and the Jurkat Input data. We carried out a quality control for the Jurkat Input data with FASTQC and found that the per-base sequence quality of this data set is low (Figure 2.10). Using FASTX-Toolkit we found that about half of the Jurkat Input reads should be filtered out due to having low sequence quality. Fortunately, this data set has very high coverage and we did not find significant changes in the results after filtering out the low quality reads.

In conclusion, there are some caveats in our study that lead to difficulties in concluding with confidence that H2AX is enriched in heterochromatin. Our study used data generated from different cell types (e.g. H2AX from U2OS cells; H3K4me3 and H3K9me3 from $CD4^+$ T cells; DNA replication timing from HeLa cells). In addition, H2B should be a good control for H2AX, but it might not be good enough for a genome-scale study. We set out to design our study similar to current standard ChIP-seq studies of histone modifications. Unfortunately, the new ChIP-seq experiments failed. We cannot claim with confidence that the conclusions of Seo *et al.* [195] are wrong and further investigation is required. In particular PCR experiments at well-known H2AX enriched regions would help to determine whether the ChIP-seq experiments of Seo *et al.* [195] or the results reported here

provide a more accurate estimate of the enrichment of H2AX along the human genome.

Table 2.2: Coefficients of Spearman correlations among different data sets

	Input	H2AX	H2B	$H2AX^*$	$H2B^*$	K4	K9	J_H2AX_{norm}	$H2AX_{norm}$	$H2AX^*_{norm}$	$K4_{norm}$	$K9_{norm}$
J_H2AX	0.2	0.7	0.8	0.7	0.8	0.7	-0.1	0.8	-0.5	-0.5	0.5	-0.6
Input	1	0.2	0.1	0.2	0.1	0.3	0.4	-0.5	0.3	0.3	0.1	0.1
H2AX		1	0.9	1	0.9	0.7	-0.2	0.6	-0.4	-0.4	0.6	-0.6
H2B			1	0.9	1	0.7	-0.2	0.7	-0.7	-0.7	0.6	-0.6
$H2AX^*$				1	0.9	0.7	-0.2	0.5	-0.4	-0.4	0.6	-0.6
$H2B^*$					1	0.7	-0.2	0.6	-0.7	-0.7	0.6	-0.6
K4						1	-0.2	0.5	-0.4	-0.4	0.9	-0.6
K9							1	-0.5	0.4	0.4	-0.5	0.9
J_H2AX_{norm}								1	-0.6	-0.6	0.4	-0.6
$H2AX_{norm}$									1	0.9	-0.4	0.4
$H2AX^*_{norm}$										1	-0.4	0.4
$K4_{norm}$											1	-0.5
$K9_{norm}$												1

J_H2AX, Input are the Jurkat H2AX and Input DNA data of Seo *et al.* [195]. J_H2AX_norm is J_H2AX normalized to Input. H2AX, H2B are our U2OS H2AX and H2B data sets. $H2AX_{norm}$ is H2AX normalized to H2B. $H2AX^*$, $H2B^*$ correspond to H2AX and H2B data sets after removing duplicate reads. $H2AX^*_{norm}$ is $H2AX^*$ normalized to $H2B^*$. K4, K9 are CD4+ H3K4me3 and H3K9me3 of Barski *et al.* [57]. $K4_{norm}$, $K9_{norm}$ are H3K4me3 and H3K9me3, respectively, normalized to the CD4+ nucleosome data of Schones et al. [65].

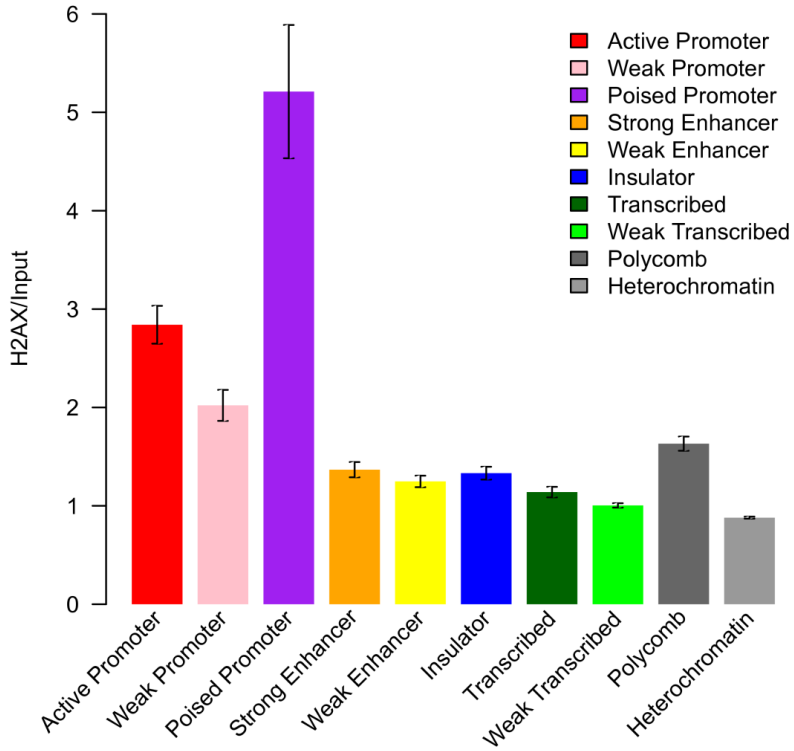


Figure 2.9: Enrichment of Jurkat H2AX in different chromatin states. H2AX and Input ChIP-seq data were obtained from Seo *et al.* [192]. H2AX/Input was calculated as the total number of H2AX reads divided by the total number of Input reads within each of ten chromatin states. Error bars represent one standard error.

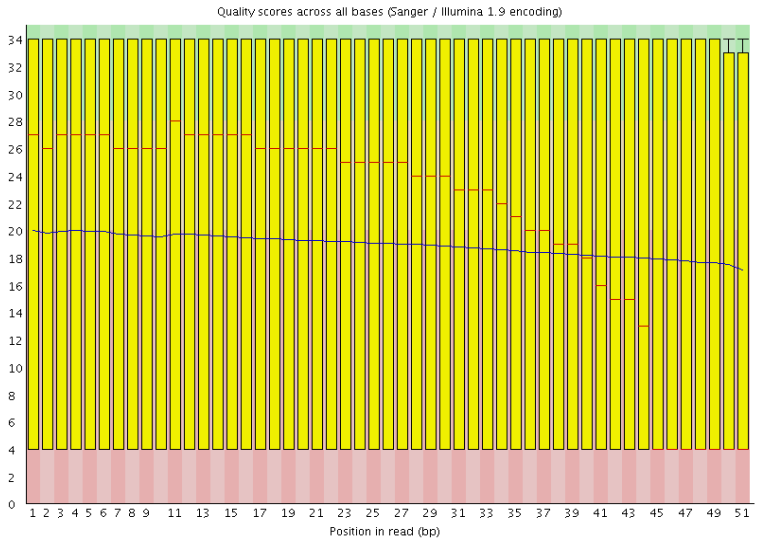


Figure 2.10: Per base sequence quality of the Jurkat Input data that was obtained from Seo *et al.* [192].

2.5 Methods

2.5.1 Data

ChIP-seq libraries for H2AX and H2B were generated by the Centre for Chromosome Biology at National University of Ireland Galway. ChIP-seq data for two chromatin marks H3K4me3 and H3K9me3 in $CD4^+$ T cells were downloaded from Barski *et al.* [57]; MNase-seq (Micrococcal Nuclease) data for mono-nucleosome in $CD4^+$ T cells were downloaded from Schones *et al.* [65]. We obtained whole-genome chromatin state segmentation data of Ernst *et al.* [123] from the UCSC genome browser [182] for nine ENCODE cell types: GM12878, HEPG2, HSMM, K562, NHLF, H1-hESC, HMEC, HUVEC and NHEK.

2.5.2 Mapping

We used Bowtie (version 0.12.7) to map all the ChIP-seq data sets used in this study to the human genome version HG18 [137]. The parameters used are "-l 60 -a --best --strata -m 1" to take into account of sequencing quality and to yield the best mapping rate. We chose first 60 bp from the 5' read ends as seed regions (-l 60) where read quality is considered and two mismatches are allowed during the mapping. We set output parameters to report all alignments for each read and partition them into best alignment stratum (-a --best --strata). Only uniquely aligning reads were kept for further analyses, with potential duplicates being removed from the alignments using Picard (http://picard.sourceforge.net/).

2.5.3 Signal normalization

We used 100 kb non-overlapping windows across the HG18 and calculated number of H2AX reads and number of H2B reads in each window, denoted as x_i and y_i respectively. The enrichment of H2AX relative to H2B in the i^{th} window, w_i, is calculated as:

$$w_i = (N_y/N_x)(x_i + 1) / (y_i + 1) \qquad (2.1)$$

Where N_x, N_y are the total number of unique-mapped reads of H2AX and H2B, respectively. The factor N_y/N_x is included to correct for the difference in the library size between H2AX and H2B. The ratio $(x_i + 1) / (y_i + 1)$ is the enrichment of H2AX relative to H2B; we added a pseudocount of 1 to each side to avoid the

ratio being zero or infinity. We used the same method to calculate the enrichment of chromatin marks H3K4me3 and H3K9me3 relative to the nucleosome in the $CD4^+$ T cells data sets [65].

To normalize H2AX enrichment against H2B in a given chromatin state [123], we calculated total number of H2AX reads and total number H2B reads in all genomic intervals annotated by that chromatin state. The normalized enrichment of H2AX in the i^{th} chromatin state, s_i, is calculated as:

$$s_i = \frac{N_y}{N_x}\frac{\sum_{j=1}^{n} x_j}{\sum_{j=1}^{n} y_j} \qquad (2.2)$$

$\sum_{j=1}^{n} x_j$ and $\sum_{j=1}^{n} y_j$ $\sum_{j=1}^{n} x_j$ represent the total number of H2AX reads and the total H2B reads, respectively, across n genomic intervals annotated by the i^{th} chromatin state. We used this same method to normalize the Jurkat H2AX to the Input DNA that were obtained from Seo *et al.* [192].

Chapter 3: The Shared Genomic Architecture of Human Nucleolar Organizer Regions

The content of this chapter was published as part of a joint first author paper:

Floutsakou I[*], Agrawal S[*], Nguyen TT[*], Seoighe C, Ganley ARD, McStay B: The shared genomic architecture of human nucleolar organizer regions, *Genome Research*, (2013) doi: 10.1101/gr.157941.113 (accepted preprint). [*] Indicates co-first authorship.

My contribution: performed chromatin profiling and transcripts mapping.

3.1 Abstract

The human genome includes around 400 copies of a ribosomal DNA (rDNA) repeat tandemly clustered in nucleolar organizer regions (NORs) on the short arms of five acrocentric chromosomes. In human cells not all NORs are actively participating in the formation of nucleoli and even within active NORs not all copies of the rDNA repeat are transcribed. While it is known how individual rDNA repeats within an active NOR can toggle between active and silent states, the selection mechanism operating at the level of whole NORs is not understood. The acrocentric short arms are known to contain elements that direct the form and function of the whole NOR but they are still missing from the latest human genome assembly. Here we describe the genomic architecture of human NORs. We established 380 kb of the sequence distal to the NOR (termed as distal junction – DJ) and 200 kb of the sequence proximal to the NOR (termed as proximal junction – PJ). We found that both DJ and PJ are highly conserved among the five acrocentric chromosomes, suggesting they are sites of frequent recombination. The PJ sequence, similar to the regions adjacent to centromeres, has a high level of segmental duplication. Although previously believed to be heterochromatic, our integrative analysis of ChIP-seq, RNA-seq, FAIRE-seq and DNase-seq reveals that the DJ is likely to have an open chromatin structure and is actively transcribed by

RNA polymerase II. Additionally, experiments showed that the DJ is localized to the periphery of the nucleolus, where it anchors the rDNA arrays. Our findings enable study of the role of NORs in nucleolar formation and function and investigation of the link between nucleoli and human pathologies.

3.2 Introduction

The ribosome is a complex molecular machine within all living cells that serves as the primary site of protein synthesis. Ribosome biogenesis is undertaking in a distinct cellular compartment, the nucleolus. The nucleolus has distinct structure and is responsible for ribosomal RNA gene (rDNA) transcription, pre-ribosomal RNA (pre-rRNA) processing and pre-ribosome assembly [196]. Nucleoli form around clusters of repeated ribosomal DNA (rDNA) that encode rRNA, termed nucleolar organizer region (NOR). Nucleolar size is linked to cancer disease severity and nucleolar lesions have been shown to have a causative role in various cancers [197,198]. Transcriptional regulation of rDNA is associated with the tumor suppressor genes, oncogenes, and the development of malignancy [199,200,201,202,203]. Interestingly, recent studies have reported that the nucleolus also has impact on many other biological processes including cell cycle progression, aging, X chromosome inactivation and viral replication [204,205,206,207].

Most of the NORs in human cells are active and coalesce to form several nucleoli [208]. These formed nucleoli are separated from the rest of the nucleoplasm by a shell of heterochromatin that is positioned close telomeres or centromeres of the acrocentric short-arms [209,210,211]. The human genome includes around 400 copies of 43-kb rDNA repeat tandemly clustered in NORs on the short arms of five acrocentric chromosomes (chromosomes 13, 14, 15, 21, 22) [212,213]. The genomic sequence and transcriptional regulation of rDNA repeats within NORs have been well established and studied [212,214,215]. Alterations of chromatin structure are responsible for silencing about half of the rDNA in active cells [212]. Particularly, the nucleolar remodeling complex (NoRC) can repress rDNA transcription by recruiting histone deacetylase, histone methyltransferase and DNA methyltransferase activity to reset the rDNA promoter to a repressed state

[214]. With the introducing of high-throughput sequencing technology, the chromatin landscape of the rDNA repeat was fully revealed [215]. Nonetheless, the transcriptional regulation and organization of whole NOR is still unknown. One possibility is that NOR-adjacent sequences on acrocentric short arms may contain elements responsible for regulating NOR activity. Although high-throughput genomic technologies have largely boosted the whole-genome sequencing and consequently genome of many species are now complete, the DNA sequences along the length of short-arms of acrocentric chromosomes are still missing. Therefore, many aspects of nucleolar formation, organization and function and their contribution to human disease remain unknown.

Here, we set out to identify and characterize the NOR-adjacent regions. We established around 200 kb of the sequence proximal to the NOR (PJ) and 380 kb of the sequence distal to the NOR (DJ). Strikingly, both DJ and PJ sequences are highly conserved among the acrocentric chromosomes. The PJ sequence is highly repetitive and contains many segmental duplications. The DJ is a transcriptionally-active region and is localized to the periphery of the nucleolus, where it anchors the rDNA arrays. This study is a collaboration between three groups:(1) Myself and Cathal Seoighe in the School of Maths, NUI Galway; (2) Ioanna Floutsakou and Brian McStay in the Centre for Chromosome Biology, NUI Galway; (3) Saumya Agrawal and Austen Ganley in the Institute of Natural and Mathematical Sciences, Massey University, Auckland, New Zealand. The content and flow of this chapter are based on the paper that resulted from this collaboration [216]. To keep the content of this chapter concrete and easy to follow, we also briefly describe the results and figures produced by the other groups in the collaboration, declaring these contributions clearly throughout the rest of this chapter.

3.3 Results

3.3.1 Identification of rDNA flanking regions

This section is based on work carried out by collaborators.

To identify the PJ and DJ sequences, we designed probes using available DNA sequences adjacent to the rDNA arrays [217,218]. Based on these probes, we screened and sequenced some cosmid libraries and searched through GeneBank

using the resulting sequences. We indentified 15 BAC clones from the DJ and 3 BAC clones from the PJ, some of which contain rDNA sequence. Using these BAC clones, we finally indentified 379 kb of DJ DNA sequence and 207 kb of PJ DNA sequence (Figure 3.1A).

We carried out other experiments to validate the PJ and DJ sequences. We hybridized the DJ BAC clones to metaphase chromosome and found that they are indeed localized at places distal to the rDNA (Figure 3.1B). FISH experiment on DJ cosmid showed that the DJ is distal to the rDNA (Figure 3.1C). We are not able to apply similar approach to validate the PJ sequence because it has a high level of segmental duplication. We therefore applied a sequence-based method to seek for evidence of the PJ adjoining the rDNA.We sequenced PJ-containing cosmids and observed that the PJ is anchored to at least 16 kb of the rDNA.

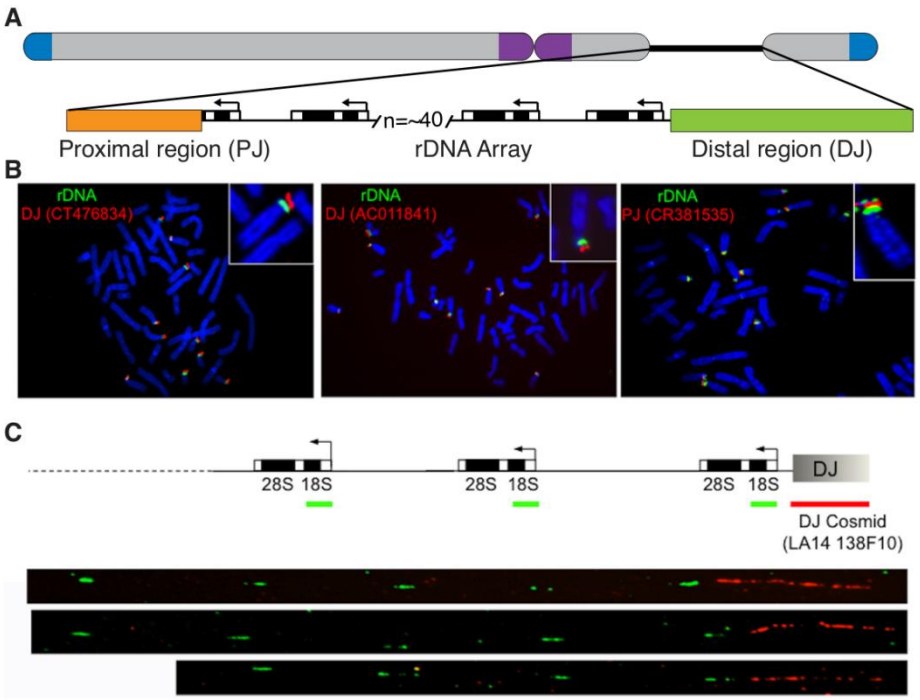


Figure 3.1: Human rDNA flanking regions. (A) Location of PJ (orange) and DJ (green) relative to telomeres (blue) and centromeres (purple), and the NOR (black line), on a human acrocentric chromosome. (B) FISH experiments show DJ and PJ localize distally and proximally to rDNA respectively on all acrocentric chromosomes. (C) DNA combing of HeLa cell nucleolar DNA shows DJ (red) is physically linked to 18S rDNA (green). This figure was produced by collaborators.

3.3.2 Interchromosomal conservation of rDNA flanking regions

This section is based on experiments carried out by collaborators.

FISH experiments showed that hybridization signals of the DJ and the PJ present in all acrocentric chromosomes (Figure 3.1B). This observation is consistent with previous findings reporting that the sequences distal to the rDNA are conserved across all acrocentric chromosomes [218,219]. To further validate the integrity of the rDNA flanking sequences, we performed PCR experiments at five locations across the DJ and observed the PCR signals at all five acrocentric chromosomes (Figure 3.2A). We also screened genomic DNA containing a chr21 translocation and confirmed the orientation of the PJ and DJ relative the rDNA (Figure 3.2A).

We found high level of identities of DJ and PJ sequences in the five acrocentric chromosomes. Intra-chromosomal sequence identities for both DJ and PJ are close to 100% as expected (Figure 3.2B). The DJ inter-chromosomal identity is also very high (~99.1%), suggesting there is an active homogenization mechanism that maintains DJ sequence identity between the acrocentric chromosomes (Figure 3.2B). The PJ inter-chromosomal identity is lower (~93.3%), likely arising from inter-chromosomal genomic variants in the rDNA junction position and Alu elements (Figure 3.2B).

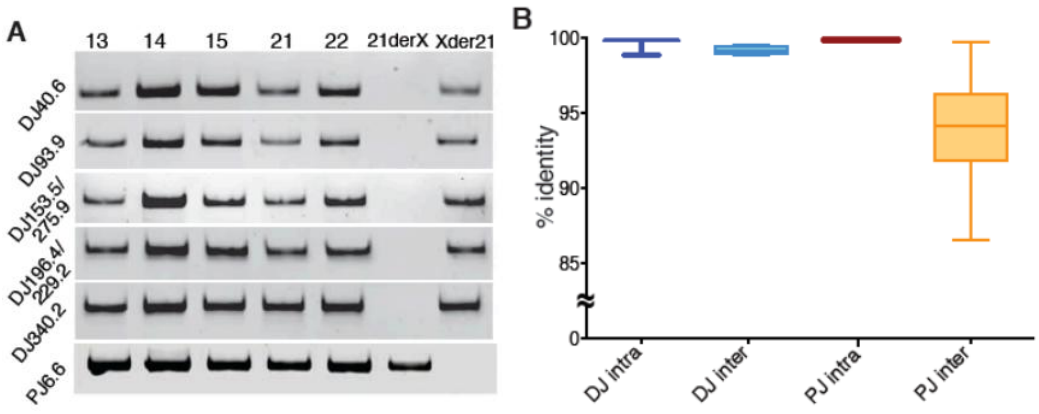


Figure 3.2: DJ and PJ acrocentric chromosome conservation. (A) PCR at five locations of the DJ on all five acrocentric chromosomes and on the reciprocal products (Xder21 and 21derX) of a chr21 translocation that originates in the rDNA. Bottom panel is for the single unique PJ region. (B) Average intra-chromosomal

and inter-chromosomal DJ and PJ sequence identities from pairwise comparisons of representative BAC and cosmid clones. This figure was produced by collaborators.

3.3.3 Localization and role of the DJ in nucleolar architecture

This section is based on experiments carried out by collaborators.

We performed 3D-immuno FISH for the DJ and found that the DJ sequences formed as separated foci, with majority localizing within the perinucleolar heterochromatin (Figure 3.3A). We cannot duplicate the same experiment for the PJ due to its high level of segmental duplication, so we will focus on studying the DJ for the rest of this study. To further look into the association between the DJ and nucleolar architecture, we inhibit rDNA transcription by introducing actinomycinD (AMD). This leads to a collapse of the rDNA repeats into nucleolar caps surrounding nucleolar periphery. Interestingly, as a consequence of AMD introduction, the nucleolar caps form immediately adjacent to the DJ foci within the perinucleolar heterochromatin, rather than the DJ foci moving towards rDNA foci within the nucleolar interior (Figure 3.3B). This suggests that the DJ sequence might be involving in its perinucleolar localization.

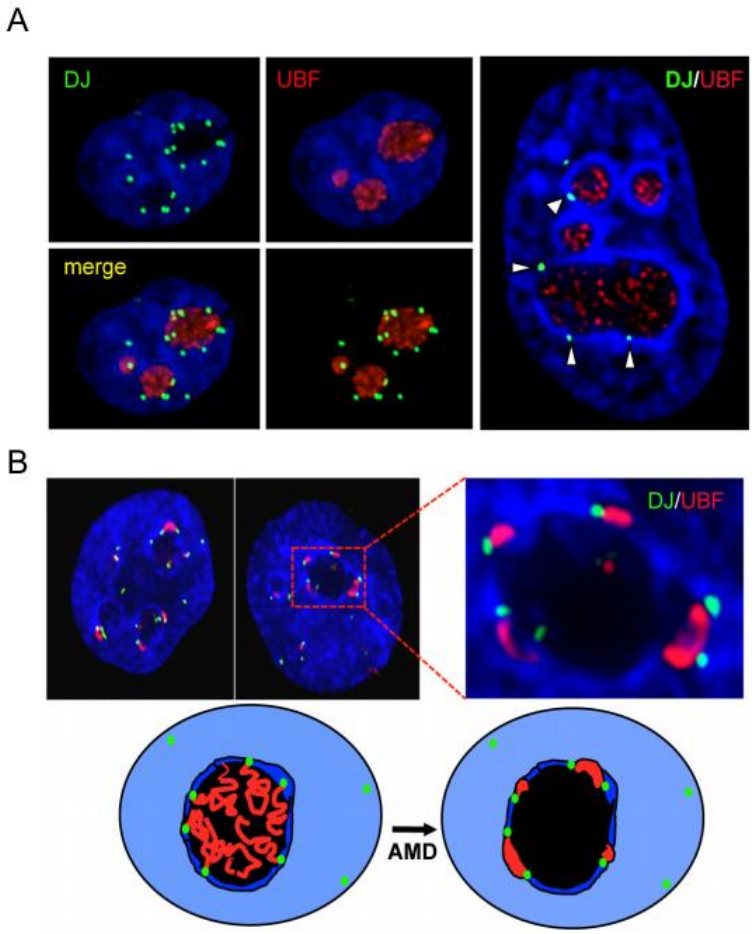
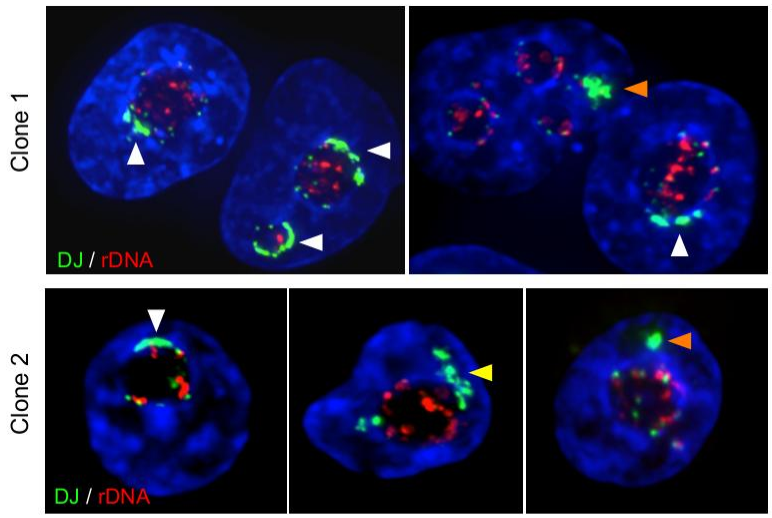


Figure 3.3: The DJ forms a perinucleolar anchor for rDNA repeats. (A) 3D-immuno FISH experiments show that DJ sequences are localized within perinucleolar heterochromatin. (B) Inhibition of rDNA transcription with AMD results in formation of nucleolar CAPs juxtaposed with DJ sequences in perinucleolar heterochromatin. Two representative cells are shown, one with an enlargement. Cartoon models the transition between active and withdrawn rDNA upon AMD treatment. rDNA (red) retreats from the nucleolus (black) to the DJ (green) that is embedded in perinucleolar heterochromatin (dark blue). This figure was produced by collaborators.

To assess the involvement of the DJ in its perinucleolar localization, we transfected cells with the DJ BACs that we described above (section 3.3.1). We selected two stable clones containing the BACs and performed the 3D-FISH in cells treated with AMD. Strikingly, the DJ arrays appear to strongly associate with perinucleolar heterochromatin for both of the clones (Figure 3.4). Further, the DJ

arrays cover a large fraction of nucleolar surface. These results indicate that the DJ specify association with perinucleolar heterochromatin.



	Associated ◁	Partial ◁	Non-associated ◁
Clone 1 cells n = 41	68%	27%	5%
Clone 2 cells n = 50	74%	18%	8%

Figure 3.4: Ectopic DJ arrays target perinucleolar heterochromatin. Positioning of DJ arrays. 3D-FISH was performed on AMD treated cells with rDNA (red) and DJ BAC CT476834 (green) probes. The large green hybridization signals identified by arrowheads indicate the ectopic DJ array. Endogenous DJ signals are also visible. The table below illustrates the degree of association between DJ arrays and nucleolar perinucleolar heterochromatin. This figure was produced by collaborators.

3.3.4 Chromatin profiling of the DJ

The above findings led us to a key hypothesis that the DJ sequences serve as an anchor point within perinucleolar heterochromatin for the linked rDNA array in the nucleolar interior. We hypothesized further that the chromatin structure of the DJ region could reveal functional elements that support this role. To profile the chromatin structure of the DJ region, we carried out integrative analysis of ChIP-seq, FAIRE-seq and DNase-seq data available from ENCODE project [123,130] in seven different cell types including GM12878, H1-hESC, HMEC, HSMM, K562,

NHEK and NHLF (see the Methods). FAIRE-seq (Formaldehyde Assisted Isolation of Regulatory Elements) employs formaldehyde fixation and phenol-chloroform extraction to enrich nucleosome-free regions [220]. DNase-seq uses the DNaseI enzyme to digest nucleosome-depleted sites, also referred to as DNaseI hypersensitive (HS) sites [221,222]. Interestingly, the FAIRE and DNaseI-HS signals illustrate open chromatin (nucleosome-free) regions at regular ~45 kb intervals across the DJ. This open chromatin picture is similar among different cell types (Figure 3.5).

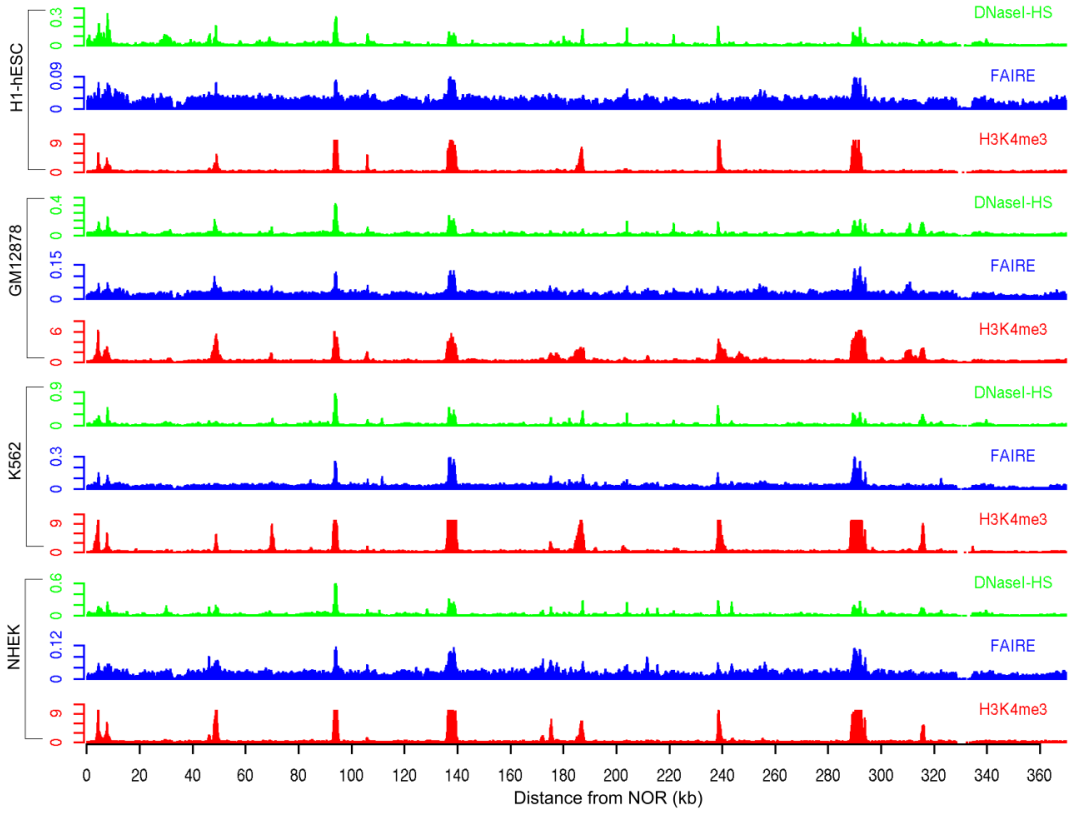


Figure 3.5: Distribution of open chromatin marks and H3K4me3 across the DJ in four different cell types.

To fully characterize the chromatin landscape of the DJ, we analyzed ChIP-seq data of nine histone modifications, seven generally associated with transcriptional activation (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H4K20me1 and H3K36me3) and two generally associated with transcriptional repression (H3K9me3 and H3K27me3) (see Table 3.1 for more details). Interestingly, we found chromatin signatures characteristic of promoters – as

evidenced from the peaks of an active promoter mark, H3K4me3 – coinciding with the open chromatin peaks, (Figure 3.5). The other four active chromatin marks (H3K9ac, H3K27ac, H3K4me1, and H3K4me2) are enriched at these promoter regions in different cell types, further implicating that the DJ contains regions of active chromatin (Figure 3.6A). Consistent with this result, the heterochromatin mark (H3K9me3) and Polycomb-repressed mark (H3K27me3) have very few peaks across the DJ compared to the active chromatin marks.

Table 3.1: Summary of 9 chromatin marks used to profile chromatin in the DJ

Chromatin mark	Full name	Associated with	REFs
H3K4me1	H3 lysine 4 monomethylation	Enhancer	[57,113]
H3K4me2	H3 lysine 4 dimethylation	Enhancer; promoter	[57]
H3K4me3	H3 lysine 4 trimethylation	Promoter	[223,224]
H3K36me3	H3 lysine 36 trimethylation	Transcribed region	[57,224]
H4K20me1	H4 lysine 20 monomethylation	Transcribed region	[57,224]
H3K9ac	H3 lysine 9 acetylation	Active regulatory region	[225]
H3K27ac	H3 lysine 27 acetylation	Active regulatory region	[225]
H3K9me3	H3 lysine 9 trimethylation	Heterochromatin	[57,223]
H3K27me3	H3 lysine 27 trimethylation	Polycomb-repressed	[57,223]

Chromatin marks associated with actively transcribed gene bodies (H3K36me3, H4K20me1) are observed extending leftward and rightward from the promoters at 187 kb and 238 kb respectively (Figure 3.6A), suggesting that transcription activities are present in the DJ. Indeed, a previous study by Guttman *et al.* observed that genes actively transcribed by RNA polymerase II (RNA Pol II) are marked by H3K4me3 at their promoter and H3K36me3 along the transcribed region [120].

To integrate datasets from all of the chromatin marks and profile an overall chromatin landscape of the DJ, we carried out a multivariate Hidden Markov Model analysis using ChromHMM [129]. We found that chromatin landscape of the DJ is largely conserved among different cell types; particularly for the promoter regions and transcription regions (Figure 3.6BC). Our experimental validation confirmed that the H3K4me3 and FAIRE peaks are present in HT1080 cell line (Figure 3.6D).

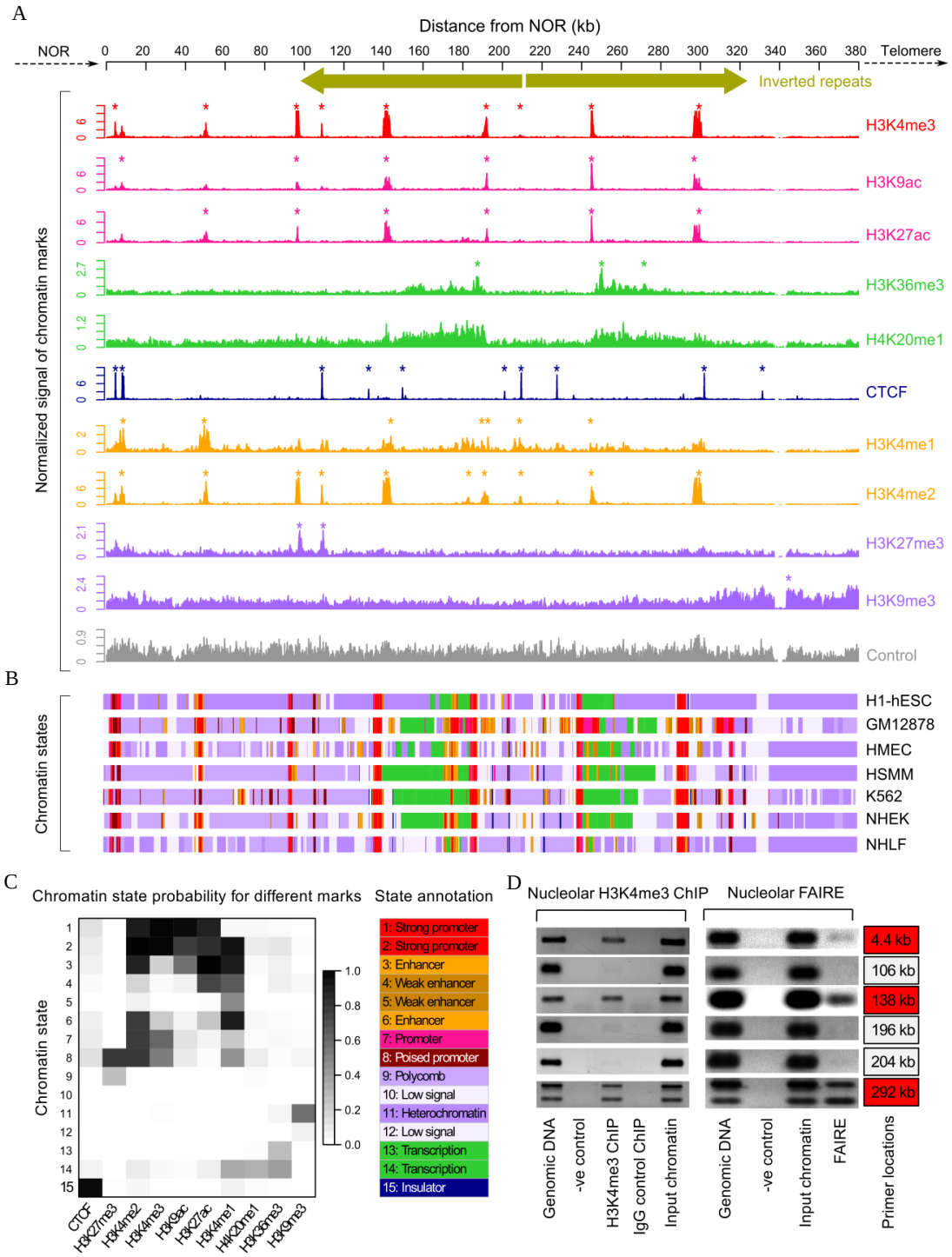


Figure 3.6: Chromatin landscape of the DJ. (A) ChIP-seq signals of different chromatin features (indicated on the right) in H1-hESC cells, normalized to tags per million mapped reads are shown below a schematic representation of the DJ, including the inverted repeats. Asterisks indicate enrichment sites. The control signal is shown in gray (bottom). (B) Chromatin states derived from the

multivariate HMM analysis for seven different cell types (indicated on the right). Each colored bar represents a specific chromatin state, as annotated in (C). (C) The chromatin state probabilities for different marks outputted from the HMM analysis are shown on the left. The mapping between chromatin states and known genomic features is shown on the right. (D) Nucleolar H3K4me3 ChIP-PCR and nucleolar FAIRE-PCR performed by McStay's lab validates the presence of H3K4me3 and FAIRE in the DJ. DJ positions of the primers used are shown to the right, and red boxes correspond to peaks of H3K4me3 from (A). Genomic DNA (gDNA), input and negative controls (-ve and IgG) are shown alongside the treatments.

CTCF is a multivalent DNA binding protein involved in many cellular processes including transcriptional regulation, insulator activity, recombination, and regulation of chromatin architecture [78]. Recently, CTCF has been shown to be involved in the transcriptional regulation of ribosomal genes [226] and nucleolar organization in human cells [227]. Here, we mapped CTCF ChIP-seq data obtained from the ENCODE project and found sharp peaks of CTCF across the DJ that are positioned close to the DJ/rDNA boundary and frame the DJ gene bodies described above (Figure 3.6A). A motif analysis for sequences of the CTCF peaks revealed a motif matched the CTCF model of JASPAR [228], supporting the existence of CTCF binding in the DJ (Figure 3.7).

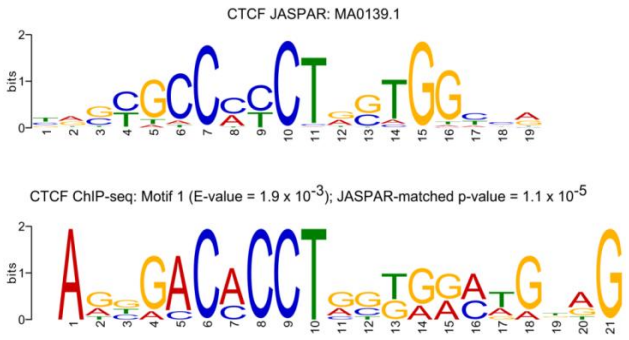


Figure 3.7: Occurrence of CTCF DNA motifs within the DJ. Upper panel shows a known motif model of CTCF from JASPAR. Lower panel shows a CTCF motif found from ChIP-seq data. The Y-axis shows the information content of the position, in bits.

3.3.5 Transcription profiling of the DJ

The observed chromatin landscape suggests that despite being embedded in perinucleolar heterochromatin, the DJ is transcriptionally active. We first investigated the distribution of the key proteins that are involved in transcriptional regulation such as RNA Pol II and TAF1. During the transcription initiation, TATA-binding protein (TBP) targets most promoters as part of the multi-subunit TFIID complex that includes TBP-associated factors (TAFs). Together these proteins recruit RNA Pol II and other transcription factors to the transcription start sites (TSSs) of genes to start the transcription process [68]. We mapped the TAF1 and RNA Pol II ChIP-seq data obtained from the ENCODE project to the DJ and found that peaks of TAF1 and RNA Pol II coincide with the DJ promoters marked by H3K4me3 peaks (Figure 3.8).

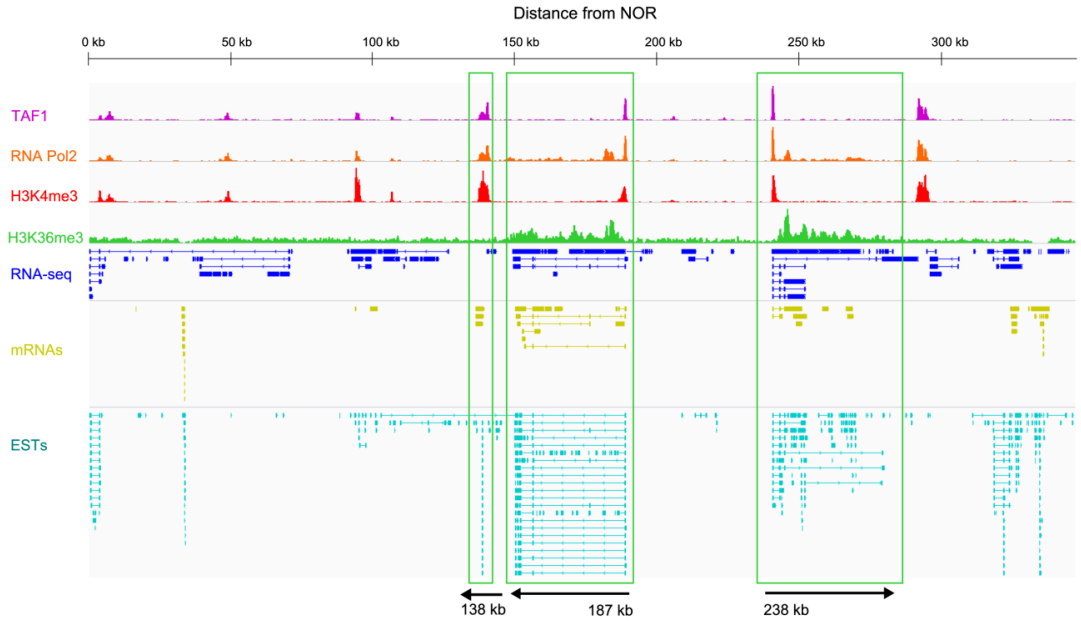


Figure 3.8: Comprehensive transcriptome profiling of the distal junction. The top four tracks show ChIP-seq signals of four chromatin marks (TAF1, RNA Pol II, H3K4me3, and H3K36me3) in H1-hESC cells; and the next track shows the structure of all DJ transcripts assembled from RNA-seq data in H1-hESC cells. Mapping of the mRNA data and EST data (obtained from GenBank) onto the DJ is shown in the two bottom tracks.

To investigate transcriptional activity in the DJ we searched for genes across the DJ using different data including mRNA-seq (from the Caltech ENCODE project), mRNA and Expressed Sequence Tags (EST) (from GenBank) (Figure 3.8) (see the Methods). Navigating to the putative promoters at 187 kb and 238 kb, we found multiple lines of evidence strongly supporting transcriptional activity in the DJ (Figure 3.9A). Mapping of the RNA-seq data indicates that the two transcripts at these promoters are spliced, and their cDNA clones are present in GenBank (accession numbers AK026938 and BX647680). Moreover, we estimated their relative expression levels using RNA-seq data with Cufflinks [229] (see the Methods), and found that these transcripts are expressed at low to medium levels (Figure 3.9B). Our RT-PCR experiments confirmed the existence of these spliced polyadenylated transcripts, which we termed disnor187 and disnor238 (Figure 3.9C). The largest open reading frames present in disnor187 and disnor238 are 120 and 144 amino acids respectively. Their size, chromatin signatures K4-K36 and limited coding capacity suggest that they can be long non-coding RNAs (lncRNAs) [120].

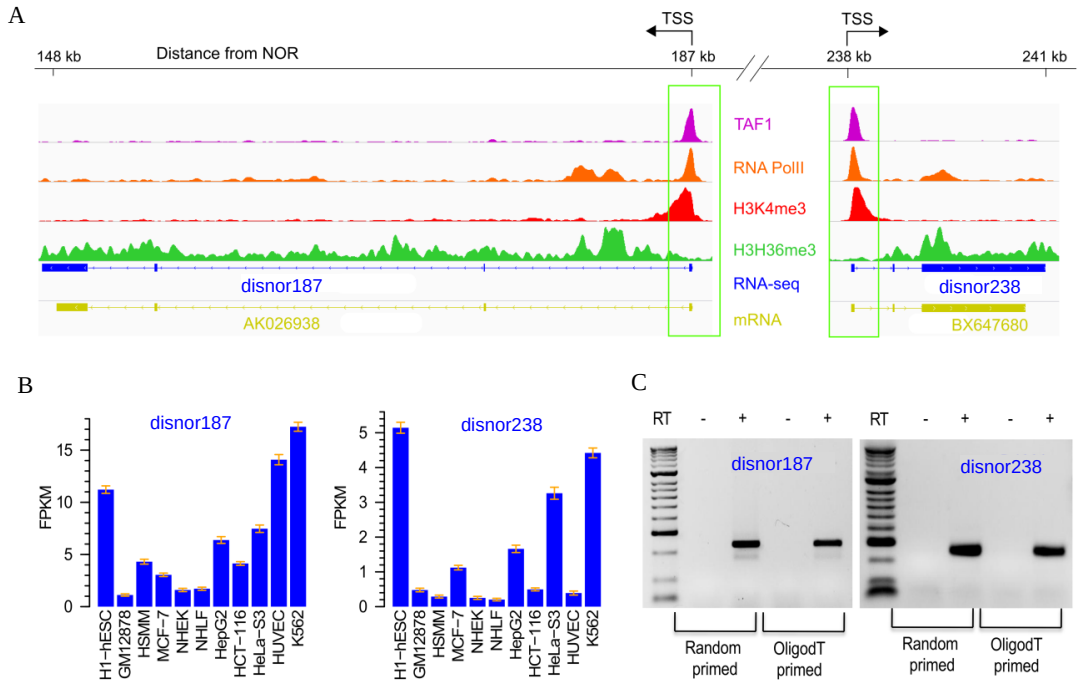


Figure 3.9: Transcription profiling of DJ transcripts. (A) ChIP-seq reveals chromatin features consistent with transcription originating from promoters at 187 kb and 238 kb in the DJ. The top four tracks represent an expansion of selected

chromatin features from Figure 3.6A. The bottom two tracks show RNA-seq and cDNA mapping results. These identify spliced transcripts (disnor187 and disnor238) that are similar to cDNA clones AK026938 and BX647690. Exons are indicated by blocks. (B) Quantitation of DJ transcript levels. Transcript abundances for disnor187 and disnor238 were estimated from RNA-seq data, measured as FPKM (Fragments Per Kilobase of transcript per Million mapped reads) from a variety of different cell types (bottom). This shows that disnor187 and disnor238 are transcribed at low to moderate levels. The error bars show the 95% confidence interval of the FPKM. (C) RT-PCR using primers to detect the disnor187 and disnor238 transcripts in HT1080 cells. Products are observed for random and oligoT-primed reverse transcription, showing the transcripts are polyadenylated, and they are of the expected sizes for spliced transcripts.

3.4 Discussion

Here, we established 580 kb of sequence from the regions flanking rDNA array and carried out deep analyses of these regions. We found that the sequences both proximal and distal to the rDNA are conserved and show high level of identity among five acrocentric chromosomes. The structure of the proximal sequences is similar to the sequences bordering centromeres that are very repetitive and contain large numbers of segmental duplicates [230]. These results strongly suggest that the PJ is a region of active and frequent recombination. Further, this recombination occurs predominantly with other centromeric regions in the genome, indicating the co-localization of these regions might be involved in translocations arising from the acrocentric short arms [231].

The sequences distal to rRNA are, surprisingly, highly unique and contain functional elements – like those in transcription-active regions. These sequences form discrete foci localizing within perinucleolar heterochromatin where they anchor the rDNA array to this region. Although these regions are packed within perinucleolar heterochromatic shell, they are euchromatic and have distinct genomic characters. The DJ region is, in fact, transcriptionally active and has an open chromatin landscape. Interestingly, the DJ contains of polyA and spliced transcripts. In particular, the two transcripts disnor187 and disnor238 starting at 187 kb and 238

kb are potential candidates for NOR regulation. We propose that the DJ acts as a master regulator of the entire NOR where it determines the transcriptional status of the linked rDNA array. The active NORs may be localized to perinucleolar heterochromatin where they form nucleoli, while inactive NORs may lose this association and form silent arrays that do not participate in nucleoli.

The identification of the PJ and DJ sequences has provided insights into understanding nucleolar biology and started to unravel a big part of the human genome that are still missing - the short arms of the acrocentric chromosomes. The DJ can help designing experiments to check the evidence of multiple NORs in human nucleoli – a major interest of the nucleolar biology community but remains difficult to assess. The high level of segmental duplication of the PJ likely suggests that nucleolus-associated chromatin domains indentified previously might be segmentally duplicated [232]. Derenzini *et al.* [233] reported that many types of cancers are associated with the heterogeneity in nucleolar morphology, suggesting direct transcriptional regulation of rRNA genes is not responsible for the development of malignancy. We proposed that genetic changes in the short arms of acrocentric chromosomes can lead to various cancers and diseases, as these short arms largely determine the nucleolar form and function. Therefore the findings of the NOR flanking sequences plus their characteristics help understanding the roles that genetic and epigenetic alterations in the DJ and PJ play in human disease.

3.5 Methods

Data and methods from other groups have been described in detail in our paper [216]. Here, we only report details of the data and methods we used for chromatin profiling and transcriptome profiling for the DJ.

3.5.1 Data

ChIP-seq data for ten chromatin marks (CTCF, H3K4me1, H3K4me2, H3K4me3, H3K36me3, H3K9me3, H3K27me3, H3K9ac, H3K27ac and H4K20me1) and Input were obtained for seven different cell types (GM12878, H1-hESC, HMEC, HSMM, K562, NHEK and NHLF) from ENCODE Broad Histone [123]. DNase-seq and FAIRE-seq data were obtained from ENCODE UNC/Duke [234]. PolyA tailed RNA-seq data were obtained for 11 different cell types

(GM12878, H1-hESC1, HCT-116, HeLa-S3, HepG2, HSMM, HUVEC, K562, MCF-7, NHEK, NHLF) from ENCODE Caltech RNA-seq. GenBank mRNAs and ESTs data were downloaded from the UCSC genome browser [235] on 20/01/ 2012. These data were mapped to the human genome to which the DJ sequences had been added, as described in the Methods. More information about these data is described in Table 3.2.

Table 3.2: Datasets used to profile the chromatin and transcripts in the DJ

Data	Details	Source	Data type
Histone modifications	Histone modifications that mark different chromatin states, such as promoter (H3K4me3), heterochromatin (H3K9me3)	Broad Histone/ENCODE	ChIP-seq
Transcription factors	The insulator CTCF, RNA-Pol II, transcription initiation factor TAF1 and other factors that indicate transcriptional activity.	Yale TFBS/ENCODE	ChIP-seq
FAIRE	Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) mark open chromatin regions	UNC FAIRE/ENCODE	FAIRE-seq
DNaseI HS	DNaseI hypersensitive sites (DNaseI HS) mark open chromatin regions.	Duke DNaseI HS/ENCODE	DNase-seq
Transcripts	ESTs and cDNAs data to find evidence of transcription.	UCSC Genome Browser	RNA sequences

Transcripts	RNA-seq data to assemble the DJ transcriptome.	Caltech RNA-seq/ENCODE	RNA-seq

3.5.2 Chromatin profiling

Mapping

The short read length (36 bp for ChIP-seq and 75 bp for RNA -seq) means it is likely that many reads will map to both the DJ and the human genome. To map the reads uniquely to the DJ, we created a custom human genome that includes the latest human assembly hg19 and DNA sequences of the DJ and human rDNA repeat (extracted from BAC clone RP11-337M7 , GenBank accession number AL592188). We mapped the ChIP-seq data onto this custom genome using bowtie (version 0.12.7) [137] with parameters -l 34 -n 2 -a --best --strata -m 1 to take into account of sequencing quality and to yield the best mapping rate. The options "-l 34 -n 2" indicate that a maximum of two mismatches are allowed in the first 34 bases (referred to as the seed) of each read. The options "-a --best --strata" report only those alignments in the best alignment stratum. Further, specifying "-m 1" causes bowtie to report those reads having unique alignment in the best stratum. Potential alignment duplicates were removed from the alignments using Picard (http://picard.sourceforge.net/). Mapped reads from all replicates for each chromatin mark were combined to obtain the highest read coverage.

Signal profiling

For each mapped read, we created a tag by extending 200 bp (the known expected fragment size) from the 5' end towards 3' end of the read. We then calculated the number of tags overlapping each base (also known as coverage depth) across the custom genome. This coverage depth was then normalized per million total mapped reads. Finally, to smooth the signal profile, we ran a sliding 200 bp window (with step size of 10 bp) across the DJ and calculated the average normalized coverage depth for each window. Signal profiles for open chromatin

markers (FAIRE and DNaseI) were processed using F-Seq [236] following a published procedure [234].

Peak calling

We used the peak calling software MACS [143] with mostly default parameters (-g hs --nomodel --shiftsize 100 -p 1e-5) to identify enriched regions of each chromatin mark across the custom genome. Since the expected fragment size is known we did not apply shifting model in MACS (--nomodel), but instead applied a shift size of 100 bp (--shiftsize 100) that is a half of the fragment size. We ran MACS for each chromatin mark in each cell type separately and used the corresponding Input DNA as a control.

Chromatin state analysis

To depict the chromatin landscape from the combination of chromatin marks, the multivariate Hidden Markov Model (HMM) software, ChromHMM, was used with default parameters [129] to segment the custom genome into different chromatin states. The seven cell types we used were chosen because these have comprehensive data for all ten chromatin marks. We ran the HMM model with 15 states as this number is enough to characterize the whole human genome [123].

3.5.3 Transcriptome profiling

Paired-end RNA-seq data from the 11 cell types was mapped to the custom genome using Tophat (v1.2.0) [157] with mostly default parameters (-r 50 -a 8). We then merged the output alignments from all replicates using samtools [237]. The result is used as the input for Cufflinks (v1.3.0), with default parameters, to assemble the transcriptome of the custom genome. Finally we merged all 11 assembled transcriptomes, corresponding to the 11 cell types, using Cuffmerge [229] to obtain the final transcriptome. We also used this final transcriptome to estimate the abundance of the DJ transcripts. We used BLAT [238] to map the mRNA and EST data to the DJ using the parameters "-fine -q=rna -minIdentity=95 -maxIntron=70000", and "-minIdentity=97 -maxIntron=70000", respectively. Specifying "-minIdentity=95" causes BLAT to report those alignments having at least 95% of the identity level. We set higher minimum identity level when mapping the EST data (97%) because they normally contain shorter sequences. The

option "-maxIntron=70000" reports only alignments with intron length less than 70 kb.

Chapter 4: Integrative Analysis of mRNA Expression and Half-life Data Reveals *Trans*-acting Genetic Variants Associated with Increased Expression of Stable Transcripts

The content of this chapter was published as:

Nguyen TT, Seoighe C (2013) Integrative Analysis of mRNA Expression and Half-Life Data Reveals *Trans*-Acting Genetic Variants Associated with Increased Expression of Stable Transcripts. PLoS ONE 8(11): e79627. doi:10.1371/journal.pone.0079627

4.1 Abstract

Genetic variation in gene expression makes an important contribution to phenotypic variation and susceptibility to disease. Recently, a subset of *cis*-acting expression quantitative loci (eQTLs) has been found to result from polymorphisms that affect RNA stability.

Here we carried out a search for *trans*-acting variants that influence RNA stability. We first demonstrate that differences in the activity of *trans*-acting factors that stabilize RNA can be detected by comparing the expression levels of long-lived (stable) and short-lived (unstable) transcripts in high-throughput gene expression experiments. Using gene expression microarray data generated from eight HapMap3 populations, we calculated the relative expression ranks of long-lived transcripts versus short-lived transcripts in each sample. Treating this as a quantitative trait, we applied genome-wide association and identified a single nucleotide polymorphism (SNP), rs6137010, on chromosome 20p13 with which it is strongly associated in two Asian populations ($p = 4 \times 10^{-10}$ in CHB – Han Chinese from Beijing; $p = 1 \times 10^{-4}$ in JPT – Japanese from Tokyo). This SNP is a *cis*-eQTL for *SNRPB* in CHB and JPT but not in the other six HapMap3 populations. *SNRPB* is a core component of the spliceosome, and has previously been shown to affect the expression of many RNA processing factors.

We propose that a *cis*-eQTL of *SNRPB* may be directly responsible for inter-individual variation in relative expression of long-lived versus short-lived transcript in Asian populations. In support of this hypothesis, knockdown of *SNRPB* results in a significant reduction in the relative expression of long-lived versus short-lived transcripts. Samples with higher relative expression of long-lived transcripts also had higher relative expression of coding compared to non-coding RNA and of RNA from housekeeping compared to non-housekeeping genes, due to the lower decay rates of coding RNAs, particularly those that perform housekeeping functions, compared to non-coding RNAs.

4.2 Introduction

RNA stability plays a major role in gene expression regulation in virtually all organisms, from bacteria to mammals [239,240,241]. Indeed, steady-state gene expression levels represent the equilibrium of two opposing biological processes: RNA transcription and RNA decay. Changes in gene expression levels can result from alteration in either of these processes [239,242]. Recent studies have investigated RNA stability using high-throughput techniques in diverse organisms, from yeast [243,244] to *Arabidopsis* [245], mouse [246,247,248], and human [248,249,250,251,252], and for both coding and non-coding RNAs [247,253]. Several of these studies have reported strong correlations between RNA stability and steady-state gene expression levels. In addition, RNA stability has been shown to be related to physiological function [246,250]. For example, genes encoding proteins involved in housekeeping functions tend to have stable mRNAs [248,253]. The modulation of RNA stability can, in turn, have a major impact on cellular processes, including proliferation, differentiation, and adaptation to environmental stimuli [239,240,241]. Dysregulation of RNA stability has been linked to several human diseases, such as chronic inflammation [254], cardiovascular disease and cancer [28,255,256].

The regulation of RNA stability is achieved through interactions between *trans*-acting RNA-binding proteins and *cis*-acting elements within RNAs [257,258]. Among RNA-binding proteins, heterogeneous nuclear ribonucleoproteins (hnRNPs) are key factors that regulate major steps of gene expression, including pre-mRNA

processing, RNA stability, and translation [259,260,261]. For example, *HNRNPA2B1*, a member of the hnRNP family, was found to stabilize a large number of target transcripts carrying a conserved structural RNA element in the 3' untranslated regions [251]. Knockdown of *HNRNPA2B1* resulted in a remarkable increase in the relative decay rate of the target transcripts and, consequently, a significant decrease in their expression levels [251]. The contribution of RNA decay to gene expression levels was also investigated in a recent study where a subset of *cis*-acting expression quantitative loci (*cis*-eQTLs) was found to be a consequence of variation in decay rates [262]. A moderate number of genetic variants were found to significantly associate with inter-individual variation in both gene expression and RNA decay, for which variation in RNA decay could explain the association with gene expression level [262]. Despite increased appreciation of the role of RNA stabilization in determining gene expression levels there has been no investigation of *trans*-acting genetic variants that affect the stabilization of RNA.

Here we investigate factors that affect RNA stability in *trans*. We first show that perturbation of RNA stabilization factors that affect multiple genes can be inferred from gene expression data. Given a dataset of RNA decay rates and expression levels, we define the RNA stability score (RS-score), based on the expression of long-lived transcripts relative to short-lived transcripts. Knocking down *HNRNPA2B1*, which has been shown to be involved in stabilization of a large proportion of RNAs [251], leads to a significant reduction in the RS-score. Using gene expression microarray data generated from eight HapMap3 populations [263], we identified a SNP, rs6137010, on chromosome 20p13 that is strongly associated with the RS-score in Asian populations. This SNP is a *cis*-eQTL of *SNRPB*, a gene that encodes a core component of the spliceosome and has been shown to modulate the expression of many RNA processing factors [264]. The C allele of rs6137010 is associated both with higher expression of *SNRPB* and higher RS-score. Knockdown of *SNRPB* results in a significant decrease in the RS-score, suggesting that the *cis*-eQTL for *SNRPB* is responsible for the observed genetic variation in RS-score in Asian populations.

4.3 Results and Discussion

4.3.1 Perturbation of RNA stabilization is detectable from expression data

We hypothesized that changes in the activity of *trans*-acting factors that are involved in stabilizing multiple RNAs could be detectable by analyzing gene expression profiles. To test this hypothesis we obtained gene expression data from a published study in which the heterogeneous ribonucleoprotein, *HNRNPA2B1,* was knocked down [251]. In the original study this gene was shown to play a role in the stabilization of RNAs containing an abundant structural motif and RNAs containing this motif were downregulated in the knockdown samples compared to controls [251]. However, even in the absence of knowledge of the specific *trans*-acting factor and target RNAs involved it is possible to infer the effects of the knockdown on RNA stability. This is because stable, long-lived transcripts are enriched among the genes that are targeted by *HNRNPA2B1* [251].

We divided genes into two groups by using RNA decay rate data from Goodarzi *et al.* [251]. The first group contains genes expressing long-lived RNAs (decay rate lower than the mean across genes) and the second group contains genes expressing short-lived RNAs (decay rate higher than the mean). We then defined the RS-score for a sample as the difference in the mean expression rank between these two groups of genes in the sample (see Methods for more details). A higher RS-score implies relatively higher expression levels of long-lived or stable RNAs. A similar idea has previously been used to infer the impact of miRNA regulation on target genes using gene expression data [265]. The regulatory effect score (RE-score) of a miRNA was defined as the difference in the mean expression rank between targets of the miRNA and non-targets. A higher RE-score indicates lower expression levels of target genes and, thereby, a stronger effect of the corresponding miRNA. Analogously, a higher RS-score implies that the long-lived RNAs that are more likely to be subject to stabilization by *trans*-acting factors are relatively more highly expressed in a sample.

The RS-score of the *HNRNPA2B1* knockdown was significantly lower than RS-score of the control in three independent replicates (p=3.7 x 10^{-3}; paired t test) (Figure 4.1). This is consistent with expectations because *HNRNPA2B1* is one of the heterogeneous nuclear ribonucleoproteins that influence pre-mRNA processing

and other aspects of RNA metabolism and transport. More importantly, *HNRNPA2B1* is involved in stabilizing a large number of genes, particularly genes expressing long-lived RNAs, by binding to a structural RNA motif of target genes [251]. *HNRNPA2B1* knockdown caused a significant reduction in the expression levels of long-lived RNAs (Figure S1 in Appendix A), resulting in lower RS-scores in the knockdown samples. These observations suggest that gene expression levels can be used to infer the effects of *trans*-acting factors that are involved in stabilizing large numbers of genes.

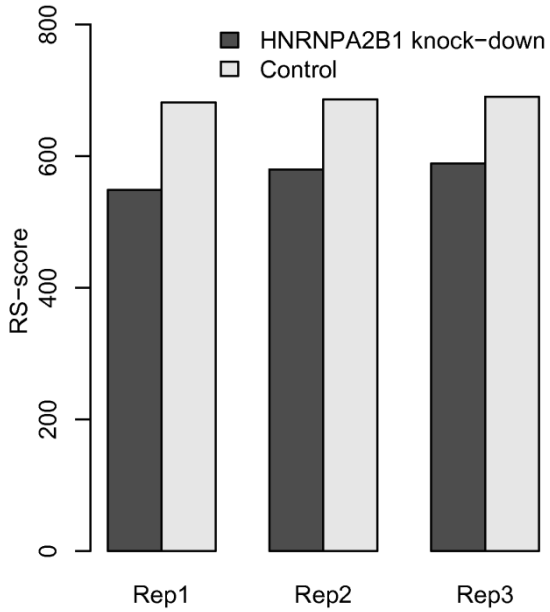


Figure 4.1: *HNRNPA2B1* knock-down results in reduced RS-score. RS-score was calculated for *HNRNPA2B1* knockdown samples and control samples separately in three independent replicates (Rep1, Rep2, and Rep3).

4.3.2 The genetics of *trans*-acting factors that affect RNA stability

We obtained gene expression data generated from lymphoblastoid cell lines of 726 individuals in eight HapMap3 populations [263] (Table S1 in Appendix A). Using the half-life data from HeLa cells [253], we calculated the RS-score for each of these individuals (see Methods). Interestingly, the RS-score was well correlated with the expression level of *HNRNPA2B1* in most of the populations (Table S2 in Appendix A), with the strongest correlation in CHB (Spearman *rho* = 0.48; p = 8.4 x 10^{-6}). Because the experimental knock down of *HNRNPA2B1* results in a

reduction in the RS-score, we hypothesized that *cis*-eQTLs affecting the expression level of *HNRNPA2B1* should also be associated with RS-score. This is the case for four *cis*-eQTLs of this gene in two of the HapMap3 populations (Table S3 in Appendix A).

To search more generally for genetic variants associated with the RS-score we used a genome-wide association study (GWAS) approach, treating the RS-score as a quantitative trait. We carried out additive tests of association between single nucleotide polymorphisms (SNPs) genotyped as part of the HapMap3 project and the RS-score in each population separately (see the Methods section for more details). We found one strong association between a SNP, rs6137010, on chromosome 20p13 and RS-score in the CHB population ($p = 4.4 \times 10^{-10}$; Figure 4.2). Interestingly, this association is replicated in the other Asian population – JPT ($p = 1.2 \times 10^{-4}$). We used a label permutation procedure to check the robustness of this result to failures in modelling assumptions (see Methods). The association between rs6137010 and RS-score in CHB was stronger than the best associations in each of 1,000 label permutations. Furthermore, the Bonferroni-adjusted p-value of this association is very significant (Bonferroni $p = 5.9 \times 10^{-3}$). Therefore, the association between rs6137010 and RS-score in CHB is robust, genome-wide significant, and replicated in a second population (JPT).

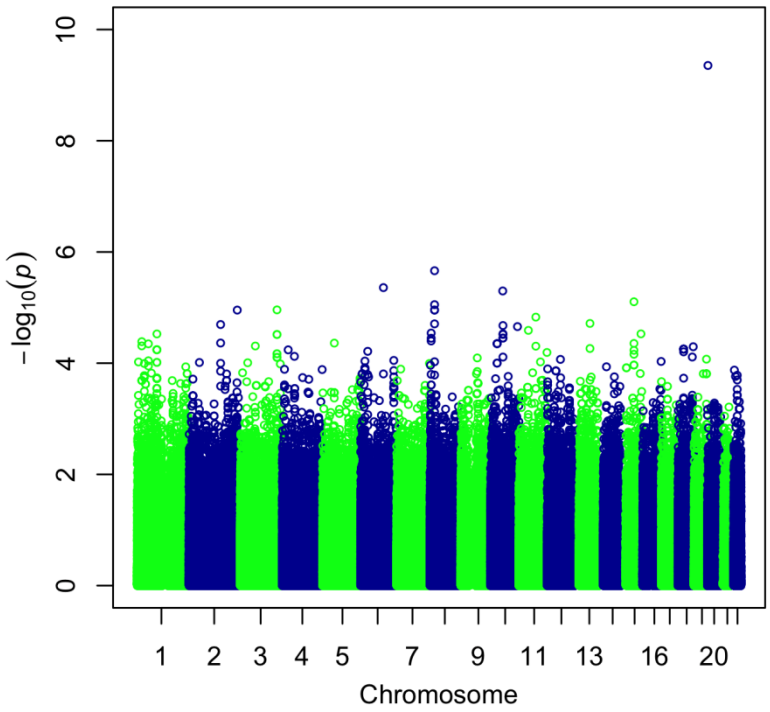


Figure 4.2: Manhattan plot for GWA with RS-score in CHB. The plot shows -$\log_{10}$ of P-values from tests of association between individual SNP markers and the RS-score. Successive chromosomes are shown in different colors.

To increase the statistical power of the association tests, we combined individuals from different populations. Because different populations have different ancestries combining individuals from these populations can lead to spurious associations, resulting from structure in the combined population. To tackle this problem, we applied a principal components analysis (PCA) approach [266] (see Methods for more details) to model ancestry differences among all 726 individuals. In a scatter plot of the first and second principal components (Figure S2 in Appendix A) three broad clusters are evident, consisting of the African populations, the Asian populations and CEU, MEX, GIH. Given these clusters, we considered four ways of combining populations: CHB + JPT (Asian populations), YRI + MKK + LWK (African populations), CEU + GIH + MEX, and finally all 8 populations (ALL). For each combination, we performed a principal components analysis and included the first five principal components as covariates in the GWAS regression models (see Methods). The SNP rs6137010 was strongly associated with the RS-score in CHB + JPT (p = 2.0 x 10^{-12}; Figure S3 in Appendix A). This association is

also the best among 1000 permutations and is genome-wide significant (Bonferroni $p = 2.7 \times 10^{-5}$). In total, 6 genetic markers showed genome-wide significant association (Bonferroni $p < 0.05$) but the association with rs6137010 in CHB + JPT was the strongest (Table 4.1). The P-P plots showed that the p-value of the association with the RS-score at rs6137010 is very different to other loci in the Asian populations (Figure 4.3). We found no evidence of population stratification in the GWAS tests of the Asian populations as their genomic inflation factors are less than 1.05 (Table S4 in Appendix A). However, unsurprisingly there was evidence of population stratification in three combined populations: CEU+GIH+MEX, YRI+LWK+MKK and ALL (Table S4 and Figure S4 in Appendix A).

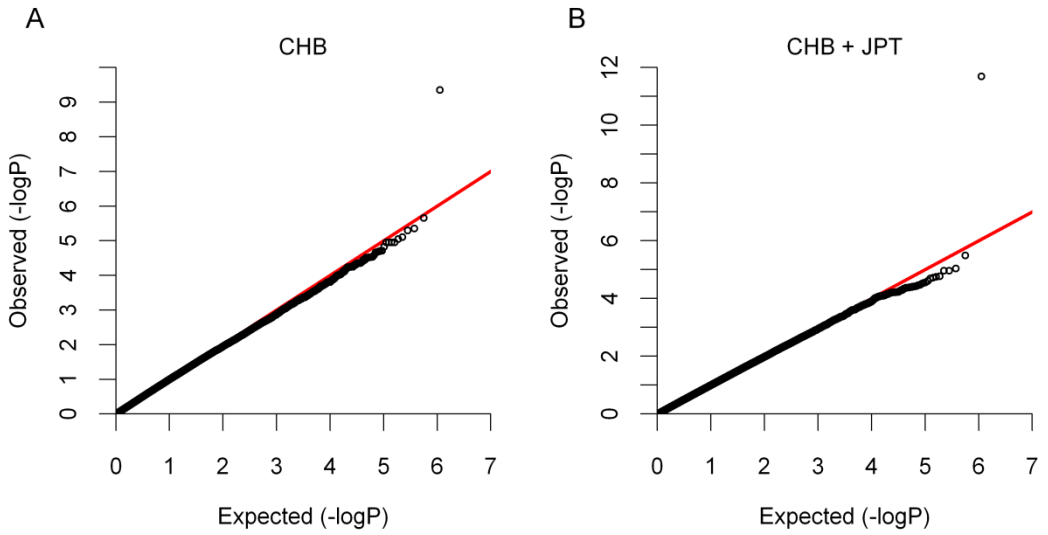


Figure 4.3: P-P plots of the association with RS-score in A) CHB and B) CHB + JPT. This figure compares the observed distribution of the $-\log_{10}$ P-values to the expected distribution, given that the P-values come from a uniform distribution in the interval zero to one (as expected under the null hypothesis). The Y-axis shows quantiles of the observed distribution and the X-axis shows the corresponding quantiles under the uniform distribution. The red line is used to compare the expected and observed values.

Table 4.1. Markers associated with the RS-score at Bonferroni $p < 0.05$

SNP	**Location**	**Function**	**Associated gene**	**Population**	**P-value**	**Bonferroni**
rs6137010	20:2038118	*cis*-eQTL	SNRPB	CHB+JPT	2.0×10^{-12}	2.7×10^{-5}
rs6137010	20:2038118	Intron	STK35	CHB+JPT	2.0×10^{-12}	2.7×10^{-5}
rs6137010	20:2038118	*cis*-eQTL	SIRPA	ALL	5.4×10^{-11}	9.0×10^{-4}
rs6137010	20:2038118	intron	STK35	ALL	5.4×10^{-11}	9.0×10^{-4}
rs11136253	8:145179783	*cis*-eQTL	ZNF707	ALL	1.5×10^{-10}	2.5×10^{-3}
rs11136253	8:145179783	coding-synon	OPLAH	ALL	1.5×10^{-10}	2.5×10^{-3}
rs6137010	20:2038118	*cis*-eQTL	SNRPB	CHB	4.4×10^{-10}	5.9×10^{-3}
rs6137010	20:2038118	Intron	STK35	CHB	4.4×10^{-10}	5.9×10^{-3}
rs4466324	7:85113458	unknown	None	ALL	6.0×10^{-10}	1.1×10^{-2}
rs17127419	11:122878168	unknown	HSPA8	ALL	9.8×10^{-10}	1.6×10^{-2}
rs12034707	1:178400832	*cis*-eQTL	TOR1AIP1	ALL	1.8×10^{-9}	3.0×10^{-2}
rs12034707	1:178400832	intron	QSOX1	ALL	1.8×10^{-9}	3.0×10^{-2}
rs10997765	10:69066422	intron	CTNNA3	ALL	1.9×10^{-9}	3.3×10^{-2}

To check the effect of the choice of half-life data on this result, we compared RS-scores calculated using half-life data from HeLa cells and RS-scores calculated using B-cell half-life data [252] and found that they were highly correlated in all populations (Spearman *rho*=0.73 ± 0.15). It has previously been reported that RNAs involved in housekeeping functions tend to have long half-life [248,253]. As an alternative to using half-life data, which has the caveat that it may be cell type dependent, we calculated the RS-score by grouping genes based on whether they are housekeeping or not, using data from Chang *et al.* [267]. We found that the RS-score calculated by grouping the genes in this way was highly correlated with the RS-score based on the half-life in HeLa cells in all populations (Spearman *rho*=0.70 ± 0.09). Moreover, the RS-score (based on the housekeeping data) was significantly associated with rs6137010 in the combined CHB+JPT population ($p = 7.1 \times 10^{-13}$; Bonferroni $p = 9.5 \times 10^{-6}$). We also calculated an equivalent score by considering protein-coding versus non-coding genes. Non-coding genes have been found to have shorter half-life than protein-coding genes [247,253]. This score was also highly correlated with the RS-score calculated from the half-life data and,

again, significantly associated with rs6137010 in CHB + JPT (p = 6.2 x 10^{-10}; Bonferroni p = 8.3 x 10^{-3}). These two results are of interest, beyond providing an alternative way to group genes that is not dependent on RNA half-life data that may differ between cell types. They suggest that the proportion of the RNA pool corresponding to non-coding and tissue-specific genes is associated with rs6137010 in Asian populations.

4.3.3 Searching for causal SNPs and causal genes

To search for causal SNPs that may explain the GWAS results we mapped each SNP that shows genome-wide significant association with the RS-score to a gene if the SNP is either within the gene or is a *cis*-eQTL (*cis*-expression Quantitative Trait Locus) of the gene using *cis*-eQTL data from Stranger *et al.* [263] (Table 4.1). We found that rs6137010, the SNP with the strongest GWAS signal, mapped close to the *SNRPB* gene, which is involved in RNA processing. *SNRPB* encodes part of the core small nuclear ribonucleoprotein particles (snRNPs) that are major components of the spliceosome complex. Although it is 352 kb downstream, rs6137010 is significantly associated with the expression level of *SNRPB* in both CHB (*rho* = 0.50; p = 2.3 x 10^{-6}) and JPT (*rho* = 0.32; p = 3.7 x 10^{-3}), but not significantly associated with *SNRPB* expression in any of the other populations studied. The association between rs6137010 and *SNRPB* is strongest among all genes within 1 Mb-window centered on the SNP. Furthermore, the SNP is within an enhancer region as evidenced from whole-genome chromatin state segmentation data [123] available through the UCSC genome browser [182]. These results show that rs6137010 is a *cis*-eQTL of *SNRPB* in Asian populations. Changes in the expression level of *SNRNPB* have been reported to affect alternative splicing and abundance of a large number of RNA processing factors [264]. rs6137010 has two alleles, T and C, with C the minor allele in Asian populations but the major allele in the other HapMap3 populations. Asian individuals carrying the C allele at this SNP had higher expression levels of *SNRPB* (Figure 4.4A) and higher RS-scores (Figure 4.4B). This suggests that the association between rs6137010 and inter-individual variation in RNA stability could be mediated by changes in *SNRPB* expression levels.

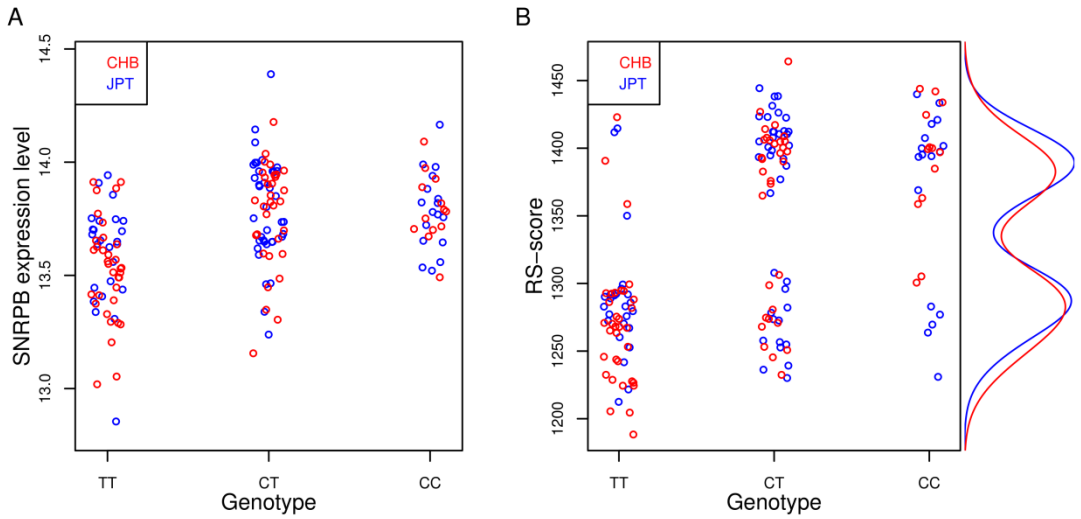


Figure 4.4: Stripcharts of *SNRPB* expression levels and the RS-score against the genotype of rs6137010 in CHB and JPT. A) *SNRPB* expression levels are significantly different among the three genotypes TT, CT and CC (p = 1.2 x 10^{-5} in CHB and p =1.9 x 10^{-3} in JPT from one-way ANOVA). B) RS-scores are significantly different among the three genotypes (p = 4.3 x 10^{-10} in CHB and p =7.0 x 10^{-5} in JPT from one-way ANOVA). The bimodal distributions of the RS-score in CHB and JPT are displayed in red and blue lines, respectively.

To identify genes across the human genome whose expression levels are significantly associated with rs6137010, we carried out *trans*-eQTL mapping for this SNP by fitting Spearman rank correlation models and considering only associations with FDR < 0.1. FDRs were calculated using the Benjamini and Hochberg procedure [268] as implemented in R [269]. We found 6,396 and 2,585 genes associated with rs6137010 in CHB and JPT, respectively. Among these, 3,194 (in CHB) and 429 (in JPT) genes were positively correlated with the minor allele count of rs6137010. Among the genes that were associated with rs6137010, 25.2% were putative targets for AU-rich element decay, compared to 17.6% of other genes (p = 0.01, Fisher exact test). We did not find any genes significantly associated with the SNP in other populations using the same FDR threshold. We carried out Gene Ontology (GO) analyses using DAVID [270] for the positively correlated genes and, interestingly, found that they were enriched for the GO term ribonucleoprotein complex in both CHB (p = 1.9 x 10^{-25}; Table S5 in Appendix A) and JPT (p = 3.7 x 10^{-5}). The ribonucleoprotein complex is known to be involved in

many steps of RNA processing such as pre-mRNA splicing and RNA transportation and stabilization. Both *HNRNPA2B1* and *SNRPB* mentioned above belong to the ribonucleoprotein complex. These results indicate that rs6137010 is a *trans*-eQTL cluster that is disproportionately associated with the expression levels of ribonucleoprotein complex genes.

We next turned to investigating further the possible role of *SNRPB* in mediating the association of rs6137010 with the RS-score. We obtained gene expression microarray data generated from HeLa cells in which *SNRPB* was knocked down and compared to controls [264]. Using the HeLa half-life data [253] we calculated and compared RS-scores between the two conditions and found a significant reduction of the RS-score in *SNRPB* knockdown ($p = 1.2 \times 10^{-6}$ from a two-tailed *t* test; Figure 4.5). This is consistent with expectations because depletion of *SNRPB* reduces the levels of many RNA processing genes [264], potentially affecting the stability of RNA across the transcriptome. Furthermore, the genes that were differentially expressed upon *SNRPB* knockdown were enriched for genes that showed the strongest association (FDR < 0.01) with rs6137010 in CHB ($p = 0.002$ from two-tailed Fisher's exact test). These results suggest that rs6137010, by modulating the expression of *SNRPB*, may be directly responsible for inter-individual variation in the RS-score in CHB. Interestingly, the distribution of the RS-score was bi-modal in both CHB and JPT (Figure 4.4B), consistent with the existence of an associated locus with a large effect size. It is tempting to speculate that an ungenotyped causal SNP in strong linkage disequilibrium with rs6137010 may stratify the samples between the two modes of the distribution. Higher resolution genotype data will be necessary to test this hypothesis.

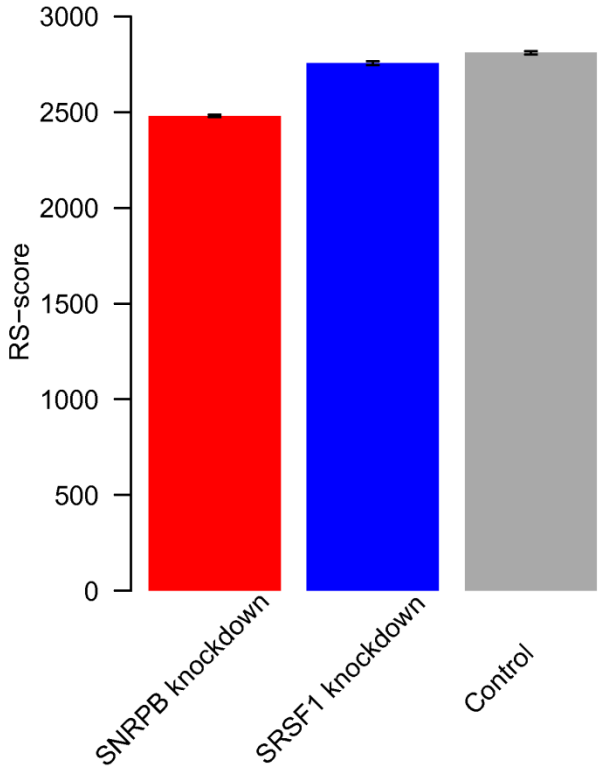


Figure 4.5: RS-scores calculated from three samples - *SNRPB* knockdown, *SRSF1* knockdown and control. The control corresponds to the sample transfected with nontargeting siRNA. Error bars represent two standard errors.

The RS-score of the knockdown of another splicing factor, *SRSF1*, is also significantly lower than of the control (p = 3.4 x10^{-4} from a two-tailed t test), but significantly higher than of the *SNRPB* knockdown (p = 1.1 x10^{-9} from a two-tailed t test) (Figure 4.5). This indicates that knocking down *SNRPB* has stronger effect on the RS-score than knocking down *SRSF1*. This is not surprising because *SNRPB* has been found to have a stronger impact than *SRSF1* on the inclusion levels of alternative exons that are enriched for genes encoding RNA processing [264]. *SNRPB*, which plays a central role in modulating expression levels of many RNA processing factors [264], might therefore have the strongest influence in the RS-score among RNA processing factors. Previous studies discovered the involvement of several splicing factors in RNA stability [271,272]. Thus, the core splicing factor *SNRPB* may have an important role in RNA stability as well.

4.4 Conclusions

Genetic variants that affect RNA stability in *cis* have been shown to contribute to inter-individual variation in gene expression [262]. Here we demonstrate that the effects of knocking down the expression of *HNRNPA2B1* that stabilizes a large number of RNAs can be detected from gene expression data. In particular, the expression of genes expressing transcripts with a long half-life is reduced relative to genes with short half-life transcripts. We defined the RS-score to summarize the relative expression of long-lived compared to short-lived transcripts. Treating the RS-score as a quantitative trait, we performed genome-wide association and identified a locus on chromosome 20p13 that is strongly associated with the RS-score in two Asian populations. This locus is a *cis*-eQTL for *SNRPB*, a core component of the spliceosome that has previously been shown to affect the expression of many RNA processing factors [264]. We propose that the *cis*-eQTL of *SNRPB* may be directly responsible for the association of the RS-score with this locus. Consistent with this model, knockdown of *SNRPB* results in a significant reduction in the RS-score.

4.5 Methods

4.5.1 Data

Processed gene expression data generated using the Illumina whole genome expression array from 726 lymphoblastoid cell lines (LCLs) in eight HapMap3 populations (CEU, CHB, GIH, JPT, LWK, MEX, MKK, and YRI) by [263] were downloaded from ArrayExpress [273]. Single nucleotide polymorphisms (SNPs) for the same 726 individuals were obtained from HapMap3 (release 2) [274]. SNPs with minor allele frequency (MAF) $\leq$ 1% in a population were excluded. This resulted in between 1.1 million and 1.3 million SNPs per population. Half-life data for 11,052 mRNAs and 1,418 ncRNAs in HeLa cells, and for 8,344 genes in B-cells were obtained from Tani *et al.* [253] and Friedel *et al.* [252], respectively.

4.5.2 RNA stability score

We defined the RNA stability score (RS-score), as a measure of the relative expression levels of long-lived and short-lived transcripts in a sample. We first

classified all genes as either expressing long or short lived RNAs, by setting a threshold on an available RNA half-life or decay rate data set. Specifically, for the HeLa half-life data [253], we chose the same threshold used by the authors to determine whether a gene expresses long-lived (half-life $\geq$ 4 hours) or short-lived (half-life < 4 hours) RNA. For the RNA decay rate data [251], a gene was considered as expressing long-lived RNA if its decay rate was greater than the average across genes (corresponding to a relative decay rate greater than 0) and as short-lived if its decay rate was less than average (corresponding to values less than 0). We then ranked all genes in the sample by their expression levels (a higher expression level corresponds to higher rank value). Finally, the RS-score is defined as the difference in the mean rank of genes expressing long-lived RNAs and genes expressing short-lived RNAs. Therefore, higher RS-scores correspond to higher relative expression of genes with longer half-life, consistent with more efficient stabilization of RNA.

4.5.3 Genome-wide association test

Assuming an additive mode of inheritance, we performed linear regression analysis to assess association of RS-score with SNP genotypes, using PLINK v1.07 [275]. We included gender as a covariate in the linear model to correct for any sex bias. To combine samples from different populations, we carried out a principal component analysis (PCA) as implemented in the Eigensoft 4.2 [266,276]. To correct for population stratification in genome-wide association tests, we included the first five principal components in addition to gender as covariates in the linear models. To check whether population stratification exists, we calculated the genomic inflation factor λ (λ_{GC}) as the median χ^2 association statistic across SNPs divided by 0.456, the predicted median χ^2 when there is no stratification [277]. Values of $\lambda_{GC} \leq 1.05$ generally indicate no population stratification [278].

4.5.4 Permutation testing

Applying a permutation testing procedure by Hirschhorn and Daly [279], in each GWAS test, we carried out 1000 permutations. In each permutation, we randomly shuffled the phenotype values, re-ran the GWAS and recorded the best (lowest) p-value from each run. Finally, we counted how many of these 1000 lowest

p-values are less than or equal to the original p-value being evaluated. The permutation *p* is defined as this number divided by 1000 (i.e. the proportion of the 1000 lowest p-values that are less than or equal to the original p-value).

4.5.5 Analysis of RNA-seq data from *SNRPB* knockdown samples

We downloaded RNA-seq data in HeLa cells generated by Saltzman *et al.* [264] from samples in which *SNRPB* or *SRSF1* was knocked down, and the control samples that were transfected with nontargeting siRNA. The data consisted of three samples for each knock down and three control samples. We mapped the RNA-seq reads to the human genome, build hg19, using Tophat 1.4.1 (with default parameters) [157] and estimated expression levels of RefSeq genes using Cufflinks 1.3.0 (with default parameters) [229]. Using the HeLa half-life data [253], we calculated the RS-score for each of the three samples.

Chapter 5: Identification of Human Genetic Variants Affecting mRNA Translation Rate

The contents of this chapter have been prepared for publication as:

Nguyen TT, Seoighe C: Identification of human genetic variants affecting mRNA translation rate.

5.1 Abstract

The mRNA translation rate has a major impact on steady-state protein levels and mutations that affect the rate of translation can cause genetic diseases. Some examples of *cis*-acting variants that contribute to inter-individual variation in protein levels have recently been reported. These variants are referred to as protein quantitative trait loci (*cis*-pQTLs). The existence of a *cis*-pQTL can be inferred from differences in the abundance of protein translated from alternative alleles in heterozygous individuals. This is referred to as allele-specific translation (AST).

Here we developed a computational pipeline for studying AST. We first demonstrated that our pipeline can call the genotypes accurately for a specific human cell type using existing high-throughput sequencing data. Applying this method, we called genotypes for HeLa cells at the common single nucleotide polymorphism (SNPs) from the HapMap3 project. We next phase the resulting genotypes to obtain a haplotype-resolved genome of HeLa cells. Mapping a Ribo-seq data set from HeLa cells to the haplotype-resolved genome, we indentified 171 genes with evidence of AST. Examination of the heterozygous SNPs from 5'UTRs of these AST genes revealed two interesting cases. First, the SNP rs9960, located within the translation initiation sites of the *ATP5H* gene, has two alleles A and G where the A allele is associated with much higher translation rate than the G allele. Thus the mutation A→G at this locus likely suppresses the translation of *ATP5H*. Second, the SNP rs6122080 is located at the conserved binding sites of a translation initiation factor. The mutation G→A at this SNP dramatically changes the

secondary structure of the 5'UTR of the *SLCO4A1* gene and, consequently, silences the translation of *SLCO4A1*.

This study presents an analysis pipeline to identify AST genes and reports novel *cis*-pQTLs in humans.

5.2 Introduction

Gene expression is regulated through the processes of transcription, mRNA decay, mRNA translation and protein degradation [280]. Genetic variants responsible for inter-individual variation in mRNA expression level have been studied in detail [263,281,282,283]. These variants are referred to as expression quantitative trait loci (eQTLs). In addition, a moderate number of genetic variants were found to associate with inter-individual variation in both RNA expression level and RNA decay, for which variation in RNA decay can explain the association with RNA expression level [262]. Recently, a study of Wu *et al.* [284] investigated the genetic control of gene expression at the translation level and identified some *cis*-acting variants that contribute to inter-individual variation in protein abundance (*cis*-pQTLs). Interestingly, many *cis*-pQTLs do not correspond to *cis*-eQTLs [284]. This indicates that variation in protein levels can be caused by distinct genetic mechanisms that are not responsible for variation in RNA levels. Two cases of *cis*-pQTLs have been shown to be linked human diseases [284].

At the transcription level, previous studies demonstrated that *cis*-eQTLs generally act by a mechanism involving allele-specific expression (ASE) [282,283,285]. Also, mapping of ASE is considered as a powerful approach to directly show that a variant acts in *cis* [285,286]. Similarly, at the translation level, *cis*-pQTLs can act by a mechanism involving AST. Mapping of AST helps discover the causal variants that influence the rate of mRNA translation.

Whole-proteome mass spectrometry has been widely used to assess directly changes in protein abundance [248,284,287]. However this method can only detect for a subset of protein-coding genes [288]. Recently, the ribosome profiling method (Ribo-seq), based on deep sequencing of ribosome-protected mRNA fragments, has been developed to monitor the protein synthesis for a much larger number of genes [289]. The abundance of protein estimated by both methods are well correlated

[289], suggesting that the Ribo-seq can be used to capture the rate of protein synthesis. More importantly, within a single Ribo-seq experiment, both mRNA levels and protein levels can be detected simultaneously by sequencing all the transcribed mRNA fragments (also referred to as RNA-seq) and by sequencing only those fragments protected by the ribosome (Ribo-seq), respectively [289,290]. This advantage allows monitoring accurately the rate at which an mRNA molecule is translated to protein and opens the door for studying of global gene expression control at unprecedented resolution.

A previous study of Guo *et al.* [290] applied the Ribo-seq approach to simultaneously measure the mRNA expression and protein abundance in the absence and in the presence of microRNAs and thereby discriminated the regulatory effect of microRNAs between the mRNA levels and the translation efficiency. Similarly, the Ribo-seq method can enable identifying the relative contributions of genetic variants to the rate at which the gene is copied into mRNA and the rate at which the mRNA is translated into protein.

The RNA-seq data have provided powerful joint-analysis of variation in mRNA levels and ASE [282,283]. In the same manner, the Ribo-seq data makes possible for the joint-analysis of variation in protein levels and AST. Therefore, among the variants that are associated with changes in protein abundance, we can pinpoint the causal ones by using the Ribo-seq data. However, this task remains difficult when using the mass spectrometry data. Furthermore, generating mass spectrometry data for the large-scale study is known to be very costly [287].

Here, we developed a novel pipeline for studying AST in humans by integratively analyzing the RNA-seq and Ribo-seq data. We aimed to find the potential disease consequences of novel *cis*-pQTLs. The analysis of allele-specific events using short read sequences is severely influenced by the biased mapping of reads to the reference allele [291]. Therefore the pipeline first constructs a haplotype-resolved genome for the cell-type from which the RNA-seq and Ribo-seq data were obtained. This step takes advantage of large volumes of high-throughput sequencing data that are publicly available for the cell-type. Then the RNA-seq and Ribo-seq data are mapped to the haplotype-resolved genome to measure the mRNA

levels and protein levels of protein-coding genes. The pipeline next applies statistical tests to compare the rate of translation from the two haplotypes and, accordingly, identify genes that are associated with AST. Applying this pipeline for the RNA-seq and Ribo-seq data of Guo *et al.* [290] from HeLa cells, we found 171 genes showing evidence of AST. Interestingly, these genes are significantly enriched among the genes with *cis*-pQTLs identified by Wu *et al.* [284]. Examining the heterozygous SNPs within the 5'UTR of these AST genes, we discovered two mutations that appear to suppress the translation initiation.

5.3 Results and Discussion

5.3.1 Overview of the pipeline

To study AST, we developed a pipeline with two starting points (Figure 5.1). The pipeline requires the haplotype information. Thus in case this information is not available, the pipeline can construct a haplotype-resolved genome for a specific cell type by using available high-throughput sequencing data for this cell type. This step was described in detail in the Methods section. Briefly, we first applied a strict filtering process to trim low quality bases and discard low quality reads from the short read sequences. The resulting reads were aligned to a custom build of the human reference genome HG19. This custom genome incorporates alternate alleles at each known SNP in order to reduce the mapping bias in favour of the reference allele. Then, the genotypes are called with SAMtools and BCFtools [237,292]. We only focused on common SNPs of the HapMap3 project [274]. The genotypes are next phased with SHAPEIT [293] by incorporating genotypes of individuals from HapMap3 [274].

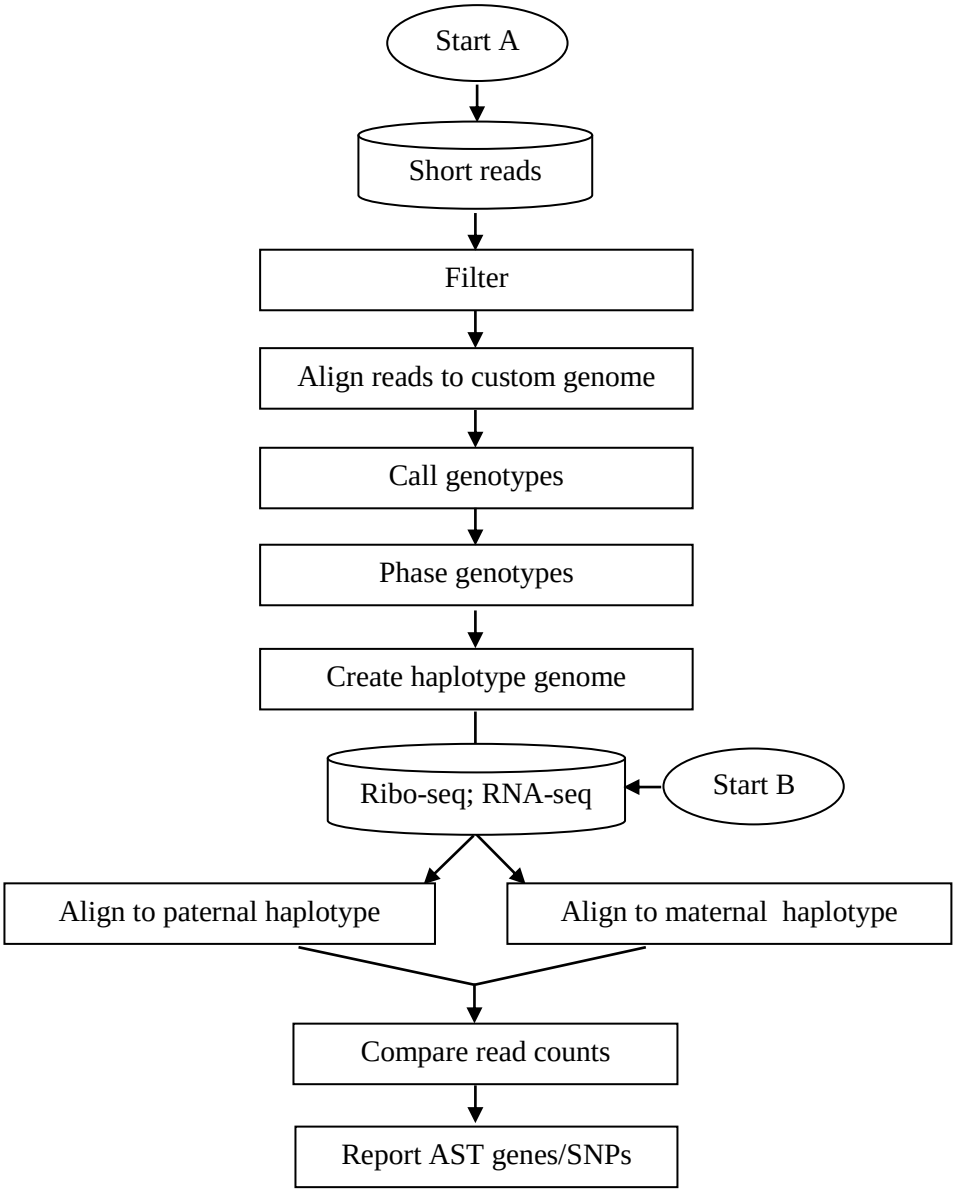


Figure 5.1: Flowchart of the pipeline for studying AST. The pipeline requires the haplotype information. In case this information is not available, a haplotype-resolved genome is constructed for a specific cell type by using the available high-throughput sequencing data for this cell type (Start A). In case the haplotype genome is available, the pipeline can start finding the AST genes immediately (Start B). RNA-seq and Ribo-seq data are mapped to the parental haplotypes. The Ribo-seq read counts are compared between the two haplotypes to identify the AST genes. The RNA-seq read counts are used to discard those AST genes that are likely to be a consequence of ASE.

The RNA-seq reads and Ribo-seq reads are mapped to each parental haplotype. The Ribo-seq read counts from the two haplotypes of each protein-coding gene are compared to check whether this gene is associated with AST. The RNA-seq read counts are used to exclude the possibility that the allelic differences in Ribo-seq counts are caused by allelic imbalance in mRNA abundance. Inspection of heterozygous SNPs within the AST genes is then carried out to identify those SNPs that cause AST. Finally, these results will be cross-referenced with databases of genes and genetic variants that have been associated with human diseases and phenotypes to identify candidate disease-causing variants.

5.3.2 Building the haplotype genome for HeLa

We first attempted to construct the haplotype-resolved genome of HeLa cells from which we obtained the RNA-seq and Ribo-seq data [290]. This cell line has been using in many genomic studies. Thus a large amount of high-through sequencing data from HeLa can be downloaded from the public databases, e.g. the Sequence Read Archive (SRA) [294]. Previous studies have successfully identified and genotyped SNPs from high throughput sequencing data [295]. The coverage and quality of sequencing have a great impact on the accuracy of genotype calling [295]. To start the construction of the HeLa haplotype-resolved genome, we downloaded high-through sequencing data for HeLa cells in 430 runs from the SRA [294]. In total, we obtained more than 11 billion reads (read length <= 60 nt) (Table 5.1). We noted that using the short-read data from factors that are differentially enriched between the two parental haplotypes can lead to inaccurate genotype calling. Thus we mainly focused on using the Input DNA/Control data.

To yield the most accurate genotype calling, we applied stringent filtering criteria to trim low quality bases and remove low quality reads that are potentially cause spurious mapping and inaccurate genotype calling (see the Methods). After filtering, more than 7 billion reads with length from 25 to 60 bp were retained, yielding coverage of ~ 100X. Because the reference genome contains only one allele (referred to as the reference allele) at any given SNP, reads carrying non-reference alleles are less likely to be mapped compared with reads carrying reference alleles. This bias can result in inaccurate genotype calling. We therefore

applied several strategies to reduce the effect of the read mapping bias. We built a custom human genome that contains two parts. The first part, called the primary part, is a modified version of the latest human reference genome (HG19). For each known SNP across the HG19, we replaced the nucleotide at this locus with the minor allele of the SNP. Here we made use of 1,458,412 SNPs avaliable from HapMap3 [274]. The second part, called the enhanced part, includes sequence tags containing the major alleles at these SNPs. Each tag extends 50 bases on either side of each of the SNPs.

We carried out a two-step mapping strategy to align the the HeLa short read sequences. The reads were first mapped to the primary part without allowing any mismatches. The unmapped reads were subsequently mapped to the enhanced part without allowing any mismatches as well. By default, the majority of short-read aligners allow up to two mismatches, which are normally due to SNPs [296]. In our stragtegy, reads that failed to map during the first step due to the mismatches at SNPs can be 'rescued' in the second step. We obtained high a proportion of reads (67%) that mapped perpectly and uniquely to the custom genome (Table 5.1).

Table 5.1: Summary of the short reads from HeLa and GM12878 cells.

	# of runs	# of raw reads	# of reads after fitering	# of mapped reads
HeLa	430	11,112,077,494	7,496,002,218	5,026,504,141 (67%)
GM12878	156	3,609,979,722	2,759,025,888	2,095,130,742 (76%)

The value in brackets within the 3rd column is the percentage of reads, following filtering, that mapped perfectly and uniquely to the custom human genome.

We used SAMtools and BCFtools [237,292] to call the HeLa genotypes at 1,450,883 SNP loci of HapMap3. Remarkably, we were able to identify the HeLa genotypes at 97.8% of the loci. Among these, 22.4% are heterozygous – that is slightly smaller than the average percentage of heterozygous SNPs of the HapMap3 individuals (27.3%). HeLa cells were taken from an African American (Henrietta Lacks) who is expected to share common ancestors with other individuals in the

African ancestry in Southwest USA (ASW) population. To check this we applied a principal components analysis (PCA) approach [266] (see the Methods) to model ancestry differences among HeLa and all individuals from 12 different populations of HapMap3. As we expected, the scatter plot of the first and second principal components showed that HeLa is clustered with the ASW population and is close to other African populations (Figure 5.2).

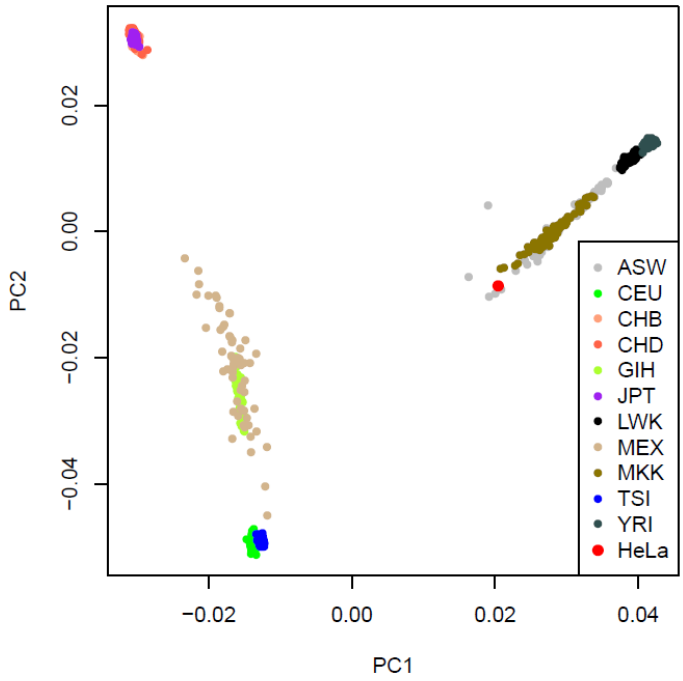


Figure 5.2. PCA of the genotypes from HeLa and HapMap3 samples.

To assess the accuracy of our genotype calling pipeline, we applied the pipeline to call genotypes for a lymphoblastoid cell line, GM12878. This is a primary cell line that has been investigated in detail by the ENCODE Consortium [297] and therefore large volumes of high-throughput sequencing data are publicly available for this cell line. We downloaded the short read sequences from 156 runs for GM12878 from the SRA (Table 5.1). Applying the genotype calling pipeline, we were able to call genotypes for GM12878 at 1,406,880 out of the 1,450,883 HapMap3 SNP loci (97%). The resulting genotype calls are compared with the known genotypes of a HapMap3 individual, NA12878 from whom the GM12878 cell line was taken. Most of the resulting genotype calls for GM12878 (98%) are concordant with the known genotypes of NA12878. As HapMap3 genotypes at many loci can be called inconsistently among different genotyping platforms [274],

98% of the resulting concordance level indicates that our pipeline can provide accurate genotype calling.

We next phased the resulting HeLa genotypes by incorporating the existing genotype data of individuals from the ASW population. Similar to the HapMap3 phasing pipeline, we used the ASW trios samples as reference panel to improve phasing accuracy. These samples usually have high-quality genotypes. The remaining ASW samples were incorporated to phase the genotypes for HeLa (see the Methods).

5.3.3 Determining AST

We developed a method to identify protein-coding genes that show evidence of AST (see the Methods). To investigate the evidence of AST, we mapped the Ribo-seq data from Guo *et al.* [290] to each parental haplotype independently. Similar to previous studies [289,290], we estimated the absolute protein abundance of a gene by counting the number of Ribo-seq reads mapped to this gene. We noted that if a gene is associated with ASE, this gene is likely to be associated with AST as well because the mRNA levels and protein levels are correlated [248]. We therefore excluded the possibility that the allelic differences in Ribo-seq counts are caused by allelic imbalance in mRNA abundance. To detect the evidence of ASE, we used the RNA-seq data obtained from the same study [290] and carried out the same analysis as we did for the Ribo-seq data.

We next applied several statistical tests to identify genes associated with AST by using the RNA-seq read counts and Ribo-seq read counts. We focused on 716 protein-coding genes that have sufficient read counts from both the RNA-seq and Ribo-seq data. For each gene we constructed a 2 x 2 contingency table containing these read counts from the two haplotypes – an example is shown in Table 5.2 for the *TRNT1* gene. For each gene, we first checked whether the proportion of reads mapping to haplotype A (or haplotype B) differs significantly between the Ribo-seq data and the RNA-seq data by applying a Fisher's exact test (see the Methods). As a result, we obtained 285 genes (40%) at the false discovery rate (FDR) of 5%. Next, for each of these genes, we used two one-sided binomial tests to evaluate the complementary hypotheses that the Ribo-seq read count from the haplotype A was

greater or less than from the haplotype B. These same tests were also applied for the RNA-seq read counts to assess the ASE evidence. We also excluded the possibility that the allelic differences in Ribo-seq counts are caused by ASE. Consequently, we identified 171 genes (60%) that are associated with AST (p <= 0.001; binomial test).

Table 5.2. The RNA-seq read counts and Ribo-seq reads counts from two haplotypes of the *TRNT1* gene.

	Haplotype A	Haplotype B
Ribo-seq read counts	360	12
RNA-seq read counts	55	42

The *cis*-pQTLs can act by a mechanism involving AST. Thus we started by looking at the overlap between a set of the 171 AST genes found above and a set of 86 genes with *cis*-pQTLs discovered by Wu *et al.* [284], bearing in mind that these two sets of genes were identified from different cell types. Wu *et al.* [284] investigated the contribution of *cis*-acting variants to the inter-individual variation in the protein abundance in lymphoblastoid cell lines from HapMap individuals [274]. As we expected, the AST genes are significantly enriched among the *cis*-pQTL genes (Figure 5.3; p = 0.04 from two-tailed Fisher's exact test). Four genes including *TRNT1, NSUN4, ALDH16A1 and ERAP1* were found to be common between the two sets.

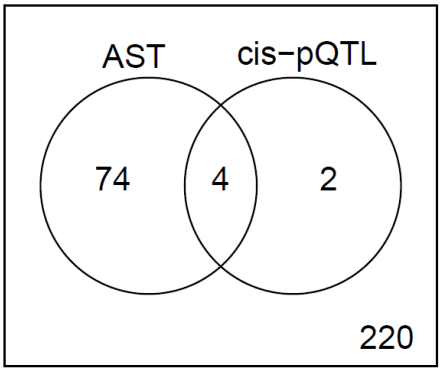


Figure 5.3. Venn diagram of the overlap between the AST genes (labeled as AST) and the *cis*-pQTLs genes (labeled as cis-pQTL).

We next turned to indentifying the candidate causal variants that influence the rate of mRNA translation. The translation is principally regulated at the initiation stage [44]. Therefore, genomic variants within the translation initiation sites (TIS) can influence the translation. In vertebrates, the TIS was first defined by Kozak as bases -6 to +4 around the start codon (where the A of the start codon AUG is designated +1, with positive and negative integers proceeding 3' and 5', respectively) [298]. Kozak showed that the consensus motif GCC(A/G)CCAUGG is the optimal context of translation initiation [299]. Searching for heterozygous SNPs within the TIS of AST genes, we found a SNP, rs9660, located 1 bp upstream of the start codon of the *ATP5H* gene (i.e. the -1 position of the Kozak motif). *ATP5H* encodes ATP synthase, H+ transporting, mitochondrial F0. The SNP, rs11870474, located downstream of this gene, was reported to be associated with Alzheimer's disease risk [300]. However, whether this association is mediated by *ATP5H* is not clear [300].

The presence of a C nucleotide at the -1 position of the Kozak motif has the strongest stimulatory effect on translation [299]. The SNP rs9660 has two alleles, A and G. The presence of A at the -1 position has a stronger stimulatory effect than that of G [299]. Consistent with this, we found that the A allele is associated with higher translation rate than the G allele. Specifically, the number of Ribo-seq reads mapping to the A allele (n = 144) is 11 times higher than for the G allele (n = 13) ($p = 2.04 \times 10^{-29}$; binomial test). While the numbers of RNA-seq reads mapping to the A allele (n = 105) and to the G allele (n = 101) are similar ($p = 0.42$; binomial test). To infer the average rate at which an mRNA molecule of the *ATP5H* gene is translated to protein, we divided the Ribo-seq read count by the RNA-seq read count. Accordingly, we observed that the average translation rate of the mRNA carrying the A allele is more than 8 times higher than that of the mRNA carrying the G allele. In HeLa genome, *ATP5H* has two copies from one haplotype and one copy from the other [301]. Accounting for this imbalance, the number of Ribo-seq reads mapping to the A allele is still much higher than for the G allele (fold-change > 5.5).

The fidelity and efficiency of translation initiation is encoded in the 5' untranslated region (5'UTR) [44,302]. The canonical pathway of eukaryotic

translation initiation requires many factors [44]. One of these factors, *eIF4B*, has been reported to be able to recognize an RNA motif within the 5'UTR of mRNAs [303]. This factor contains two binding domains: one can target a subunit of the ribosome and the other can recognize the RNA motif [303]. During translation initiation, this structure promotes ribosome binding to the 5'UTR [303]. The pattern GGA within this RNA motif ($\frac{G}{C}U\frac{U}{C}GGA\frac{A}{C}$) is highly conserved and contributes to the binding of *eIF4B* to the 5'UTR [303].

Focusing on the 5'UTR of the AST genes, we found 16 heterozygous SNPs (in 16 distinct genes). We next checked whether any of the candidate SNPs are located within the binding sites of the core translation initiation factor *eIF4B*. Scanning the 5'UTR sequences of these 16 genes with the RNA motif of *eIF4B*, we found one potential occurrence that overlaps a SNP, rs6122080, located 80 bp upstream of the start codon of the *SLCO4A1* gene. *SLCO4A1* was shown to be involved in the inflammatory response and photoreceptor death [304]. The SNP rs6122080 has two alleles G and A. The mutation G→A at this SNP changed the conserved pattern GGA to AGA and, consequently, inhibited the translation of *SLCO4A1*. Specifically, 13 Ribo-seq reads contain the G allele but no Ribo-seq reads contain the A allele. While the RNA-seq read counts are similar between G and A, 11 and 14, respectively.

In humans, the RNA motif of *eIF4B* participates in a conserved mRNA stem-loop structure that facilitates the binding of *eIF4B* to the 5'UTR [303]. We examined whether the mutation at the SNP rs6122080 can change the mRNA secondary structure of the 5'UTR of *SLCO4A1*. Using the RNAfold program [305], we predicted the secondary structure for different versions of the 5'UTR of *SLCO4A1*. Interestingly, the 5'UTR containing the G allele has a strong and conserved secondary structure in which the RNA motif of *eIF4B* participates in a stem-loop structure (Figure 5.4). Whereas, the 5'UTR containing the A allele has a weak secondary structure that differs considerably from the structure of the 5'UTR containing the G allele (Figure 5.4). The sequence context around the GGA pattern at the 5'UTR of *SLCO4A1* is CUCGGAU that contains one mismatch compared with the consensus RNA sequence of *eIF4B*, CUCGGAA, in the last position

(underlined). However, changing U to A in this position does not affect the secondary structure of the 5'UTR (Figure 5.4). Together, we proposed that the mutation G→A at the SNP rs6122080 changes the secondary structure of the 5'UTR of *SLCO4A1* and, consequently, affects the initiation of translation of this gene.

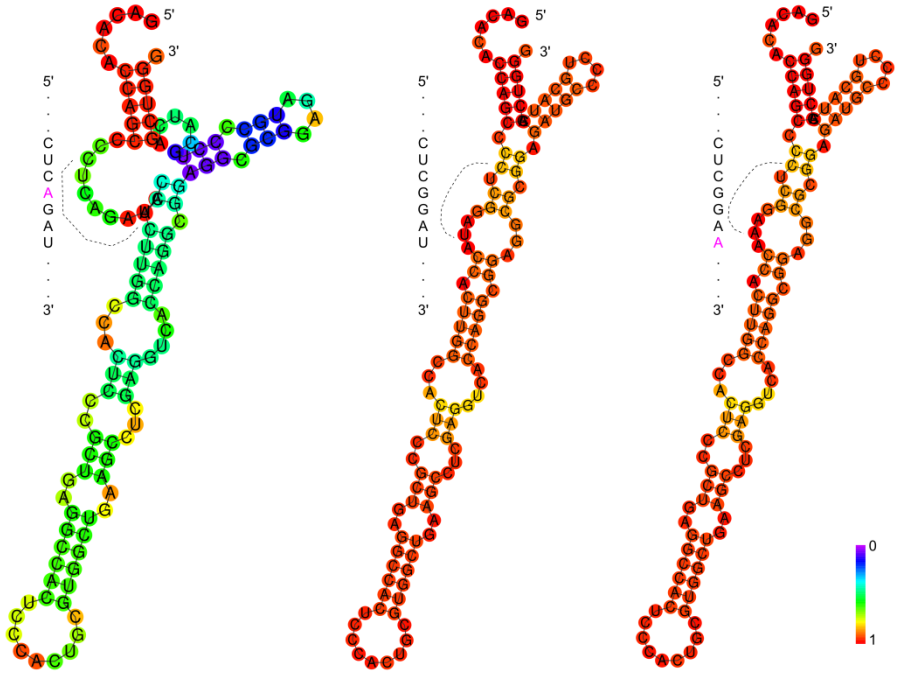


Figure 5.4. Predicted secondary structures of different versions of the 5'UTR of the *SLCO4A1* gene. Three plots show the secondary structure of three versions of the 5'UTR of *SLCO4A1*. In each plot, the occurrence of the *eIF4B*'s RNA motif is marked by the dashed curve together with the RNA sequence. The left and middle plots show the structure of the 5'UTRs that contain the A allele and the G allele at rs6122080, respectively. The right structure corresponds to the 5'UTR containing the G allele at rs6122080 but the last position of the occurrence of the *eIF4B*'s RNA motif was changed from U to A.

5.4 Conclusions and future works

The rate at which mRNA is translated to protein plays a key role in steady-state protein levels. Genetic variants that affect the rate of translation (pQTLs) can cause genetic diseases. The analysis of allele-specific mRNA translation (AST) can directly demonstrate that a pQTL acts in *cis*. This study presented a computational pipeline for identification of human genetic variants that are associated with AST in order to discover potential disease consequences of novel human pQTLs. The

pipeline has tackled two major tasks. First, it constructs a haplotype-resolved genome for a popular cell-type by making use of large volumes of high-throughput sequencing data that are publicly available for that cell-type. Second, the pipeline combines the RNA-seq and Ribo-seq data to identify genes that are associated with AST. Mapping the RNA-seq and Ribo-seq data, ideally from a single experiment, to the haplotype genome, the pipeline estimates simultaneously the mRNA levels and protein levels for protein-coding genes. It next applies several statistical methods to compare the rate of translation from the two haplotypes and, accordingly, identify genes that are associated with AST.

Applying this pipeline for the RNA-seq and Ribo-seq data of Guo *et al.* [290], we found a moderate number of AST genes that are enriched among the *cis*-pQTL genes identified by Wu *et al.* [284]. Inspection of heterozygous SNPs within the AST genes revealed interesting mutations in two cases. A mutation A→G occurred at the translation initiation sites of the *ATP5H* gene can suppress the mRNA translation of this gene. Another example, a mutation G→A occurred at the conserved binding sites of the translation initiation factor *eIF4B* within the 5'UTR of the *SLCO4A1* gene affected the initiation of translation of this gene.

This pipeline was designed to apply to any Ribo-seq dataset from popular cell-types. Recently, more Ribo-seq data have been generated from different human cell-types. For example, studies of Ingolia *et al.* [306], Lee *et al.* [307], Reid *et al.* [308] and Liu *et al.* [309] provided the Ribo-seq data from HEK293 cells; studies of Guo *et al.* [290], Liu *et al.* [309] and Stadler *et al.* [310] provided the Ribo-seq data from HeLa cells. Similar to the quick developments of the RNA-seq, large volumes of Ribo-seq data are expected to be publicly available shortly. We propose to apply our pipeline to more data sets to discover a more complete map of AST in humans.

Adey *et al.* [301] have recently discovered a comprehensive map of genomic variants and a high-quality haplotype-resolved genome from HeLa cells. We have applied for access to these data, which are subject to an agreement made between the NIH and the family of Henrietta Lacks [311,312]. We plan to validate the genotypes data and the haplotype genome of HeLa cells obtained from our pipeline (as described above) with the data of Adey *et al.* [301]. Also, as our pipeline can

start with a haplotype genome (Figure 5.1), we propose to map those RNA-seq and Ribo-seq data sets from HeLa cells [290,309,310] to the HeLa haplotype-resolved genome of Adey *et al.* [301]. This will, no doubt, yield more accurate mapping results.

Finally, we plan to carry out experimental validations for several interesting cases of AST. This is necessary because estimating the protein abundance from the Ribo-seq data is an indirect method and, consequently, may lead to spurious findings in the subsequent analyses (e.g. the AST analysis).

5.5 Methods

5.5.1 Data

The short read sequences corresponding to HeLa cells and GM12878 cells were downloaded from the SRA, with the majority Input DNA and Control data (e.g. IgG control). Both RNA-seq and Ribo-seq data in HeLa cells were obtained from the same study, Guo *et al.* [290]. Genotypes at 1,458,412 known SNP loci were downloaded from the HapMap3 project [274].

5.5.2 Haplotype phasing for HeLa

We developed a pipeline for calling and phasing genotypes in HeLa cells (Figure 5.1). Below, we described each step in details.

5.5.2.1 Filtering

To yield accurate genotype calling, we carried out a strict filtering process for the short read sequences that we downloaded from SRA for HeLa and GM12878 cells. We applied highly stringent criteria to trim low quality bases and discard low quality reads. Specifically, we first trim ambiguous nucleotides (Ns) and low quality bases from both 5' end and 3' end of reads. A base is considered to have low quality if the error rate of calling this base > 1% (i.e. the Phred score < 20). After trimming, any reads shorter than 25 nt are discarded. Reads having any low quality bases (Phred score < 13) or Ns are also discarded. This step is performed by custom python scripts.

5.5.2.2 Mapping

We focused on calling the genotypes for HeLa at the known SNPs from the HapMap3 project [274]. We first built a custom human genome by using the latest human reference genome (HG19) and the HapMap3 genotype data [274]. We obtained genotypes at 1,458,412 loci for all populations from HapMap3 [274]. We discarded those SNPs that are not bi-allelic or potentially mapped to the wrong strand by HapMap3. Following this, 1,450,883 SNPs remained, each associated with two alleles. We used this data to create the custom human genome. Because the short reads strongly favour mapping to the reference allele [291], we created two parts for the custom human genome. The first part, called the primary part, is a modified version of the latest human genome assembly (HG19). For each of the SNPs, we replaced the nucleotide at this locus with the minor allele of the SNP. The second part, called the enhanced part, was built following the method of Satya *et al.* [313]. Satya *et al.* [313] added sequence tags containing the alternative alleles to reduce the bias of reads mapping to the reference allele. Similarly, the enhanced part included sequence tags containing the major alleles at these SNPs. Each tag extends 50 bases on either side of each of the SNPs.

We carried out a two-step mapping of the short reads to the custom human genome. We first mapped the short reads to the primary part by using Bowtie 0.1.9 [137], allowing no mismatch. The un-mapped reads were mapped to the enhanced part, allowing no mismatch as well. We applied this command for both steps: "bowtie -p 10 -v 0 --sam -m 1 REF fastq_file", where -v 0 indicates allowing no mismatch and -m 1 indicates reporting only uniquely-mapped reads. Collecting the reads mapping uniquely and perfectly to the custom genome from both steps, we have final alignment data that can be used for the genotype calling.

5.5.2.3 Genotype calling

We carried out genotype calling at the 1,450,883 SNP loci mentioned above by using SAMtools and BCFtools [237,292] with almost default parameters. We set "-Q 15" to consider only bases having the Phred quality score of at least 15. High-quality genotypes were kept for the phasing step.

5.5.2.4 Haplotype Phasing

We used SHAPIT (version 2) [293,314] to phase the genotypes for HeLa. As HeLa cells are taken from Henrietta Lacks who is an African American, we used the genotype data from a corresponding HapMap3 population, African ancestry in Southwest USA (ASW). Following the HapMap3 phasing pipeline, we used the trios samples as reference panel. We incorporated genotypes of the remaining samples to phase the HeLa genotypes. We only chose SNPs that are present in both the reference panel and HeLa genotype data. The final dataset contains 1,168,786 SNPs across the human autosomes. We ran SHAPIT with default parameters.

5.5.3 Mapping Ribo-seq data

We constructed the haplotype genome for HeLa by using the genotype phasing results obtained in the previous step. We then used Tophat (version 1.4.1) [157] to map the RNA-seq and Ribo-seq data of Guo *et al.* [290] to the HeLa haplotype genome. We mapped these data to each haplotype version independently by using this command: "tophat -N 1 --segment-mismatches 1 --segment-length 60 -n 1 -G refSeq --no-novel-juncs -o outdir hap_genome". The refSeq protein-coding genes, a known gene model, downloaded from UCSC genome browser [235] on the 8th July 2013 was used to guide the mapping (-G refSeq). We only looked for reads across junctions indicated in the refSeq gene model (--no-novel-juncs). We only allowed 1 mismatch (-N 1 -n 1 --segment-mismatches 1) and reported only unique alignments (containing the tag NH:i:1). The results from different replicates were merged to produce the final alignment data that is provided for the next step.

5.5.4 Estimating allele-specific translation

For each gene containing at least one heterozygous SNP in HeLa, we counted the number of RNA-seq reads and the number of Ribo-seq reads mapping to each of the two haplotypes of that gene. We also accounted for the differences in total number of mapped reads between the RNA-seq and Ribo-seq data. At each gene, we have four count values: x_A, x_B corresponding to RNA-seq reads mapping to haplotype A and haplotype B respectively; y_A, y_B corresponding to Ribo-seq reads mapping to haplotype A and haplotype B respectively. Table 5.2 showed an example for the *TRNT1* gene. We only kept genes for which both the RNA-seq and Ribo-seq have at least 10 reads (from the two haplotypes).

We constructed a 2 x 2 contingency table for each gene from the four count values. Then a Fisher's exact test was applied to check whether the proportion of reads mapping to haplotype A (or haplotype B) is significantly different between the Ribo-seq data and the RNA-seq data. The Fisher's exact test was performed in R [269] and was adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR). Those genes that passed the test with FDR $\leq$ 5% and estimated odds ratio $\leq$ 2/3 or $\geq$ 3/2 are selected for further analysis. Similar to Degner *et al.* [291], we next applied two one-sided binomial tests to each gene to test the complementary alternative hypotheses that the read count from one haplotype is greater than or less than the read count from the other. This procedure is applied separately for the Ribo-seq to check for the evidence of allele-specific translation (AST) and for RNA-seq data to check for the evidence of allele-specific expression (ASE) of a gene. Finally, we excluded those AST genes that have the ASE in the same direction (i.e. x_A > x_B and y_A > y_B; or x_A < x_B and y_A y < y_B).

Chapter 6: Conclusions

Chromatin structure and gene regulation are implicated in nearly all aspects of human growth, development and diseases. Research in these areas has revealed mechanisms underlying how chromatin is formed and regulated, and how gene expression is controlled. This thesis participates in these exciting research topics and discovers novel insights into the DNA damage response, nucleolar formation and function, and genetic variation on mRNA stability and mRNA translation.

Chapter 2 characterized the global genomic distribution of the histone H2AX, a key factor of the double strand break (DSB) repair pathway. We generated H2AX and H2B ChIP-seq libraries in human U2OS cells and characterized genome-wide landscapes of these histones. We found multiple lines of evidence supporting that H2AX is more abundant in heterochromatin than in euchromatin. It has been reported that heterochromatic DSBs are repaired with slower kinetics and less efficiency than euchromatic DSBs [189,190]. Also, H2AX is not efficiently phosphorylated in response to heterochromatic DSBs [174,191]. Therefore, we propose that the greater enrichment of H2AX in heterochromatin can guarantee sufficient H2AX phosphorylation to signal heterochromatic DSBs. However this study contains several caveats. The H2AX and H2B ChIP-seq libraries have too high GC contents, high read duplicate levels, and insufficient sequencing coverage. Due to these concerns, we postponed submitting this study for a publication. A recent study by Seo *et al.* [192] reported that H2AX is concentrated on the transcription start site of actively transcribed genes. The data of Seo *et al.* [192] suggested that H2AX is preferentially enriched in euchromatin – that does not agree with our findings mentioned above. It also is unknown whether the distribution of H2AX is cell-type specific. Thus, more experimental efforts are necessary to resolve the discrepancy between our findings and the results of [192]. Also, it will be interesting to examine how dynamic the enrichment of H2AX is among different cell-types, particularly between embryonic stem cells and cancer cells.

A second study in the area of chromatin biology was presented in Chapter 3, which investigated a hidden part of the human genome – the nucleolar organizer

region (NOR). The NOR harbours an array of rDNA repeats that are transcribed to produce ribosomal RNAs. The sequence of each rDNA repeat is known but sequences anchoring both sides of the rDNA array are still missing from the current human genome assembly. Our collaborators established 380 kb of the sequence distal to the rDNA array (termed as distal junction – DJ) and 200 kb of the sequence proximal to the rDNA array (termed as proximal junction – PJ). The PJ structure resembles the sequences bordering centromeres that are very repetitive and contain large numbers of segmental duplicates. These results strongly suggest that the PJ is a region of active and frequent recombination. In contrast, the DJ region is replete with unique sequences. Although FISH experiments from previous studies showed that the DJ is embedded within perinucleolar heterochromatin, we found that the DJ likely has an open chromatin structure. An integrative analysis of FAIRE-seq, DNase-seq and ChIP-seq (of nine chromatin marks) data from different cell types revealed a conserved euchromatic landscape of the DJ. Importantly, the DJ chromatin profile indicates transcriptional activities and therefore we set out to characterize the DJ transcriptome. We discovered two RNA Pol II-transcribed transcripts that are supported by multiple lines of evidence and were confirmed by RT-PCR experiments. These surprising results led us to further functional experiments. The DJ was found to be localized to the periphery of the nucleolus, where it anchors the rDNA array. Thus, our results yielded insights into the role of NORs in nucleolar formation and function, and open the door for investigating of the link between nucleoli and human pathologies. However, further analysis (e.g. by using Hi-C data) is necessary to reveal the mechanisms underlying the interaction between the rDNA array and its flanking elements.

Chromatin structure is a major determinant of gene expression variation. Chromatin accessibility at the promoter of a given gene can control the rate at which the gene is transcribed [315]. Notably, genetic variants within the regulatory regions (e.g. promoters and enhancers) can change transcription factor binding affinity and chromatin accessibility, thereby leading to changes in gene expression [316,317]. We were therefore motivated to study genetic variation in gene expression. The next two chapters particularly focused on the contribution of

genetic variants to mRNA stability and mRNA translation, which are currently poorly understood.

In Chapter 4, we searched for *trans*-acting variants that affect transcriptome-wide RNA stability. Previous studies have measured RNA stability at the single gene level [248,253] and identified variants that influence the stability of nearby genes [262]. Our study aimed to identify genetic variants that change the activity of *trans*-acting RNA-stabilizing factors, thereby affecting stability of many transcripts, particularly the long-lived ones. We first demonstrated that differences in the activity of RNA-stabilizing factors can be detected by measuring the relative expression levels of long-lived versus short-lived transcripts in high-throughput gene expression experiments. We referred to this measure as the RNA stability score (RS-score). We calculated the RS-score for 726 HapMap3 samples using expression microarray data and carried out genome-wide tests of association between SNP markers and the RS-score. The SNP rs6137010 was found to be strongly associated with the RS-score in two Asian populations: Han Chinese from Beijing (CHB) and Japanese from Tokyo (JPT). Interestingly, this SNP appears to be a *cis*-eQTL for *SNRPB* in CHB and JPT. *SNRPB* is a core component of the spliceosome, and has previously been shown to affect the expression of many RNA processing factors. In addition, knockdown of *SNRPB* leads to significant decrease in the RS-score. Thus, these results suggest that the *cis*-eQTL of *SNRPB* may be directly responsible for inter-individual variation in the RS-score in Asian populations. We noted that more efforts are necessary to draw stronger conclusions from this study. The association between rs6137010 and *SNRPB* expression level should be supported by chromatin loop data (e.g. Hi-C data) because rs6137010 is localized more than 300 kb far from *SNRPB*. In addition, functional experiments (e.g. qPCR) should be carried out to confirm the role of *SNRPB* in RNA stability.

Chapter 5 developed a computational pipeline to pinpoint genetic variants that influence the rate of mRNA translation. The analysis of allele-specific translation (AST) can be affected by the biased mapping of short read sequences to the reference allele. To overcome this, we attempted to construct a haplotype-resolved genome for a given cell-type by incorporating high-throughput sequencing data that are publicly available for that cell-type. We then mapped both RNA-seq and Ribo-

seq data to the resulting haplotype-resolved genome and counted the number of reads from each haplotype separately. The Ribo-seq read counts were used to find genes associated with AST and the RNA-seq read counts were used to exclude the possibility that AST may be a consequence of allele-specific mRNA expression (ASE). Applying this pipeline for the datasets corresponding to HeLa cells, we identified 171 genes showing evidence of AST. Two heterozygous SNPs, located within the 5'UTR of two AST genes, were found to likely cause the AST. First, the SNP rs9960, with two alleles A and G, is located within the translation initiation sites of the *ATP5H* gene. The A allele is associated with much higher translation rate than the G allele. This suggests that the mutation A→G at this locus disturbs the translation initiation of *ATP5H*. Second, the SNP rs6122080, with two alleles A and G, is located at the conserved binding sites of the translation initiation factor *eIF4B* within the *SLCO4A1* gene. The A allele is associated with weak secondary structure of the 5'UTR of *SLCO4A1* and with no Ribo-seq reads. These results therefore suggest that the mutation G→A at this SNP changes the secondary structure of the 5'UTR of *SLCO4A1* and, consequently, affects the translation initiation of *SLCO4A1*. Notably, we only obtained a moderate number of genes (< 800) that have sufficient read counts for the AST tests. This is likely caused by two main problems: current Ribo-seq datasets do not have very high coverage; not all SNPs that are heterozygous in HeLa cells were genotyped by the HapMap project. Thus, we propose to get more Ribo-seq datasets in the public databases and use the HeLa haplotype-resolved genome published recently by Adey *et al.* [301] to yield better mapping results. Finally, experiments are necessary to confirm several of our key findings.

In summary, this thesis has tackled important questions in the areas of gene regulation and chromatin biology. We made use of large volumes of high-throughput genomics data that are available from public databases to investigate major gaps in these areas. The thesis sheds light on the genetic basis of RNA-stabilizing factors, the genetics of mRNA translation, DSB repair, and the role of NOR in the formation and function of the nucleolus.

Bibliography

1. Grummt I (1999) Regulation of mammalian ribosomal gene transcription by RNA polymerase I. Prog Nucleic Acid Res Mol Biol 62: 109-154.

2. Lee Y, Kim M, Han J, Yeom KH, Lee S, et al. (2004) MicroRNA genes are transcribed by RNA polymerase II. EMBO J 23: 4051-4060.

3. Alberts B, Wilson JH, Hunt T (2008) Molecular biology of the cell. New York: Garland Science. xxxiii, 1601, 1690 p. p.

4. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10: 252-263.

5. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 14: 283-291.

6. Cheng C, Gerstein M (2012) Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. Nucleic Acids Res 40: 553-568.

7. Cheng C, Alexander R, Min R, Leng J, Yip KY, et al. (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res 22: 1658-1667.

8. Lee TI, Young RA (2000) Transcription of eukaryotic protein-coding genes. Annu Rev Genet 34: 77-137.

9. Kadonaga JT (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. Cell 116: 247-257.

10. Parvin JD, Sharp PA (1993) DNA topology and a minimal set of basal factors for transcription by RNA polymerase II. Cell 73: 533-540.

11. Green MR (2000) TBP-associated factors (TAFIIs): multiple, selective transcriptional mediators in common complexes. Trends Biochem Sci 25: 59-63.

12. Buratowski S, Hahn S, Guarente L, Sharp PA (1989) Five intermediate complexes in transcription initiation by RNA polymerase II. Cell 56: 549-561.

13. Roeder RG (1996) The role of general initiation factors in transcription by RNA polymerase II. Trends Biochem Sci 21: 327-335.

14. Jiang CZ, Pugh BF (2009) Nucleosome positioning and gene regulation: advances through genomics. Nature Reviews Genetics 10: 161-172.

15. Bhaumik SR, Green MR (2002) Differential requirement of SAGA components for recruitment of TATA-box-binding protein to promoters in vivo. Mol Cell Biol 22: 7365-7371.

16. Basehoar AD, Zanton SJ, Pugh BF (2004) Identification and distinct regulation of yeast TATA box-containing genes. Cell 116: 699-709.

17. Huisinga KL, Pugh BF (2004) A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae. Mol Cell 13: 573-585.

18. Venters BJ, Pugh BF (2013) Genomic organization of human transcription initiation complexes. Nature 502: 53-58.

19. Pardee TS, Bangur CS, Ponticelli AS (1998) The N-terminal region of yeast TFIIB contains two adjacent functional domains involved in stable RNA polymerase II binding and transcription start site selection. J Biol Chem 273: 17859-17864.

20. Yan Q, Moreland RJ, Conaway JW, Conaway RC (1999) Dual roles for transcription factor IIF in promoter escape by RNA polymerase II. J Biol Chem 274: 35668-35675.

21. Orphanides G, Lagrange T, Reinberg D (1996) The general transcription factors of RNA polymerase II. Genes Dev 10: 2657-2683.

22. Kim TK, Ebright RH, Reinberg D (2000) Mechanism of ATP-dependent promoter melting by transcription factor IIH. Science 288: 1418-1422.

23. Kornberg RD (2005) Mediator and the mechanism of transcriptional activation. Trends Biochem Sci 30: 235-239.

24. Svejstrup JQ (2004) The RNA polymerase II transcription cycle: cycling through chromatin. Biochimica Et Biophysica Acta-Gene Structure and Expression 1677: 64-73.

25. Chang YF, Imam JS, Wilkinson MF (2007) The nonsense-mediated decay RNA surveillance pathway. Annu Rev Biochem 76: 51-74.

26. Brogna S, Wen J (2009) Nonsense-mediated mRNA decay (NMD) mechanisms. Nature Structural & Molecular Biology 16: 107-113.

27. Maquat LE (2004) Nonsense-mediated mRNA decay: Splicing, translation and mRNP dynamics. Nature Reviews Molecular Cell Biology 5: 89-99.

28. Schoenberg DR, Maquat LE (2012) Regulation of cytoplasmic mRNA decay. Nat Rev Genet 13: 246-259.

29. Conti E, Izaurralde E (2005) Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. Current Opinion in Cell Biology 17: 316-325.

30. Grimson A, O'Connor S, Newman CL, Anderson P (2004) SMG-1 is a phosphatidylinositol kinase-related protein kinase required for nonsense-mediated mRNA Decay in Caenorhabditis elegans. Mol Cell Biol 24: 7483-7490.

31. Maquat LE (2004) Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol 5: 89-99.

32. Le Hir H, Izaurralde E, Maquat LE, Moore MJ (2000) The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. EMBO J 19: 6860-6869.

33. Isken O, Maquat LE (2008) The multiple lives of NMD factors: balancing roles in gene and genome regulation. Nat Rev Genet 9: 699-712.

34. Murray EL, Schoenberg DR (2007) A+U-rich instability elements differentially activate 5'-3' and 3'-5' mRNA decay. Mol Cell Biol 27: 2791-2799.

35. Yamashita A, Chang TC, Yamashita Y, Zhu W, Zhong Z, et al. (2005) Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover. Nature Structural & Molecular Biology 12: 1054-1063.

36. Stoecklin G, Stubbs T, Kedersha N, Wax S, Rigby WF, et al. (2004) MK2-induced tristetraprolin:14-3-3 complexes prevent stress granule association and ARE-mRNA decay. EMBO J 23: 1313-1324.

37. Abdelmohsen K, Gorospe M (2010) Posttranscriptional regulation of cancer traits by HuR. Wiley Interdiscip Rev RNA 1: 214-229.

38. Mahtani KR, Brook M, Dean JL, Sully G, Saklatvala J, et al. (2001) Mitogen-activated protein kinase p38 controls the expression and posttranslational modification of tristetraprolin, a regulator of tumor necrosis factor alpha mRNA stability. Mol Cell Biol 21: 6461-6469.

39. Chrestensen CA, Schroeder MJ, Shabanowitz J, Hunt DF, Pelo JW, et al. (2004) MAPKAP kinase 2 phosphorylates tristetraprolin on in vivo sites including Ser178, a site required for 14-3-3 binding. J Biol Chem 279: 10176-10184.

40. Maitra S, Chou CF, Luber CA, Lee KY, Mann M, et al. (2008) The AU-rich element mRNA decay-promoting activity of BRF1 is regulated by mitogen-activated protein kinase-activated protein kinase 2. Rna-a Publication of the Rna Society 14: 950-959.

41. Lykke-Andersen J, Wagner E (2005) Recruitment and activation of mRNA decay enzymes by two ARE-mediated decay activation domains in the proteins TTP and BRF-1. Genes & Development 19: 351-361.

42. White EJ, Brewer G, Wilson GM (2013) Post-transcriptional control of gene expression by AUF1: mechanisms, physiological targets, and regulation. Biochim Biophys Acta 1829: 680-688.

43. Winzen R, Thakur BK, Dittrich-Breiholz O, Shah M, Redich N, et al. (2007) Functional analysis of KSRP interaction with the AU-rich element of interleukin-8 and identification of inflammatory mRNA targets. Mol Cell Biol 27: 8388-8400.

44. Jackson RJ, Hellen CU, Pestova TV (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. Nat Rev Mol Cell Biol 11: 113-127.

45. Maag D, Fekete CA, Gryczynski Z, Lorsch JR (2005) A conformational change in the eukaryotic translation preinitiation complex and release of eIF1 signal recognition of the start codon. Mol Cell 17: 265-275.

46. Unbehaun A, Borukhov SI, Hellen CUT, Pestova TV (2004) Release of initiation factors from 48S complexes during ribosomal subunit joining and the link between establishment of codon-anticodon base-pairing and hydrolysis of eIF2-bound GTP. Genes & Development 18: 3078-3093.

47. Siridechadilok B, Fraser CS, Hall RJ, Doudna JA, Nogales E (2005) Structural roles for human translation factor eIF3 in initiation of protein synthesis. Science 310: 1513-1515.

48. Passmore LA, Schmeing TM, Maag D, Applefield DJ, Acker MG, et al. (2007) The eukaryotic translation initiation factors eIF1 and eIF1A induce an open conformation of the 40S ribosome. Molecular Cell 26: 41-50.

49. Borman AM, Michel YM, Kean KM (2000) Biochemical characterisation of cap-poly(A) synergy in rabbit reticulocyte lysates: the eIF4G-PABP interaction increases the functional affinity of eIF4E for the capped mRNA 5 '-end. Nucleic Acids Research 28: 4068-4075.

50. Schutz P, Bumann M, Oberholzer AE, Bieniossek C, Trachsel H, et al. (2008) Crystal structure of the yeast eIF4A-eIF4G complex: an RNA-helicase controlled by protein-protein interactions. Proc Natl Acad Sci U S A 105: 9564-9569.

51. Marintchev A, Edmonds KA, Marintcheva B, Hendrickson E, Oberer M, et al. (2009) Topology and regulation of the human eIF4A/4G/4H helicase complex in translation initiation. Cell 136: 447-460.

52. Muckenthaler M, Gray NK, Hentze MW (1998) IRP-1 binding to ferritin mRNA prevents the recruitment of the small ribosomal subunit by the cap-binding complex eIF4F. Mol Cell 2: 383-388.

53. Pierce B (2010) Genetics: A Conceptual Approach. New York: W. H. Freeman and Company.

54. Kamakaka RT, Biggins S (2005) Histone variants: deviants? Genes Dev 19: 295-310.

55. Sarma K, Reinberg D (2005) Histone variants meet their match. Nature Reviews Molecular Cell Biology 6: 139-149.

56. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, et al. (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. Science 309: 626-630.

57. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837.

58. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, et al. (2008) Nucleosome organization in the Drosophila genome. Nature 453: 358-362.

59. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, et al. (2011) Determinants of nucleosome organization in primary human cells. Nature 474: 516-520.

60. Lee W, Tillo D, Bray N, Morse RH, Davis RW, et al. (2007) A high-resolution atlas of nucleosome occupancy in yeast. Nature Genetics 39: 1235-1244.

61. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, et al. (2007) Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. Nature 446: 572-576.

62. Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, et al. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Research 18: 1073-1083.

63. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, et al. (2006) A genomic code for nucleosome positioning. Nature 442: 772-778.

64. Mito Y, Henikoff JG, Henikoff S (2007) Histone replacement marks the boundaries of cis-regulatory domains. Science 315: 1408-1411.

65. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, et al. (2008) Dynamic regulation of nucleosome positioning in the human genome. Cell 132: 887-898.

66. Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, et al. (2008) Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. PLoS Biol 6: e65.

67. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, et al. (2008) A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Research 18: 1051-1063.

68. Rhee HS, Pugh BF (2012) Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature 483: 295-301.

69. Lickwar CR, Mueller F, Hanlon SE, McNally JG, Lieb JD (2012) Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. Nature 484: 251-255.

70. Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, et al. (2012) Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. Genome Research 22: 1735-1747.

71. Rhee HS, Pugh BF (2012) ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. Curr Protoc Mol Biol Chapter 21: Unit 21 24.

72. Varga-Weisz PD, Wilm M, Bonte E, Dumas K, Mann M, et al. (1997) Chromatin-remodelling factor CHRAC contains the ATPases ISWI and topoisomerase II. Nature 388: 598-602.

73. Saha A, Wittmeyer J, Cairns BR (2006) Mechanisms for nucleosome movement by ATP-dependent chromatin remodeling complexes. Results Probl Cell Differ 41: 127-148.

74. Kagalwala MN, Glaus BJ, Dang WW, Zofall M, Bartholomew B (2004) Topography of the ISW2-nucleosome complex: insights into nucleosome spacing and chromatin remodeling. Embo Journal 23: 2092-2104.

75. Ferreira H, Owen-Hughes T (2006) Lighting up nucleosome spacing. Nature Structural & Molecular Biology 13: 1047-1049.

76. Fan Y, Nikitina T, Morin-Kensicki EM, Zhao J, Magnuson TR, et al. (2003) H1 linker histones are essential for mouse development and affect nucleosome spacing in vivo. Mol Cell Biol 23: 4559-4572.

77. Labrador M, Corces VG (2002) Setting the boundaries of chromatin domains and nuclear organization. Cell 111: 151-154.

78. Phillips JE, Corces VG (2009) CTCF: master weaver of the genome. Cell 137: 1194-1211.

79. Grewal SI, Jia S (2007) Heterochromatin revisited. Nat Rev Genet 8: 35-46.

80. Kouzarides T (2007) Chromatin modifications and their function. Cell 128: 693-705.

81. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, et al. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40: 897-903.

82. Jenuwein T, Allis CD (2001) Translating the histone code. Science 293: 1074-1080.

83. Nakayama J, Rice JC, Strahl BD, Allis CD, Grewal SI (2001) Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. Science 292: 110-113.

84. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315-326.

85. Zhou VW, Goren A, Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet 12: 7-18.

86. Pan G, Tian S, Nie J, Yang C, Ruotti V, et al. (2007) Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. Cell Stem Cell 1: 299-312.

87. van Attikum H, Gasser SM (2005) The histone code at DNA breaks: A guide to repair? Nature Reviews Molecular Cell Biology 6: 757-765.

88. van Attikum H, Gasser SM (2009) Crosstalk between histone modifications during the DNA damage response. Trends in Cell Biology 19: 207-217.

89. Celeste A, Petersen S, Romanienko PJ, Fernandez-Capetillo O, Chen HT, et al. (2002) Genomic instability in mice lacking histone H2AX. Science 296: 922-927.

90. Rogakou EP, Pilch DR, Orr AH, Ivanova VS, Bonner WM (1998) DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. J Biol Chem 273: 5858-5868.

91. Rusche LN, Kirchmaier AL, Rine J (2003) The establishment, inheritance, and function of silenced chromatin in Saccharomyces cerevisiae. Annu Rev Biochem 72: 481-516.

92. Sullivan B, Karpen G (2001) Centromere identity in Drosophila is not determined in vivo by replication timing. J Cell Biol 154: 683-690.

93. Ahmad K, Henikoff S (2001) Centromeres are specialized replication domains in heterochromatin. J Cell Biol 153: 101-110.

94. Dolfini S, Courgeon AM, Tiepolo L (1970) The cell cycle of an established line of Drosophila melanogaster cells in vitro. Experientia 26: 1020-1021.

95. Chen CL, Rappailles A, Duquenne L, Huvet M, Guilbaud G, et al. (2010) Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res 20: 447-457.

96. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799-816.

97. Vogelauer M, Rubbi L, Lucas I, Brewer BJ, Grunstein M (2002) Histone acetylation regulates the time of replication origin firing. Mol Cell 10: 1223-1233.

98. Venters BJ, Pugh BF (2009) A canonical promoter organization of the transcription machinery and its regulators in the Saccharomyces genome. Genome Research 19: 360-371.

99. Hartley PD, Madhani HD (2009) Mechanisms that Specify Promoter Nucleosome Location and Identity. Cell 137: 445-458.

100. Li B, Carey M, Workman JL (2007) The role of chromatin during transcription. Cell 128: 707-719.

101. Schwartz S, Meshorer E, Ast G (2009) Chromatin organization marks exon-intron structure. Nature Structural & Molecular Biology 16: 990-995.

102. Andersson R, Enroth S, Rada-Iglesias A, Wadelius C, Komorowski J (2009) Nucleosomes are well positioned in exons and carry characteristic histone modifications. Genome Research 19: 1732-1741.

103. Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, et al. (2009) Nucleosome positioning as a determinant of exon recognition. Nature Structural & Molecular Biology 16: 996-U124.

104. Zaret KS, Carroll JS (2011) Pioneer transcription factors: establishing competence for gene expression. Genes Dev 25: 2227-2241.

105. Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M (2010) Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A 107: 2926-2931.

106. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, et al. (2012) Modeling gene expression using chromatin features in various cellular contexts. Genome Biol 13: R53.

107. Shahbazian MD, Grunstein M (2007) Functions of site-specific histone acetylation and deacetylation. Annu Rev Biochem 76: 75-100.

108. Guillemette B, Gaudreau L (2006) Reuniting the contrasting functions of H2A.Z. Biochemistry and Cell Biology-Biochimie Et Biologie Cellulaire 84: 528-535.

109. Agalioti T, Chen G, Thanos D (2002) Deciphering the transcriptional histone acetylation code for a human gene. Cell 111: 381-392.

110. Joshi AA, Struhl K (2005) Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation. Mol Cell 20: 971-978.

111. Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK, et al. (2005) Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. Cell 123: 581-592.

112. Ong CT, Corces VG (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature Reviews Genetics 12: 283-293.

113. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics 39: 311-318.

114. Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO, et al. (2001) Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature 410: 120-124.

115. Schotta G, Lachner M, Sarma K, Ebert A, Sengupta R, et al. (2004) A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. Genes & Development 18: 1251-1262.

116. Talasz H, Lindner HH, Sarg B, Helliger W (2005) Histone H4-lysine 20 monomethylation is increased in promoter and coding regions of active genes and correlates with hyperacetylation. Journal of Biological Chemistry 280: 38814-38822.

117. Mizuguchi G, Tsukiyama T, Wisniewski J, Wu C (1997) Role of nucleosome remodeling factor NURF in transcriptional activation of chromatin. Mol Cell 1: 141-150.

118. Lauberth SM, Nakayama T, Wu X, Ferris AL, Tang Z, et al. (2013) H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. Cell 152: 1021-1036.

119. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, et al. (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. Cell 122: 517-527.

120. Guttman M, Amit I, Garber M, French C, Lin MF, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458: 223-227.

121. Simon JA, Kingston RE (2009) Mechanisms of polycomb gene silencing: knowns and unknowns. Nat Rev Mol Cell Biol 10: 697-708.

122. Hon G, Wang W, Ren B (2009) Discovery and annotation of functional chromatin signatures in the human genome. PLoS Comput Biol 5: e1000566.

123. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43-49.

124. Hoffman MM, Buske OJ, Wang J, Weng ZP, Bilmes JA, et al. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature Methods 9: 473-U488.

125. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459: 108-112.

126. Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol 28: 817-825.

127. Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, et al. (2010) Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell 143: 212-224.

128. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, et al. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods 9: 473-476.

129. Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. Nat Methods 9: 215-216.

130. Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, et al. (2012) ENCODE whole-genome data in the UCSC Genome Browser: update 2012. Nucleic Acids Research 40: D912-D917.

131. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10: 669-680.

132. Furey TS (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet 13: 840-852.

133. Orlando V (2000) Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. Trends Biochem Sci 25: 99-104.

134. O'Neill LP, Turner BM (2003) Immunoprecipitation of native chromatin: NChIP. Methods 31: 76-82.

135. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.

136. Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11: 473-483.

137. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

138. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359.

139. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

140. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851-1858.

141. Li R, Yu C, Li Y, Lam TW, Yiu SM, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966-1967.

142. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods 6: S22-32.

143. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9: R137.

144. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, et al. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol 26: 1293-1300.

145. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res 36: 5221-5231.

146. Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26: 1351-1359.

147. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.

148. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4: 651-657.

149. Zang C, Schones DE, Zeng C, Cui K, Zhao K, et al. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25: 1952-1958.

150. Rashid NU, Giresi PG, Ibrahim JG, Sun W, Lieb JD (2011) ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. Genome Biol 12: R67.

151. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, et al. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 27: 66-75.

152. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497-1502.

153. Bailey TL (2011) DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics 27: 1653-1659.

154. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, et al. (2013) Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. PLoS Comput Biol 9: e1003326.

155. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63.

156. Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 12: 87-98.

157. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105-1111.

158. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15-21.

159. De Bona F, Ossowski S, Schneeberger K, Ratsch G (2008) Optimal spliced alignments of short sequence reads. Bioinformatics 24: i174-180.

160. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621-628.

161. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28: 511-U174.

162. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7: 562-578.

163. Dickey JS, Redon CE, Nakamura AJ, Baird BJ, Sedelnikova OA, et al. (2009) H2AX: functional roles and potential applications. Chromosoma 118: 683-692.

164. Stucki M, Jackson SP (2006) gammaH2AX and MDC1: anchoring the DNA-damage-response machinery to broken chromosomes. DNA Repair (Amst) 5: 534-543.

165. Fernandez-Capetillo O, Lee A, Nussenzweig M, Nussenzweig A (2004) H2AX: the histone guardian of the genome. DNA Repair (Amst) 3: 959-967.

166. Fillingham J, Keogh MC, Krogan NJ (2006) GammaH2AX and its role in DNA double-strand break repair. Biochemistry and Cell Biology-Biochimie Et Biologie Cellulaire 84: 568-577.

167. Paull TT, Rogakou EP, Yamazaki V, Kirchgessner CU, Gellert M, et al. (2000) A critical role for histone H2AX in recruitment of repair factors to nuclear foci after DNA damage. Curr Biol 10: 886-895.

168. Bassing CH, Chua KF, Sekiguchi J, Suh H, Whitlow SR, et al. (2002) Increased ionizing radiation sensitivity and genomic instability in the absence of histone H2AX. Proc Natl Acad Sci U S A 99: 8173-8178.

169. Bassing CH, Suh H, Ferguson DO, Chua KF, Manis J, et al. (2003) Histone H2AX: a dosage-dependent suppressor of oncogenic translocations and tumors. Cell 114: 359-370.

170. Celeste A, Difilippantonio S, Difilippantonio MJ, Fernandez-Capetillo O, Pilch DR, et al. (2003) H2AX haploinsufficiency modifies genomic stability and tumor susceptibility. Cell 114: 371-383.

171. Cook PJ, Ju BG, Telese F, Wang X, Glass CK, et al. (2009) Tyrosine dephosphorylation of H2AX modulates apoptosis and survival decisions. Nature 458: 591-596.

172. Iacovoni JS, Caron P, Lassadi I, Nicolas E, Massip L, et al. (2010) High-resolution profiling of gammaH2AX around DNA double strand breaks in the mammalian genome. EMBO J 29: 1446-1457.

173. Cowell IG, Sunter NJ, Singh PB, Austin CA, Durkacz BW, et al. (2007) gamma H2AX Foci Form Preferentially in Euchromatin after Ionising-Radiation. PLoS One 2.

174. Kim JA, Kruhlak M, Dotiwala F, Nussenzweig A, Haber JE (2007) Heterochromatin is refractory to gamma-H2AX modification in yeast and mammals. Journal of Cell Biology 178: 209-218.

175. Kruhlak MJ, Celeste A, Dellaire G, Fernandez-Capetillo O, Muller WG, et al. (2006) Changes in chromatin structure and mobility in living cells at sites of DNA double-strand breaks. J Cell Biol 172: 823-834.

176. Niforou KM, Anagnostopoulos AK, Vougas K, Kittas C, Gorgoulis VG, et al. (2008) The proteome profile of the human osteosarcoma U2OS cell line. Cancer Genomics Proteomics 5: 63-78.

177. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research 38: 1767-1771.

178. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, et al. (2008) Whole-genome sequencing and variant discovery in C. elegans. Nat Methods 5: 183-188.

179. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, et al. (2008) A large genome center's improvements to the Illumina sequencing system. Nat Methods 5: 1005-1010.

180. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22: 1813-1831.

181. Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, et al. (2009) Human mutation rate associated with DNA replication timing. Nat Genet 41: 393-395.

182. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, et al. (2013) The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res 41: D64-69.

183. de Koning APJ, Gu WJ, Castoe TA, Batzer MA, Pollock DD (2011) Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. PLoS genetics 7.

184. Birchler JA, Bhadra MP, Bhadra U (2000) Making noise about silence: repression of repeated genes in animals. Current Opinion in Genetics & Development 10: 211-216.

185. van Rij RP, Andino R (2004) RNAi - A guide to gene silencing. Science 303: 1978-1979.

186. Martens JHA, O'Sullivan RJ, Braunschweig U, Opravil S, Radolf M, et al. (2005) The profile of repeat-associated histone lysine methylation states in the mouse epigenome. Embo Journal 24: 800-812.

187. Smit AFA, Hubley, R. and Green, P. RepeatMasker.

188. Gieni RS, Chan GKT, Hendzel MJ (2008) Epigenetics regulate centromere formation and kinetochore function. Journal of Cellular Biochemistry 104: 2027-2039.

189. Goodarzi AA, Noon AT, Jeggo PA (2009) The impact of heterochromatin on DSB repair. Biochemical Society Transactions 37: 569-576.

190. Cann KL, Dellaire G (2011) Heterochromatin and the DNA damage response: the need to relax. Biochemistry and Cell Biology-Biochimie Et Biologie Cellulaire 89: 45-60.

191. Kruhlak MJ, Celeste A, Dellaire G, Fernandez-Capetillo O, Muller WG, et al. (2006) Changes in chromatin structure and mobility in living cells at sites of DNA double-strand breaks. Journal of Cell Biology 172: 823-834.

192. Seo J, Kim SC, Lee HS, Kim JK, Shon HJ, et al. (2012) Genome-wide profiles of H2AX and gamma-H2AX differentiate endogenous and exogenous DNA damage hotspots in human cells. Nucleic Acids Res 40: 5965-5974.

193. Schneider U, Schwenk HU, Bornkamm G (1977) Characterization of EBV-genome negative "null" and "T" cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed non-Hodgkin lymphoma. Int J Cancer 19: 621-626.

194. Seo J, Kim K, Chang DY, Kang HB, Shin EC, et al. (2013) Genome-wide reorganization of histone H2AX toward particular fragile sites on cell activation. Nucleic Acids Res.

195. Seo J, Kim SC, Lee HS, Kim JK, Shon HJ, et al. (2012) Genome-wide profiles of H2AX and gamma-H2AX differentiate endogenous and exogenous DNA damage hotspots in human cells. Nucleic Acids Res.

196. Olson MOJE (2011) The Nucleolus: Springer.

197. Shiue CN, Arabi A, Wright AP (2010) Nucleolar organization, growth control and cancer. Epigenetics 5.

198. Derenzini M, Trere D, Pession A, Govoni M, Sirri V, et al. (2000) Nucleolar size indicates the rapidity of cell proliferation in cancer tissues. J Pathol 191: 181-186.

199. Grandori C, Gomez-Roman N, Felton-Edkins ZA, Ngouenet C, Galloway DA, et al. (2005) c-Myc binds to human ribosomal DNA and stimulates transcription of rRNA genes by RNA polymerase I. Nat Cell Biol 7: 311-318.

200. Budde A, Grummt I (1999) p53 represses ribosomal gene transcription. Oncogene 18: 1119-1124.

201. Hannan KM, Hannan RD, Smith SD, Jefferson LS, Lun M, et al. (2000) Rb and p130 regulate RNA polymerase I transcription: Rb disrupts the interaction between UBF and SL-1. Oncogene 19: 4988-4999.

202. Grummt I (2003) Life on a planet of its own: regulation of RNA polymerase I transcription in the nucleolus. Genes Dev 17: 1691-1702.

203. Bywater MJ, Poortinga G, Sanij E, Hein N, Peck A, et al. (2012) Inhibition of RNA polymerase I as a therapeutic strategy to promote cancer-specific activation of p53. Cancer Cell 22: 51-65.

204. Boisvert FM, van Koningsbruggen S, Navascues J, Lamond AI (2007) The multifunctional nucleolus. Nat Rev Mol Cell Biol 8: 574-585.

205. Ganley AR, Ide S, Saka K, Kobayashi T (2009) The effect of replication initiation on gene amplification in the rDNA and its relationship to aging. Mol Cell 35: 683-693.

206. Zhang LF, Huynh KD, Lee JT (2007) Perinucleolar targeting of the inactive X during S phase: evidence for a role in the maintenance of silencing. Cell 129: 693-706.

207. Visintin R, Hwang ES, Amon A (1999) Cfi1 prevents premature exit from mitosis by anchoring Cdc14 phosphatase in the nucleolus. Nature 398: 818-823.

208. Savino TM, Gebrane-Younes J, De Mey J, Sibarita JB, Hernandez-Verdun D (2001) Nucleolar assembly of the rRNA processing machinery in living cells. J Cell Biol 153: 1097-1110.

209. Henderson AS, Warburton D, Atwood KC (1972) Location of ribosomal DNA in the human chromosome complement. Proc Natl Acad Sci U S A 69: 3394-3398.

210. Britton-Davidian J, Cazaux B, Catalan J (2012) Chromosomal dynamics of nucleolar organizer regions (NORs) in the house mouse: micro-evolutionary insights. Heredity (Edinb) 108: 68-74.

211. Nemeth A, Langst G (2011) Genome organization in and around the nucleolus. Trends Genet 27: 149-156.

212. McStay B, Grummt I (2008) The epigenetics of rRNA genes: from molecular to chromosome biology. Annu Rev Cell Dev Biol 24: 131-157.

213. Stults DM, Killen MW, Pierce HH, Pierce AJ (2008) Genomic architecture and inheritance of human ribosomal RNA gene clusters. Genome Res 18: 13-18.

214. Santoro R, Li J, Grummt I (2002) The nucleolar remodeling complex NoRC mediates heterochromatin formation and silencing of ribosomal gene transcription. Nature Genetics 32: 393-396.

215. Zentner GE, Saiakhova A, Manaenkov P, Adams MD, Scacheri PC (2011) Integrative genomic analysis of human ribosomal DNA. Nucleic Acids Res 39: 4949-4960.

216. Floutsakou I, Agrawal S, Nguyen TT, Seoighe C, Ganley AR, et al. (2013) The shared genomic architecture of human nucleolar organizer regions. Genome Research.

217. Sakai K, Ohta T, Minoshima S, Kudoh J, Wang Y, et al. (1995) Human ribosomal RNA gene cluster: identification of the proximal end containing a novel tandem repeat sequence. Genomics 26: 521-526.

218. Gonzalez IL, Sylvester JE (1997) Beyond ribosomal DNA: on towards the telomere. Chromosoma 105: 431-437.

219. Worton RG, Sutherland J, Sylvester JE, Willard HF, Bodrug S, et al. (1988) Human ribosomal RNA genes: orientation of the tandem array and conservation of the 5' end. Science 239: 64-68.

220. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 17: 877-885.

221. Wu C, Bingham PM, Livak KJ, Holmgren R, Elgin SC (1979) The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. Cell 16: 797-806.

222. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, et al. (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res 16: 123-131.

223. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560.

224. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA (2007) A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130: 77-88.

225. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, et al. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 120: 169-181.

226. van de Nobelen S, Rosa-Garrido M, Leers J, Heath H, Soochit W, et al. (2010) CTCF regulates the local epigenetic state of ribosomal DNA repeats. Epigenetics Chromatin 3: 19.

227. Hernandez-Hernandez A, Soto-Reyes E, Ortiz R, Arriaga-Canon C, Echeverria-Martinez OM, et al. (2012) Changes of the nucleolus architecture in absence of the nuclear factor CTCF. Cytogenet Genome Res 136: 89-96.

228. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, et al. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res 38: D105-110.

229. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511-515.

230. She X, Horvath JE, Jiang Z, Liu G, Furey TS, et al. (2004) The structure and evolution of centromeric transition regions within the human genome. Nature 430: 857-864.

231. Therman E, Susman B, Denniston C (1989) The nonrandom participation of human acrocentric chromosomes in Robertsonian translocations. Ann Hum Genet 53: 49-65.

232. Nemeth A, Conesa A, Santoyo-Lopez J, Medina I, Montaner D, et al. (2010) Initial genomics of the human nucleolus. PLoS Genet 6: e1000889.

233. Derenzini M, Montanaro L, Trere D (2009) What the nucleolus says to a tumour pathologist. Histopathology 54: 753-762.

234. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, et al. (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res 21: 1757-1767.

235. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2011) The UCSC Genome Browser database: update 2011. Nucleic Acids Res 39: D876-882.

236. Boyle AP, Guinney J, Crawford GE, Furey TS (2008) F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics 24: 2537-2538.

237. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

238. Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12: 656-664.

239. Ross J (1995) mRNA stability in mammalian cells. Microbiol Rev 59: 423-450.

240. Guhaniyogi J, Brewer G (2001) Regulation of mRNA stability in mammalian cells. Gene 265: 11-23.

241. Cheadle C, Fan J, Cho-Chung YS, Werner T, Ray J, et al. (2005) Control of gene expression during T cell activation: alternate regulation of mRNA transcription and mRNA stability. Bmc Genomics 6: 75.

242. Lam LT, Pickeral OK, Peng AC, Rosenwald A, Hurt EM, et al. (2001) Genomic-scale measurement of mRNA turnover and the mechanisms of action of the anti-cancer drug flavopiridol. Genome Biol 2: RESEARCH0041.

243. Bregman A, Avraham-Kelbert M, Barkai O, Duek L, Guterman A, et al. (2011) Promoter elements regulate cytoplasmic mRNA decay. Cell 147: 1473-1483.

244. Trcek T, Larson DR, Moldón A, Query CC, Singer RH (2011) Single-molecule mRNA decay measurements reveal promoter- regulated mRNA stability in yeast. Cell 147: 1484-1497.

245. Narsai R, Howell KA, Millar AH, O'Toole N, Small I, et al. (2007) Genome-wide analysis of mRNA decay rates and their determinants in Arabidopsis thaliana. Plant Cell 19: 3418-3436.

246. Sharova LV, Sharov AA, Nedorezov T, Piao Y, Shaik N, et al. (2009) Database for mRNA Half-Life of 19 977 Genes Obtained by DNA Microarray Analysis of Pluripotent and Differentiating Mouse Embryonic Stem Cells. DNA Research 16: 45-58.

247. Clark MB, Johnston RL, Inostroza-Ponta M, Fox AH, Fortini E, et al. (2012) Genome-wide analysis of long noncoding RNA stability. Genome Res 22: 885-898.

248. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, et al. (2011) Global quantification of mammalian gene expression control. Nature 473: 337-342.

249. Raghavan A, Ogilvie RL, Reilly C, Abelson ML, Raghavan S, et al. (2002) Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. Nucleic Acids Res 30: 5529-5538.

250. Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, et al. (2003) Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. Genome Res 13: 1863-1872.

251. Goodarzi H, Najafabadi HS, Oikonomou P, Greco TM, Fish L, et al. (2012) Systematic discovery of structural elements governing stability of mammalian messenger RNAs. Nature 485: 264-268.

252. Friedel CC, Dolken L, Ruzsics Z, Koszinowski UH, Zimmer R (2009) Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. Nucleic Acids Research 37.

253. Tani H, Mizutani R, Salam KA, Tano K, Ijiri K, et al. (2012) Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res 22: 947-956.

254. Kontoyiannis D, Pasparakis M, Pizarro TT, Cominelli F, Kollias G (1999) Impaired on/off regulation of TNF biosynthesis in mice lacking TNF AU-rich. Immunity 10: 387-398.

255. Misquitta CM, Iyer VR, Werstiuk ES, Grover AK (2001) The role of 3'-untranslated region (3'-UTR) mediated mRNA stability in. Mol Cell Biochem 224: 53-67.

256. Eberhardt W, Doller A, Akool el S, Pfeilschifter J (2007) Modulation of mRNA stability as a novel therapeutic approach. Pharmacol Ther 114: 56-73.

257. Keene JD, Tenenbaum SA (2002) Eukaryotic mRNPs may represent posttranscriptional operons. Mol Cell 9: 1161-1167.

258. Kishore S, Luber S, Zavolan M (2010) Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. Brief Funct Genomics 9: 391-404.

259. Dreyfuss G, Matunis MJ, Pinol-Roma S, Burd CG (1993) hnRNP proteins and the biogenesis of mRNA. Annu Rev Biochem 62: 289-321.

260. Chaudhury A, Chander P, Howe PH (2010) Heterogeneous nuclear ribonucleoproteins (hnRNPs) in cellular processes: Focus on. Rna 16: 1449-1462.

261. Thiele BJ, Doller A, Kahne T, Pregla R, Hetzer R, et al. (2004) RNA-binding proteins heterogeneous nuclear ribonucleoprotein A1, E1, and K are involved in post-transcriptional control of collagen I and III synthesis. Circ Res 95: 1058-1066.

262. Pai AA, Cain CE, Mizrahi-Man O, De Leon S, Lewellen N, et al. (2012) The Contribution of RNA Decay Quantitative Trait Loci to Inter-Individual Variation in Steady-State Gene Expression Levels. PLoS genetics 8: e1003000.

263. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, et al. (2012) Patterns of cis regulatory variation in diverse human populations. PLoS Genet 8: e1002639.

264. Saltzman AL, Pan Q, Blencowe BJ (2011) Regulation of alternative splicing by the core spliceosomal machinery. Genes Dev 25: 373-384.

265. Cheng C, Fu XP, Alves P, Gerstein M (2009) mRNA expression profiles show differential regulatory effects of microRNAs between estrogen receptor-positive and estrogen receptor-negative breast cancer. Genome Biology 10.

266. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38: 904-909.

267. Chang CW, Cheng WC, Chen CR, Shu WY, Tsai ML, et al. (2011) Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. PLoS One 6: e22859.

268. Benjamini Y, Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B: 289-300.

269. R Development Core Team (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing ISBN 3-900051-07-0.

270. Huang dW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1-13.

271. Gupta SK, Carmi S, Waldman Ben-Asher H, Tkacz ID, Naboishchikov I, et al. (2013) Basal splicing factors regulate the stability of mature mRNAs in trypanosomes. J Biol Chem 288: 4991-5006.

272. Lemaire R, Prasad J, Kashima T, Gustafson J, Manley JL, et al. (2002) Stability of a PKCI-1-related mRNA is controlled by the splicing factor ASF/SF2: a novel function for SR proteins. Genes & Development 16: 594-607.

273. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, et al. (2007) ArrayExpress - a public database of microarray experiments and gene expression profiles. Nucleic Acids Research 35: D747-D750.

274. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58.

275. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.

276. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190.

277. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997-1004.

278. Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11: 459-463.

279. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6: 95-108.

280. Komili S, Silver PA (2008) Coupling and coordination in gene expression processes: a systems biology view. Nat Rev Genet 9: 38-48.

281. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nature Genetics 39: 1217-1224.

282. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464: 773-U151.

283. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768-772.

284. Wu L, Candille SI, Choi Y, Xie D, Jiang L, et al. (2013) Variation and genetic control of protein abundance in humans. Nature 499: 79-82.

285. Ge B, Pokholok DK, Kwan T, Grundberg E, Morcos L, et al. (2009) Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. Nature Genetics 41: 1216-1222.

286. Pastinen T, Ge B, Gurd S, Gaudin T, Dore C, et al. (2005) Mapping common regulatory variants to human haplotypes. Hum Mol Genet 14: 3963-3971.

287. Ong SE, Mann M (2005) Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol 1: 252-262.

288. Hinkson IV, Elias JE (2011) The dynamic state of protein turnover: It's about time. Trends in Cell Biology 21: 293-303.

289. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science 324: 218-223.

290. Guo H, Ingolia NT, Weissman JS, Bartel DP (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 466: 835-840.

291. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, et al. (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics 25: 3207-3212.

292. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987-2993.

293. Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. Nat Methods 9: 179-181.

294. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2008) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 36: D13-21.

295. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12: 443-451.

296. Trapnell C, Salzberg SL (2009) How to map billions of short reads onto genomes. Nature Biotechnology 27: 455-457.

297. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57-74.

298. Kozak M (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res 15: 8125-8148.

299. Kozak M (1986) Point Mutations Define a Sequence Flanking the Aug Initiator Codon That Modulates Translation by Eukaryotic Ribosomes. Cell 44: 283-292.

300. Boada M, Antúnez C, Ramírez-Lorca R, DeStefano A, González-Pérez A, et al. (2013) ATP5H/KCTD2 locus is associated with Alzheimer's disease risk. Molecular psychiatry.

301. Adey A, Burton JN, Kitzman JO, Hiatt JB, Lewis AP, et al. (2013) The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. Nature 500: 207-211.

302. Zur H, Tuller T (2013) New universal rules of eukaryotic translation initiation fidelity. PLoS Comput Biol 9: e1003136.

303. Methot N, Pickett G, Keene JD, Sonenberg N (1996) In vitro RNA selection identifies RNA ligands that specifically bind to eukaryotic translation initiation factor 4B: the role of the RNA remotif. Rna 2: 38-50.

304. Delyfer M-N, Raffelsberger W, Mercier D, Korobelnik J-F, Gaudric A, et al. (2011) Transcriptomic analysis of human retinal detachment reveals both inflammatory response and photoreceptor death. PLoS One 6: e28791.

305. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, et al. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proceedings of the National Academy of Sciences of the United States of America 101: 7287-7292.

306. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc 7: 1534-1550.

307. Lee S, Liu B, Huang SX, Shen B, Qian SB (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A 109: E2424-2432.

308. Reid DW, Nicchitta CV (2012) Primary role for endoplasmic reticulum-bound ribosomes in cellular translation identified by ribosome profiling. J Biol Chem 287: 5518-5527.

309. Liu B, Han Y, Qian SB (2013) Cotranslational response to proteotoxic stress by elongation pausing of ribosomes. Mol Cell 49: 453-463.

310. Stadler M, Fire A (2011) Wobble base-pairing slows in vivo translation elongation in metazoans. Rna 17: 2063-2073.

311. McCarthy M (2013) NIH and family of Henrietta Lacks reach agreement on access to HeLa genome. BMJ 347: f5041.

312. Hudson KL, Collins FS (2013) Biospecimen policy: Family matters. Nature 500: 141-142.

313. Satya RV, Zavaljevski N, Reifman J (2012) A new strategy to reduce allelic bias in RNA-Seq readmapping. Nucleic Acids Res 40: e127.

314. Delaneau O, Zagury JF, Marchini J (2013) Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 10: 5-6.

315. Song LY, Zhang ZC, Grasfeder LL, Boyle AP, Giresi PG, et al. (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Research 21: 1757-1767.

316. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, et al. (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482: 390-394.

317. Wang D, Rendon A, Wernisch L (2013) Transcription factor and chromatin features predict genes associated with eQTLs. Nucleic Acids Res 41: 1450-1463.

Appendix A – Integrative Analysis of mRNA Expression and Half-life Data Reveals *Trans*-acting Genetic Variants Associated with Increased Expression of Stable Transcripts

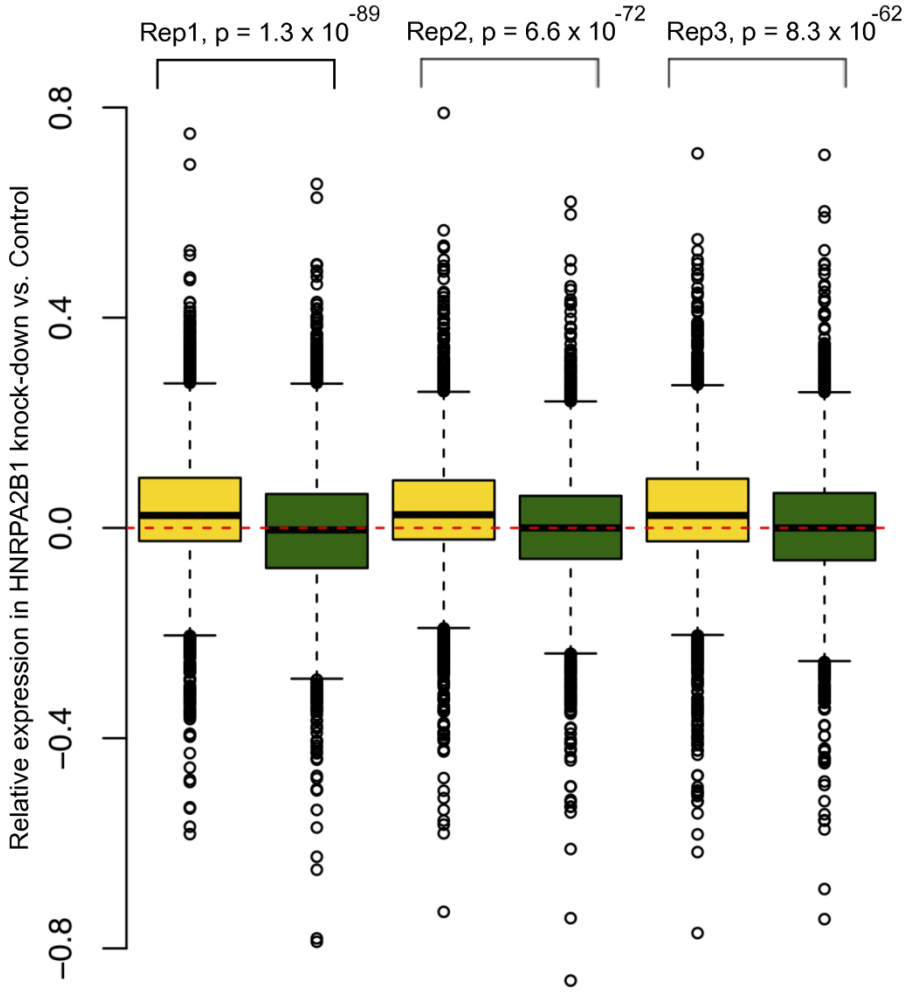


Figure S1: Gene expression levels in *HNRNPA2B1* knockdown relative to control are shown separately for genes expressing short-lived (golden) and long-lived (dark green) RNAs in three independent replicates (Rep1, Rep2, and Rep3). P-values are from Wilcoxon rank sum tests that were used to compare expression levels between these two groups of genes.

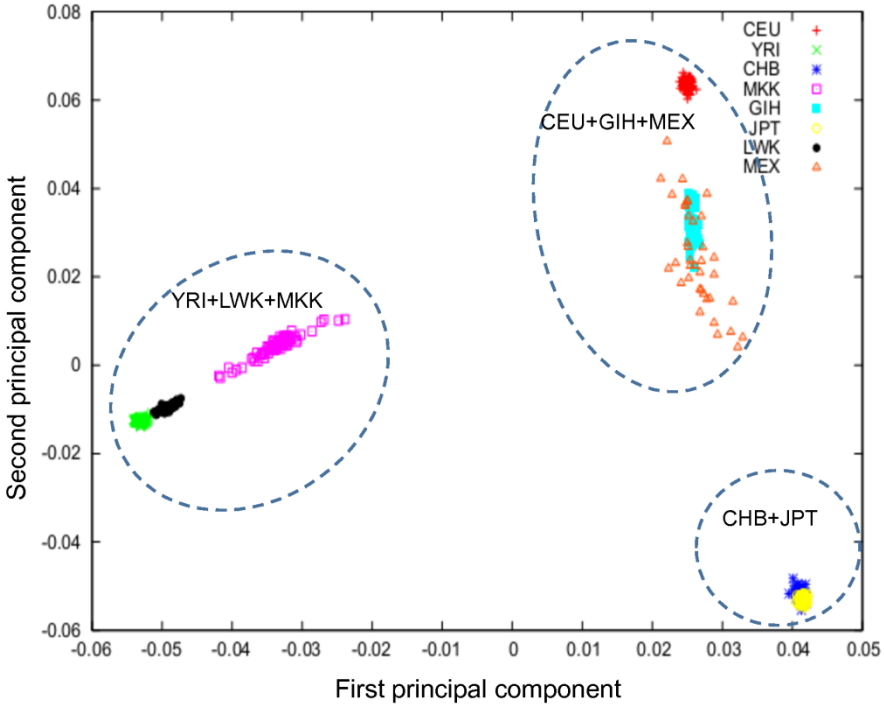


Figure S2: First principal component (PC1) versus second principal component (PC2) for all 726 individuals from 8 populations. The PC1 and PC2 accounted for 72.9% and 26.7% of total variation, respectively.

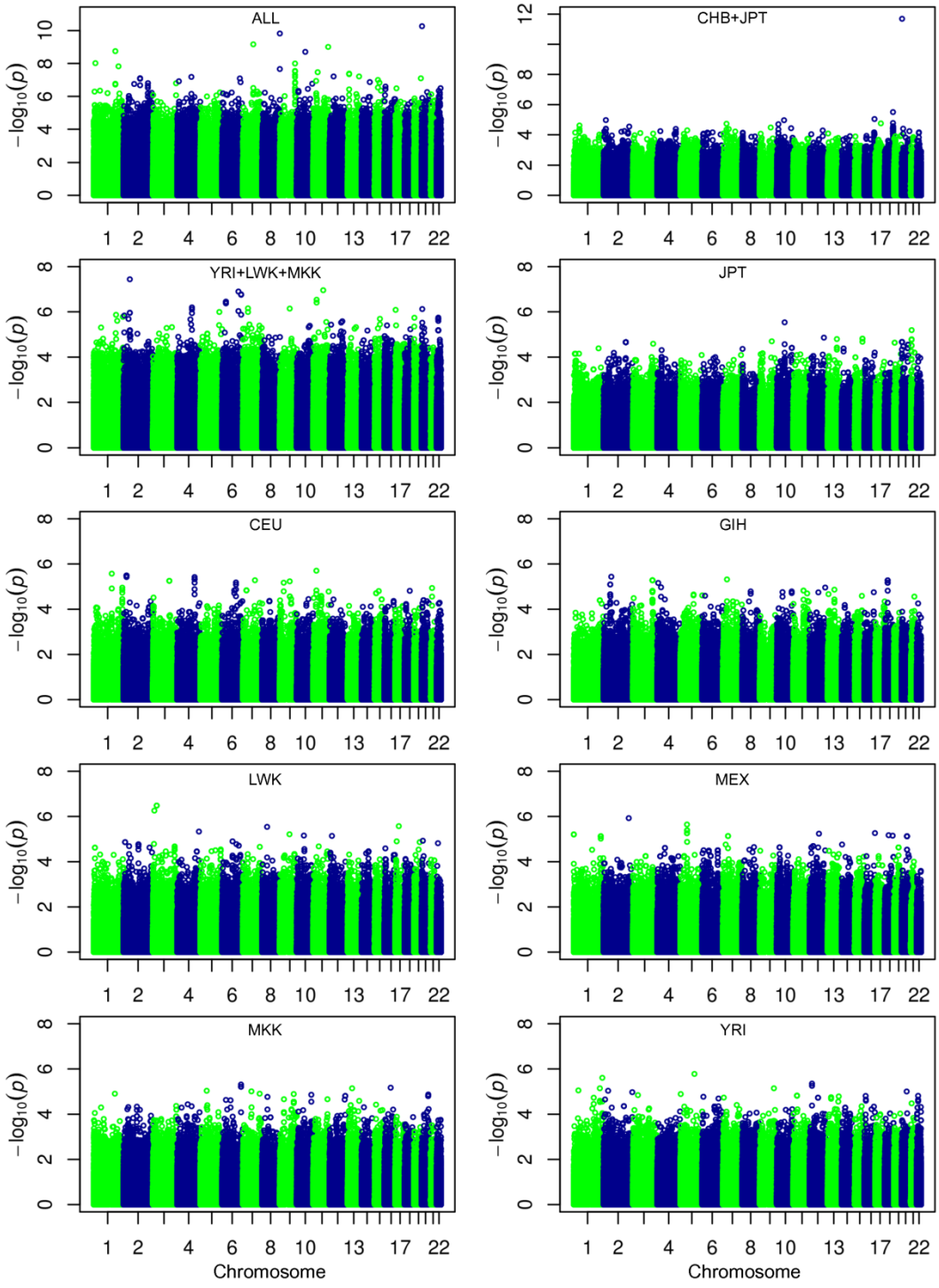


Figure S3: Manhattan plots for GWA with RS-score in different populations and combined populations. Each Manhattan plot shows the distribution of -$\log_{10}$ of the P-values from tests of association between individual SNP markers and the RS-score.

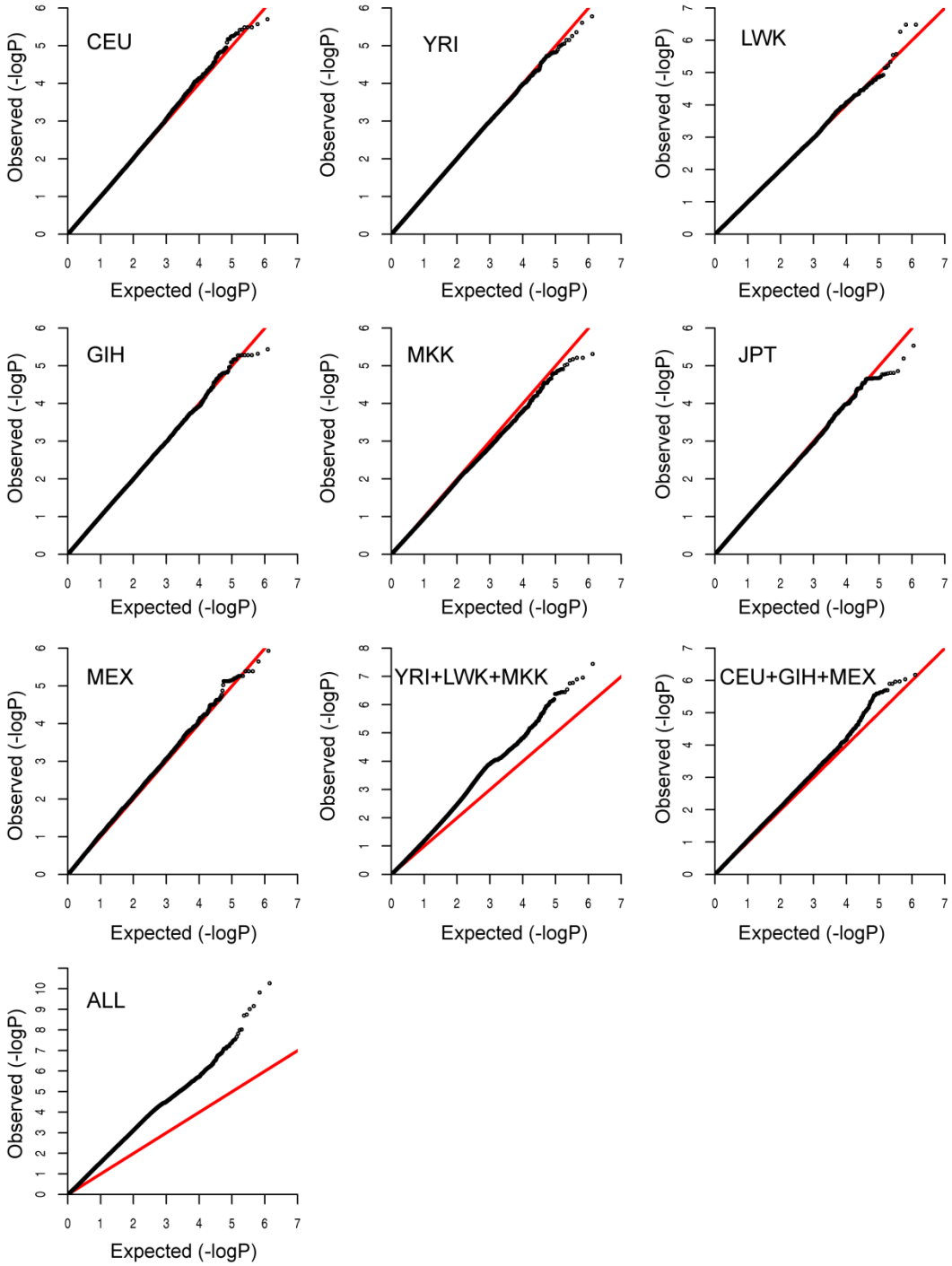


Figure S4: P-P plots of the association with RS-score. The expected (X-axis) shows $-\log_{10}$ of random values, drawn from the uniform distribution. The observed (Y-axis) shows $-\log_{10}$ of the P-values from tests of association between individual SNP markers and the RS-score. The red line is used to compare the expected and observed values.

Table S1: Summary of samples in the eight Hapmap3 populations

Population	**Population detail**	**Number of samples**
CEU	Caucasians living in Utah USA, of northern and western European ancestry	109
CHB	Han Chinese from Beijing, China	80
GIH	Gujarati Indians in Houston, TX, USA	82
JPT	Japanese in Tokyo, Japan	82
LWK	Luhya in Webuye, Kenya	82
MEX	Mexican ancestry in Los Angeles, CA, USA	45
MKK	Maasai in Kinyawa, Kenya	138
YRI	Yoruba in Ibadan, Nigeria	108

Table S2. Spearman correlation between *HNRNPA2B1* and the RS-score in each population

Population	**Rho**	**P-value**
YRI	0.38	5.0×10^{-05}
CHB	0.48	8.4×10^{-06}
MKK	0.18	3.3×10^{-02}
GIH	0.30	6.1×10^{-03}
JPT	0.28	1.2×10^{-02}
LWK	0.30	5.7×10^{-03}
MEX	0.33	2.7×10^{-02}
CEU	0.06	5.5×10^{-01}

Table S3. Association between *cis*-eQTL of *HNRNPA2B1* and the RS-score

Population	**SNP**	**P1**	**P2**	**P2_BH**
CHB	rs17153827	4.5×10^{-05}	2.2×10^{-03}	1.6×10^{-02}
CHB	rs17154015	6.5×10^{-04}	5.0×10^{-03}	2.0×10^{-02}
LWK	rs10242687	1.9×10^{-04}	2.7×10^{-03}	1.6×10^{-02}
LWK	rs1125542	3.6×10^{-05}	8.4×10^{-03}	2.5×10^{-02}

The second column shows four *cis*-eQTLs for *HNRNPA2B1* that are significantly associated with RS-score. *cis*-eQTL mapping for *HNRNPA2B1* was carried out, following a pipeline from Stranger *et al.* [263], by performing Spearman correlation tests between SNPs located within 500 Kb of the transcription start site of *HNRNPA2B1* and its expression level. Only SNPs corresponding to tests having P-values ≤ 0.001 are considered as *cis*-eQTLs, and these P-values are shown in the third column (P1). The fourth column (P2) contains the association P-values (only those values ≤ 0.01 are shown) between these *cis*-eQTLs and the RS-score. The final column (P2_BH) is the P-values of P2, corrected for multiple testing using the Benjamini and Hochberg procedure.

Table S4. Genomic inflation factors (lambda) in different populations.

Population	**Lambda**
YRI	1.005
CHB	1.017
MKK	1.000
GIH	1.024
JPT	1.014
LWK	1.000
MEX	1.060
CEU	1.011
CHB+JPT	1.020
CEU+GIH+MEX	1.086
YRI+LWK+MKK	1.262
ALL	1.900

Table S5. Top GO terms for genes positively correlated with rs6137010 in CHB

Term	**P-value**	**Bonferroni**
membrane-enclosed lumen	8.3×10^{-39}	5.9×10^{-36}
intracellular organelle lumen	1.0×10^{-38}	7.3×10^{-36}
organelle lumen	4.2×10^{-36}	3.0×10^{-33}
Mitochondrion	2.3×10^{-35}	1.6×10^{-32}
nuclear lumen	7.2×10^{-30}	5.1×10^{-27}
mitotic cell cycle	9.9×10^{-28}	4.3×10^{-24}
cell cycle	2.8×10^{-26}	1.2×10^{-22}
mitochondrial part	5.8×10^{-26}	4.1×10^{-23}
ribonucleoprotein complex	1.9×10^{-25}	1.3×10^{-22}
Nucleoplasm	2.2×10^{-25}	1.6×10^{-22}
cell cycle process	7.6×10^{-24}	3.3×10^{-20}
organelle envelope	2.8×10^{-23}	2.0×10^{-20}
Envelope	5.1×10^{-23}	3.6×10^{-20}
RNA processing	6.6×10^{-20}	2.8×10^{-16}