Chapter 2
An Overview of Gene Regulation

Andrew Harrison and Hugh Shanahan

Abstract It is not unreasonable to assume that in the near future next-generation sequencing techniques will allow the sequencing of all the DNA and expressed types of RNA involved in a given response or process. Such a range of data will be necessary to unravel the complexities of the multiple layers involved in the regulation of gene expression.

In this article we discuss a broad range of studies about gene regulation. These involve studies of processes such as transcription and splicing, the production of a variety of transcripts, and the involvement of protein–nucleic acid composites such as chromatin. We seek to shed light on common themes that are beginning to develop in these rapidly evolving, but intimately related, fields.

Keywords Gene expression • Post-transcriptional processing • Epigenetics • Non-coding RNA • Genome tertiary structure

Acronyms

CPSF	cleavage and polyadenylation specificity factor
CstF	cleavage stimulation factor
CTCF	CCCTC-binding factor
CTD	carboxy-terminal domain
dsDNA	double-stranded DNA
dsRNA	double-stranded RNA

A. Harrison (✉)
Department of Mathematical Sciences and School of Biological Sciences,
Essex University, Essex, UK
e-mail: harry@essex.ac.uk

H. Shanahan
Department of Computer Science, Royal Holloway, University of London, Surrey, UK

M. Chen and R. Hofestädt (eds.), *Approaches in Integrative Bioinformatics: Towards the Virtual Cell*, DOI 10.1007/978-3-642-41281-3_2,
© Springer-Verlag Berlin Heidelberg 2014

EJC	exon junction complex
NDR	nucleosome-depleted region
ncRNA	non-coding RNA
miRNA	microRNA
Poly(dA:dT)	double-stranded sequence of DNA composed of AT pairs
PAP	poly(A) polymerase
PASR	promoter-associated sRNA
PROMPTS	promoter-associated transcripts
PTM	posttranslational modification
RNAi	RNA interference
RNP	ribonucleoprotein
hnRNP	heterogeneous nuclear ribonucleoprotein
mRNP	messenger ribonucleoprotein
RNAPII	RNA polymerase II
ssDNA	single-stranded DNA
ssRNA	single-stranded RNA
sRNA	short RNA
siRNA	small interfering RNA
SR Protein	serine-rich protein
TSS	transcription start site
TSSa-RNA	transcription start-site-associated RNAs

2.1 Introduction

The development of sequencing technologies has resulted in dramatic reductions in sequencing costs over the last decade [1]. There are already a broad range of high-throughput sequencing technologies [2], with others, such as nanopore technology [3], expected to arrive in the very near future. Our increasing ability to sequence nucleic acids quickly and cheaply will transform many biological areas of research [4]. This includes medicine, and the sequencing, and resequencing, of individuals is already helping to illuminate the genetic changes responsible for cancer progression [5]. The new sequencing technologies are also being used increasingly in fields previously dominated by microarrays. Deep sequencing of RNA and recording its abundance in the sample, referred to as RNA-Seq [6–8], has generated much excitement and it has been claimed that it represents a revolutionary tool for transcriptomics [9]. We are still in the early days of the revolution and many of the RNA-Seq studies to date have been of a descriptive nature with basic data analysis [10]. However, there is a rapid growth in techniques and software to analyse next-generation RNA-Seq datasets [11] and increasingly sophisticated analyses are likely to become the norm. Even in the absence of sophisticated analysis techniques, there have been some fascinating results; for example, such experiments suggest that, for humans, approximately 75 % of the total mRNAs within a cell are common to all tissues, with about 8,000 protein-coding genes ubiquitously expressed [12].

However, transcriptome complexity is observed to vary between tissues, with areas such as the brain, kidney and testis expressing a greater diversity of mRNA than tissues such as the muscle and liver. Other techniques whether sequencing is being utilised include the measurement of protein–DNA interactions via ChIP-Seq [13].

Nonetheless, current next-generation sequencing presents challenges in assembly and sequence accuracy due to short read lengths and method-specific sequencing errors [14]. Understanding the physical causes impacting upon the fidelity of sequencing is important in establishing the error composition of any sequence. For example, a limitation of the 454 technology relates to sequences containing consecutive instances of the same base, such as AAA or GGG [2]. With this technology, the length of homopolymers is inferred from the signal intensity because there is no terminal molecule preventing multiple contiguous additions at a particular cycle. This results in a greater error rate than results from discriminating between incorporation and nonincorporation. The major error type for the 454 platform is insertion-deletion rather than substitution, whereas the dominant error for Illumina/Solexa is substitution, rather than insertions or deletions [2]. There are also other biases in RNA-Seq data which may limit its adoption for large-scale systems biology experiments [15]. For some applications, microarrays are more sensitive than the current sequencing technologies. This is leading to many groups using hybrids of sequencing and microarrays together, utilising the advantage of both approaches whilst minimising the disadvantages of each technology's limitations [14, 16].

The meta-analysis of large datasets of gene expression is now helping to underpin systems biology models, increasingly pointing to how the interactions between groups of closely coupled proteins underpin gene expression in humans and other higher eukaryotes [17, 18]. The implicit assumption of many current models in systems biology is that regulation is for the most part mediated by transcriptional regulatory networks [19]. However, this view has faced significant difficulties and blind studies to perform the high-throughput identification of transcription factor targets have provided very poor results [20]. Part of the problem is that the regulation of gene expression in eukaryotes is very complex and strongly modulated by a number of mechanisms beyond simple transcription factor complex formation. Our understanding of the components involved in gene regulation, their complexity as well as the interplay between different layers of regulation utilised within cells has expanded rapidly in parallel with our ability to utilise high-throughput sequencing.

Systems analysis of gene expression is identifying coordination and coupling in transcription, coordination among transcription factors, coupling among transcription factors and chromatin remodelers, a nuclear organisation coupled to transcription, interwoven layers of mRNA processing involving the coupling of transcription and splicing, coupling of transcription and export with quality control processes, dynamic messenger ribonucleoprotein complexes, regulation of cytoplasmic events from within the nucleus, coupling between transcription and ribosomal synthesis and links between protein synthesis and degradation [21]. Many of the molecular interactions responsible for coordination are being

mapped out biochemically [22], detailing the lines of feedforward and feedback between chromatin, RNA, multifunctional proteins and ribonucleoprotein (RNP) complexes.

These complexities are leading to designs of next-generation sequencing experiments increasingly requiring integrative approaches to bring together knowledge of the multiple layers of regulation of gene expression. There are many threads linking these layers and in this review we give a broad overview about the rapid progress in understanding the regulation of genes via several mechanisms. We will begin with transcriptional and post-transcriptional events. We will then discuss how the structure of chromatin radically affects transcription. Following this, the major role that non-coding RNA plays in regulation will be discussed. We will also highlight a number of common themes that are emerging across these different layers of regulation. Finally, we discuss how next-generation sequencing is poised to play a significant role in the systems biology of the future, the huge data management problem we face and how it will likely transform how we work together to better understand gene regulation.

2.2 Transcription and Beyond

The transcription cycle begins with preinitiation complex formation, RNA polymerase II (RNAPII) recruitment, a transition to an initiating and then an elongating RNAPII, and progressing to termination [23]. RNAPII will do work as it progresses through transcription and the amount of energy required to break and make bonds depends upon tertiary interactions between RNAPII, chromatin, nascent RNA and ribonucleoproteins (RNPs).

2.2.1 The Dynamic Nature of RNAPII

It is increasingly clear that subtle changes in the structure of RNAPII occur as it progresses through the transcription cycle. In particular, a relatively unstructured protein domain lies below the RNA exit channel [22], the carboxy-terminal domain (CTD) of RNAPII, and this serves as a binding pad for many nuclear factors, playing a key role during transcriptional and co-transcriptional processing, including terminating transcription.

The CTD has a simple heptad repeat structure, Tyr-Ser-Pro-Thr-Ser-Pro-Ser ($Y_1S_2P_3T_4S_5P_6S_7$), with 52 repeats in mammals [24]. The last repeat of the CTD in vertebrates is followed by a conserved ten amino acid extension. Thirty-one of the fifty-two repeats in the human CTD differ from the consensus heptad in at least one position, with most of the nonconsensus repeats towards the carboxy terminal of the CTD [24]. The presence of these divergent repeats enables additional functionality. As shown in Fig. 2.1, dynamic and reversible modifications to CTD

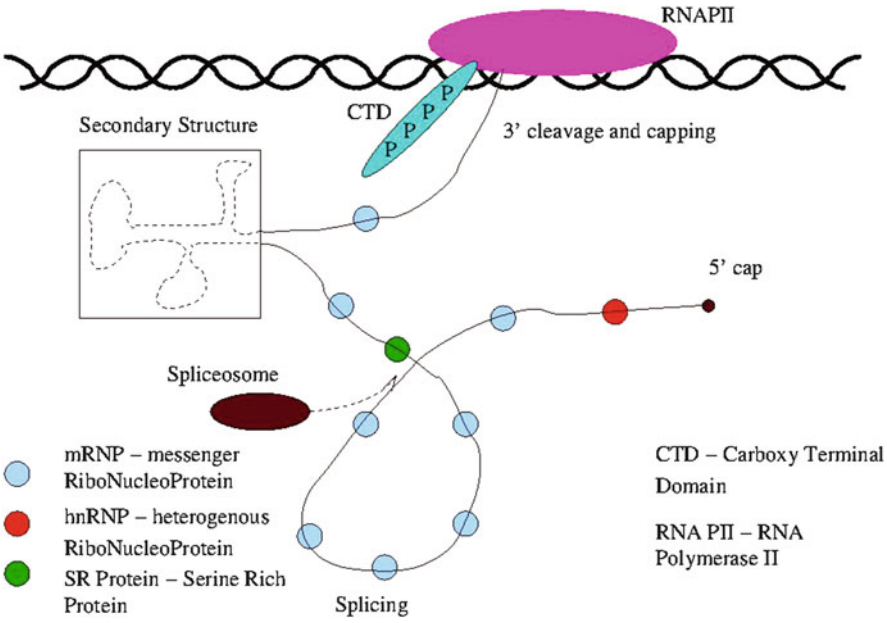

Fig. 2.1 A schematic diagram of the processes and interactions that occur in pre-mRNA and how splicing is implemented (described in Sects. 2.2.1, 2.2.2, and 2.2.3). The formation of bonds between the pre-mRNA and ssDNA is carried out by the formation of RNA secondary structure or binding with mRNPs. Splicing is enhanced by SR proteins or inhibited by hnRNPs. Initiation and termination of the transcript is aided by complex formation triggered by PTMs in the CTD region of RNA PII

occur during the transcription cycle, including phosphorylation, glycosylation as well as changes to the isomeric state of prolines. The appropriate recruitment of factors at different stages of the cycle is closely related to these modifications and a CTD code describing these coordinated changes is being actively sought [24].

All three serines of the CTD consensus repeat can undergo phosphorylation [24]. Ser2 and Ser5 are dynamically phosphorylated and dephosphorylated during the transcription cycle. Phosphorylation of Ser2 residues plays a major role in enabling RNAPII to progress into an elongating form, as well as being involved in splicing and polyadenylation events. Phosphorylation of Ser5 residues is greatest near the 5′ end of genes, with Ser5 phosphorylation helping in the addition of a methylguanosine cap to the 5′ end of the newly synthesised RNA. The CTD loses most of its Ser5 phosphorylation before RNAPII reaches the polyadenylation signals at the 3′ ends of protein-coding genes. The dephosphorylation of Ser2 and Ser5 during the transcription cycle is required for recycling RNAPII. Dynamic phosphorylation of Ser7 has a role in some protein-coding genes at their 3′ termini, involved in either terminating transcription or 3′ processing. Tyrosine and threonine can also undergo phosphorylation, but it is presently unclear what functions these play. Experiments to unravel the role of threonine are complicated by it being

found in 15 positions in the nonconsensus repeats, as well as in its canonical position 4 in the consensus repeats. Serines and threonines can also be glycosylated, with phosphorylation and glycosylation appearing to be mutually exclusive [24]. Isomerisation of the two peptide-prolyl bonds, at positions 3 and 6, also occurs, resulting in four possible configurations in each repeat.

2.2.2 *The Folding and Binding of Proteins to Nascent Pre-mRNA*

Alterations in RNA structures represent a regulatory mechanism for many cellular processes [25]. There is an intimate relationship between the binding of messenger ribonucleoproteins (mRNPs) and RNA secondary structure, with some proteins binding to single-stranded RNA (ssRNA) sequences [26] and others to double-stranded RNA (dsRNA) sequences [27]. Heterogeneous RNPs (hnRNPs) are very abundant in the cell and RNA–protein interactions act to modify the form of RNA secondary structures and may act to inhibit the existence of structures in some cases [28].

Pre-mRNA is free to fold only within a limited period after transcription, with an upper limit of ∼100 nucleotides [29]. It is likely that co-transcriptional wrapping up of RNA by folding, or through binding by mRNPs, occurs rapidly in order to minimise the possibility of genomic mutations induced by the formation of R-loops during transcription [30]. As shown in Fig. 2.1, an R-loop is a structure in which an RNA molecule is partially or completely hybridised with one strand of a double-stranded DNA, leaving the other strand unpaired [31]. Transcriptional R-loop formation in higher eukaryotes is highly correlated with chromosome instability. Little is known about the molecular mechanisms responsible for R-loops influencing genome stability, but single-stranded DNA is more vulnerable to mutations than double-stranded DNA [32]. Thus, extensive R-loop formation will result in these transcribed regions being more susceptible to DNA-damaging agents by increasing the frequency of single-stranded regions. R-loops will also act to slow down elongation of RNAPII [22].

2.2.3 *Post-transcriptional Splicing*

RNA-Seq results suggest that almost all human protein-coding genes undergo alternative splicing [33]. Furthermore, over 80 % of genes produce a minor isoform with a relative abundance of 15 % or more of the major isoform. It has been recently proposed that most alternative splicing is a consequence of noise in the splicing machinery [34]. However, alternative splicing and polyadenylation are observed to vary significantly between tissues, with coordinated changes in alternative

splicing and polyadenylation between many genes being observed, suggesting that alternative splicing provides a central contribution to the evolution of phenotypic complexity in mammals [33].

Pre-mRNA splicing occurs co-transcriptionally in all eukaryotes [22]. However, there is little overlap between groups of genes that are differentially spliced and those that are differentially expressed [35]. As shown in Fig. 2.1, a small number of RNA-binding proteins, usually members of the serine-rich protein (SR protein) and hnRNP families, are involved in splicing regulation and the interplay of these positive and negative factors acts to modulate the inclusion, or otherwise, of exons [36]. SR proteins help to activate splicing by binding to exons and recruiting the spliceosome. Most members of the SR protein family have their binding to RNA affected by the conformation of the target RNA [37]. SR proteins exert some of their stimulatory effect through stabilising RNA–RNA interactions during spliceosome assembly and splicing catalysis [38]. HnRNP proteins, in contrast, usually repress splicing by interfering with the spliceosome's interactions with splice sites. In particular hnRNP proteins may disrupt RNA–RNA interactions through sequestering sequences [38].

The binding of these positive [39] and negative [40] regulators of splicing has been shown to depend on RNA secondary structures. There seem to be two mechanisms involved in how RNA secondary structure affects the choice of 5′ and 3′ splice site and branch point elements. The most common process results from the presence of structural elements which may hinder the accessibility of selected sequences by splicing factors [37] – depending on the system analysed, this inhibition has been observed to target only the acceptor site, the donor site or both. The second mechanism occurs when RNA secondary structures that do not involve the conserved splicing sequences can vary the relative distance between these elements – these changes then result in considerable variation in splice site usage or efficiency [37]. Structural constraints also affect less-defined cis-acting sequences such as exonic/intronic splicing enhancers or silencer elements [41]. Furthermore, RNA secondary structure has been proposed to influence splicing. For example, secondary structural elements involving both exonic and intronic sequences have been found in the *dystrophin* gene [42].

The advent of high-throughput sequencing experiments, in conjunction with exon arrays, enables observations of co-regulated splicing events in groups of genes, as well as the determination of sequence motifs associated with these events [35, 43]. Some of the sequence motifs now being associated with tissue-specific alternative splicing are consistent with the binding patterns previously identified for known splicing regulators, such as NOVA and FOX [35]. This suggests that as catalogues of isoform expression profiles increase, they will provide sufficient sensitivity to enable the discovery of weaker motifs indicative of novel splicing regulators. There is a need for such analysis as there are more than 300 RNA-binding proteins in mammalian genomes that may act as splicing regulators, yet little is presently known about their binding specificity or their involvement in particular splicing events [35]. Mining of the existing datasets is already highlighting positional dependencies in the

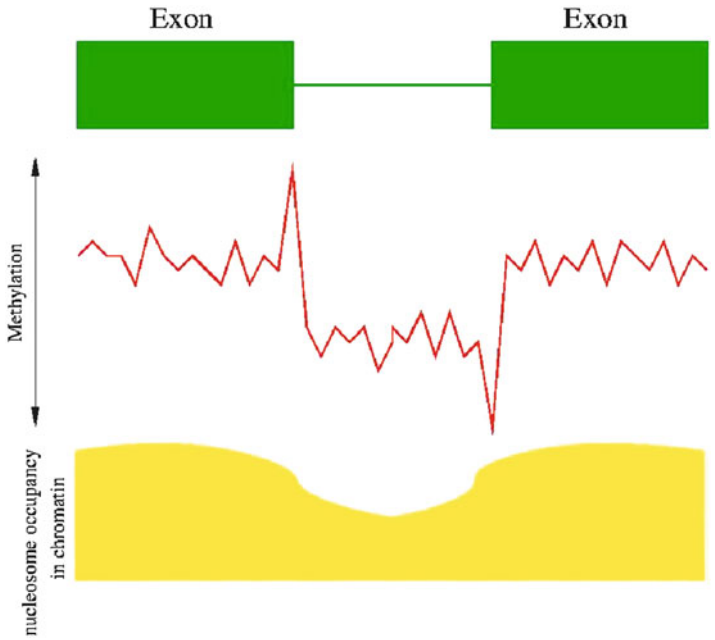

Fig. 2.2 A schematic diagram of the relationship between exon–intron boundaries, methylation and nucleosome occupancy as described in Sect. 2.2.3. As noted in Sect. 2.3.6, there is a noticeable peak in methylation (specifically CpG) at exon–intron boundaries and a trough at intron–exon boundaries [48]

binding of regulators, with both NOVA and FOX binding as enhancers when they are downstream of an alternative splicing exon, whereas they act as repressors when they bind on the upstream side [35]. Combinations of hundreds of RNA features are being assembled as part of large data-mining efforts to identify the principal components of the splicing code [44].

Unravelling evidence of co-regulated splicing events in several genes is non-trivial as it is very likely that splicing regulation can occur at every possible step of the spliceosome assembly and catalysis pathway. Furthermore, there are large numbers of factors involved in the splicing of each transcript and stochastic events may be important during splicing because simple binding kinetics determines the assembly pathways for a given pre-mRNA substrate [38]. Spliceosome assembly is also modulated in response to transcriptional events and chromatin structure [38]. The rate of elongation affects splice site selection and exon skipping and, thereby, the nature of the information expressed from a gene [45, 46]. Post-transcriptional processing also involves a close relationship with how DNA is modified in its accessibility during transcription [36, 47]. As shown in Fig. 2.2, chromatin organisation marks exon–intron structure and chromatin structure, via histone modifications, modulates exon selection [48, 49]. Transitions in DNA methylation across junctions of exons and introns may also be involved in splicing [50]. The differences in

transcription rates that result from these chromatin modifications, as well changes in nucleosome density [47], may be the principal cause for a large proportion of tissue-specific, or development-specific, alternative splicing events [36]. It is now possible to use high-throughput sequencing technologies to map histone methylation states across the human genome [51]. We are therefore likely to see high-throughput sequencing used in further integrative studies of the dynamic interplay between proteins modifying chromatin, interacting with RNA, and their resulting impact on alternative splicing. Efforts to crack the splicing code, e.g. [44], are likely to be enhanced by knowledge about the tissue-specific modifications that chromatin undergoes within particular genes of interest.

2.2.4 RNA Editing

RNA editing can provide a source of sequence variation between transcripts from the same gene. The most common form of editing in eukaryotes is A-to-I, in which adenosine is converted into inosine within double-stranded RNA and the inosine is subsequently treated as guanosine by the spliceosome and ribosome [52]. Such editing is apparent because of differences in the RNA sequence and the DNA sequence. A-to-I editing is essential for the maintenance of normal life in mammals [53]. Editing can undergo spatiotemporal regulation [52]. Furthermore, RNA editing and alternative splicing are coupled, as modifying the RNA sequence can result in novel splice sites [54]. Moreover, as shown in Fig. 2.3, multiple editing sites within the same transcript are weakly correlated and so results in the production of diverse transcriptomes, eclipsing the variety resulting from alternative splicing but with less impact on the protein composition within cells [53]. The diversity resulting from RNA editing may be a principal contributor to the adaptive evolution of phenotypic complexity in mammals and be a dominant source of transcript diversity in the brain [55]. However, editing has also been associated with a number of human pathologies [56]. In particular, alterations in RNA editing impact upon a number of psychiatric disorders [57], in particular upon an individual's responsiveness to serotonergic drugs. Polymorphisms in editing genes have also been recently associated with extreme old age in humans [58]. High-throughput sequencing has already been used to identify RNA-editing sites [59], and we are likely to see many further studies in this area.

Meta-analysis of sequence differences in the small RNA component of rice and Arabidopsis [60] indicates that sequences of many transcripts are likely modified in vivo. These include N1-methyl modified purine nucleotides in tRNA, potential deamination or base substitutions in microRNAs, 3′ microRNA uridine extensions and 5′ microRNA deletions. However, the impact of editing, and other post-transcriptional modifications, can mimic RNA-sequencing errors and a number of sequence variations previously classed as sequencing errors may in fact result from editing and other modification [60].

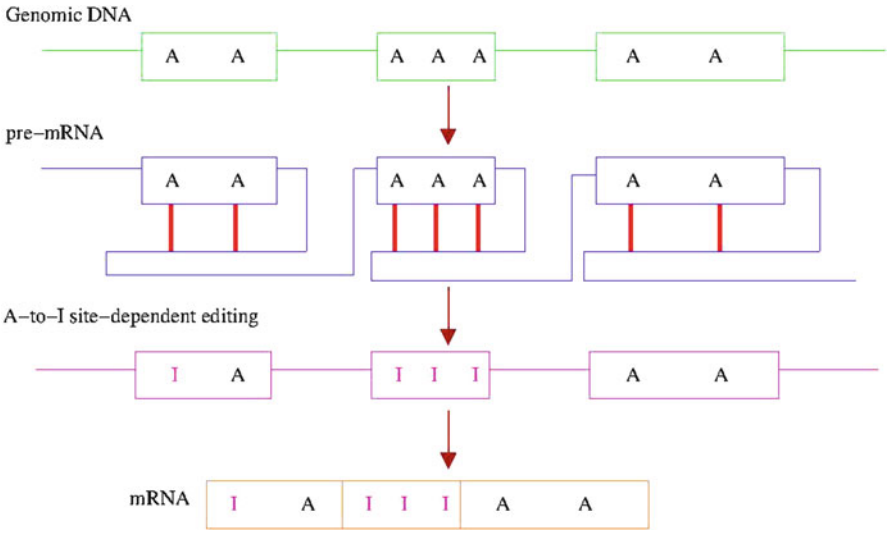

Fig. 2.3 A summary of RNA editing as described in Sect. 2.2.4. The actual editing occurs in double-stranded pre-mRNA which can then be edited by ADAR. We note that A-to-I editing is site dependent, i.e. not every A is edited to an I and the editing depends on the site and condition

2.2.5 The Processing of the 3′ Ends of Transcripts

There are a number of molecular mechanisms involved in processing the 3′ ends of pre-mRNAs in metazoans [61]. Transcripts are cleaved before acquiring a polyadenylation (poly(A)) tail and the efficiency and specificity of this 3′ processing is regulated by large protein complexes, involving many factors. Transcription factors and activators affect 3′ processing and there is also crosstalk between factors involved in transcription, splicing and this processing machinery. Furthermore, the CTD of RNAPII helps to couple this regulatory network through acting as a site for gathering and delivering polyadenylation factors [61].

The molecular machinery involved in 3′ processing has a complex architecture, containing over 80 proteins. As outlined in Fig. 2.4, there are several sub-complexes, including cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF) and poly(A) polymerase (PAP) [61]. The poly(A) signal consists of two sequence elements: an AAUAAA hexamer, or a variant such as AUUAAA, is found 10–30 nucleotides upstream of the cleavage site that binds CPSF; a U/GU-rich region is located approximately 30nt downstream of the cleavage site and associates with CstF. The majority of transcriptional units contain more than one poly(A) signal and the alternative choices act to change the coding sequence or the sequences of the 3′ untranslated region. This results in alternative protein isoforms or transcripts that differ in their stability, localisation, transport and translation properties [61]. Tissue-specific regulation of alternative polyadenylation has a higher frequency than other types of alternative splicing [33].

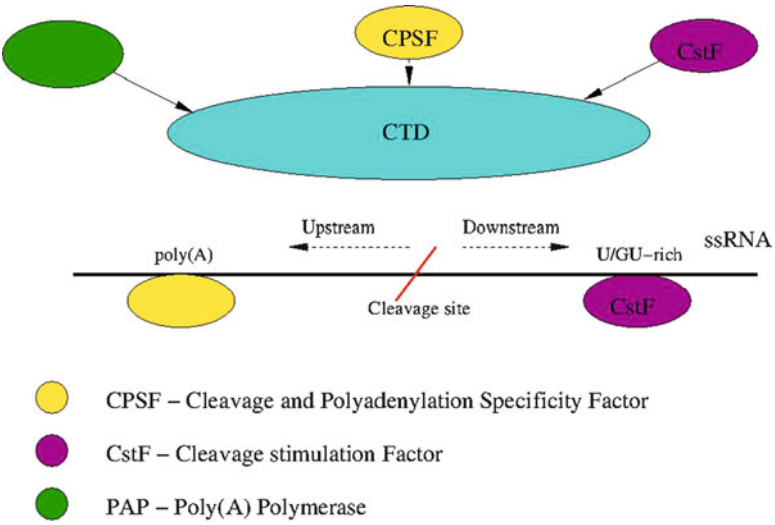

Fig. 2.4 A summary of the processing that occurs during cleavage of the 3′ end of a transcript as described in Sect. 2.2.5. The sub-complexes PAP, CPSF and CstF are not an exhaustive list of the sub-complexes required for 3′ processing. The CTD of RNA PII gathers and delivers polyadenylation factors. The cleavage site lies between an upstream poly(A) region (10–30 nucleotides of the cleavage site), which CPSF binds to, and a downstream, U/GU-rich region (∼30 nucleotides from the cleavage site) that CstF associates to

The interplay between several mechanisms involved in regulating 3′ end processing determines which of the transcriptional unit's sites are chosen to be polyadenylated [61]. Regulatory factors can compete with CPSF and CstF binding to their sequence elements. There can also be cooperative interactions, resulting from proteins bound to the transcript increasing the rate at which CPSF and CstF are able to bind their respective elements. Factors bound to the pre-mRNA can inactivate PAP. The rate of transcriptional elongation can shift the kinetic competition between processes, resulting in not enough time for upstream sites to be chosen and therefore the subsequent polyadenylation of downstream sites. Differential expression of individual proteins which make up part of the large 3′ processing complexes will act to preferentially select suboptimal cis-elements. Factors involved in polyadenylation can also be sequestered to the cytoplasm. The factors can also become bound into other complexes in the nucleus, which can result in different choices of site. The factors can also be posttranslationally modified, again altering which of the several sites are chosen to be polyadenylated.

Chromatin structure also impacts upon the regulation of alternative polyadenylation [62]. The canonical polyadenylation signal 6-mer, AATAAA, is a poly(dA:dT) tract, and such tracts act to stiffen DNA and deplete nucleosomes. Indeed, [62] find that human polyadenylation sites (PAS) have strong nucleosome depletion in conjunction with downstream nucleosome enrichment. Moreover, the downstream nucleosome affinity is associated with increased usage of the PAS when there are multiple sites available.

2.2.6 *Post-transcriptional Modifications and Folds Used in Quality Control and Regulation*

A number of post-transcriptional modifications are used by the cell to check the fidelity of transcripts as they are produced. The addition of a cap to the 5′ of the nascent transcript is likely a switch that enables RNAPII to move from an abortive state into a fully elongating state [22]. The poly(A) tail acts to enable transport of mRNAs from the nucleus to the cytoplasm and affects both their stability and the rate at which they are translated [61].

A key quality control process is the nonsense-mediated mRNA decay (NMD) pathway [22]. This involves the exon junction complex (EJC), a group of proteins which are deposited on spliced transcripts about 20 nucleotides upstream of exon–exon junctions [22]. During the first round of translation for a newly synthesised transcript, the presence of at least one EJC which is 50 or more nucleotides downstream of a stop codon results in the transcript and recently translated peptide being rapidly degraded. This targets those transcripts in which the first in-frame stop codon is poorly placed for transcript termination, resulting in the constitutive stop codon being either in the last exon or within 50 nucleotides of the final exon–exon junction [22].

EJC deposition possibly evolved to enhance protein production and mRNA surveillance [22]. However, NMD is used to play several regulatory roles in the cell, other than just simply removing aberrant transcripts. For example, a number of splicing factors appear to alter the production of their own isoforms in order to target their transcripts to the NMD pathway whenever their intracellular concentrations become too high. Moreover, splicing makes for better translation resulting from the interactions between EJCs and complexes associated with ribosomes [22]. Furthermore, the EJC also interacts with proteins involved in directing mRNA localisation.

It is not just the EJC that acts to modulate the efficiency of an mRNA's localisation and translation efficiency. A number of features of the untranslated regions of mRNAs control their metabolism [63], the regulation of which is likely to depend on the tertiary structure of RNA as well as trans-acting factors. For example, *cis*-acting elements in the mRNA, usually in the 3′ untranslated regions (UTR), mediate the subcellular region to which the transcript is localised. Whereas, other elements in the 3′ UTR, such as AU-rich elements, regulate mRNA decay. Translation efficiency also depends on structures in the 5′ UTRs as well as the length of the 5′ UTR.

2.2.7 *Sequence Variations*

Genome resequencing of individuals will identify the differences with other genome sequences, and identify single-nucleotide variations, and whether the individual is homozygous or heterozygous for such variations. Individual alleles may contain

distinctive sequences and heterozygous individuals may produce expression of different RNA sequences. RNA-Seq has now been used to detect single-nucleotide variations in expressed exons of the human genome [64].

2.3 The Structure of Chromatin Impacts upon Gene Regulation

Transcription and post-transcriptional processing occurs whilst RNAPII is progressing through chromatin. Rather than just being a naked strand of DNA, instead chromatin is a complex mixture of nucleic acid, proteins and covalently bound modifications.

2.3.1 Nucleosomes

Constraints on DNA arise from its interactions with group of eight basic histone proteins, collectively known as nucleosomes [65]. DNA and nucleosomes are arranged as beads on a string, with a linker of naked DNA sequence bridging two neighbouring DNA-wrapped nucleosomes. The nucleosomes act to neutralise the self-repulsion of DNA resulting from the negatively charged phosphates in its backbone, enabling DNA to be packaged efficiently and fit into the confined space of the nucleus. As shown in Fig. 2.5, the histone core is composed of two copies of four histone proteins (H2A, H2B, H3 and H4). Each octamer consists of two H3–H4 histone dimers bridged together as a stable tetramer that is flanked by two separate H2A–H2B dimers [66]. DNA coils through a left-hand toroid around the histone core, with approximately 147 bases looping 1.65 times around each nucleosome, with each histone core anchoring 34–36 DNA base pairs through electrostatic, hydrogen and nonpolar interactions [66]. A further linker histone, H1, protects internucleosomal linker DNA near the nucleosome entry-exit point [66]. DNA and nucleosomes may undergo further compacting into transcriptionally inactive 30 nm fibres [65], as well as other high-order compactions.

A short basic stretch flanking lysine around position 16 of the histone H4 N-terminal domain directs internucleosomal contacts, which modifies high-order chromatin structures [66]. The interaction between residues 16 and 20 of histone H4 and two acidic patches on the C-terminal α-helices of histone H2A present on an adjacent nucleosome mediates in salt-dependent folding of chromatin. Acetylation of lysine residues relieves positive charges, perturbing histone-DNA contacts and affecting nucleosome stability [66]. Indeed, acetylation of lysine 16 on H4 (H4K16ac) prevents the compaction of nucleosome arrays in vitro, likely via electrostatic repulsion and hindering H2A contacts [66]. The acetylation of H4K16 also repels ATP-dependent chromatin-remodelling complexes, such as ACF, which will only interact with histones in the absence of H4K16ac [66].

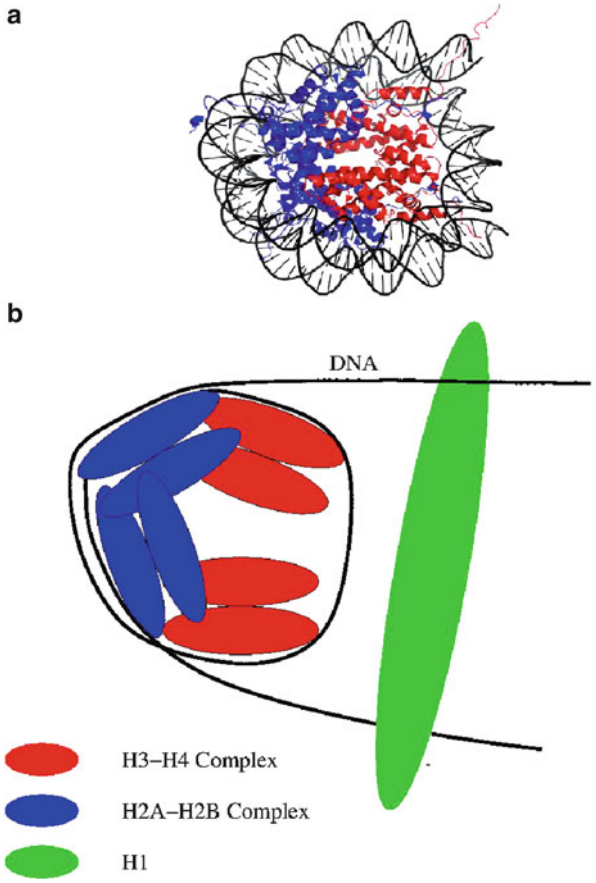

Fig. 2.5 The nucleosome as described in Sect. 2.3.1. In (**a**) a cartoon representation of a nucleosome structure determined from x-ray crystallography is shown (PDB code 1aoi) [223]. The histone structure H1 was not determined in the structure. In (**b**) a schematic diagram to represent the entire nucleosome, including the histone H1 structure, is shown. The core nucleosome structure is composed of eight domains which are composed of four dimers of H2, H3 and H4 histones. The DNA loops around this structure following this order of dimers: H2A–H2B, H3–H4, H3–H4 and H2A–H2B. The H1 histone binds to the entry and exit DNA giving the structure stability. The DNA turns 1.65 times and is comprised of 147 bases

2.3.2 Nucleosome Variants

Variants of histone combinations contribute to the properties of the nucleosomal core particle and its role in building specialised structures as well as altering transcriptional activity [66]. Histones H4 and H2B are largely invariant, whereas there is more variety with H3 and H2A.

The non-canonical H2A.Z is conserved from lower to higher eukaryotes. Nucleosomes can only incorporate one type of H2A variant because of steric clashes between loops in H2A and H2A.Z. H2A.Z impacts upon nucleosome stability and chromatin folding, resulting from a small destabilisation within H2A.Z-H3 interactions and a longer H2A.Z acidic patch, relative to H2A, used in H4 NTD binding. Despite its conservation, there remains uncertainty about the function of H2A.Z resulting from the rapid turnover rates of H2A.Z-containing nucleosomes [66]. Another H2A variant unique to mammals is H2A.Bbd. H2A.Bbd–H2B dimers dock on the $(H3–H4)_2$ tetramer, producing nucleosome core particles that organise about 118 base pairs of DNA but which are considerably less stable than the canonical nucleosomes. The variant H2A.Bbd lacks the ubiquitinatable C-terminal domain as well as the acidic patch that contacts the H4 N-terminal domain, making nucleosomes containing H2A.Bbd resistant to salt-induced chromatin folding. H2A.Bbd may reside within active chromatin [66].

Mammals have evolved a replication-dependent H3 variant, H3.1, that only differs from the non-canonical variant, H3.2, by the substitution of a single amino acid [66]. H3.2-containing nucleosomes are probably associated with heterochromatin. Whereas, the H3.3 histone variant differs from H3.1 by five amino acids and is associated with euchromatin. H3.3-containing nucleosomes are unstable, with the H3.3 histone undergoing rapid turnover. The displacement of nucleosomes during transcription appears to be the primary role for H3.3 [66]. Cysteines that are found in H3 variants may act to stabilise H3–H4 tetramers through disulphide bridges, particularly under oxidative conditions [67]. A further cysteine in H3.1 variants may also result in stabilising disulphide bridges between neighbouring nucleosomes which both contain H3.1s, helping to compact higher-order structures of chromatin [67]. There is also an H3.CenH3 variant that is involved in chromatin structures associated with kinetochore assembly and function [66]. The different forms of chromatin resulting from H3 variants and posttranslational modifications may result in chromosomes having a “barcode structure” [67], influencing epigenetic states during cellular differentiation and development.

The linker histone H1 acts to seal the two turns of nucleosomal DNA and is required for changes in conformation between extended and compact chromatin [68]. H1 also plays a role in establishing the spacing between nucleosomes, maintaining the level of methylation in particular regions of the genome, regulating a subset of cellular genes and acting to control development [68]. There are 11 variants of the linker histone H1 which is more than twice greater than the variability of any core histone. H1 variants also show a greater degree of divergence from each other than do the variants for other histones [68]. The variants are distinctive about when they appear in the cell cycle, but there is presently considerable uncertainty in their functionality [68].

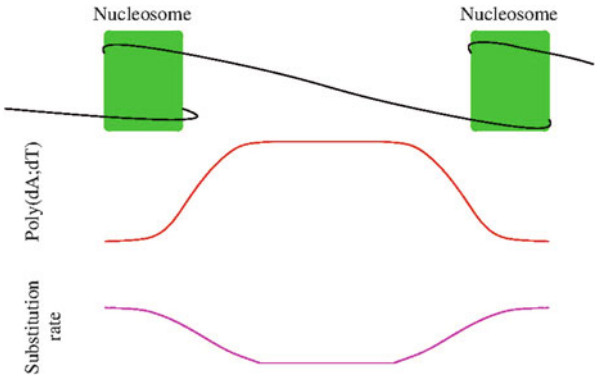

Fig. 2.6 A schematic diagram of the relationship between poly(dA:dT) density, substitution rates and its position between nucleosomes as described in Sect. 2.3.3

2.3.3 Nucleosome Positioning, Promoters and Gene Regulation

The code through which the genome sequence acts to position nucleosomes is increasingly understood [69–71] and our knowledge of the role nucleosomes play in gene regulation is rapidly evolving. DNA sequences have different abilities to bend and modify their helical twist and these differences are amplified when wrapping around the sharp bends of the nucleosome [70]. The bending around nucleosomes is facilitated through approximately 10 bp periodicity of specific dinucleotides. However, tracts of poly(dA:dT) are rigid and predicted to be unable to efficiently loop around histones. As shown in Fig. 2.6, such tracts are, as expected, observed to be free of nucleosomes [72] and play an important role in regulating nucleosome positioning within neighbouring genomic sequences. The nucleosome positioning code works in tandem with other regulatory codes in DNA [73], and amino acid content of proteins are likely to be modified as a function of nucleosome occupancy [74]. Moreover, [75] (see Fig. 2.6) have observed that substitution rates in linker regions are approximately 10–15 % lower than in nucleosomal DNA, which may be associated with higher DNA repair efficiencies in linker regions compared to nucleosomes. The roles that nucleosomes play in regulating transcriptional start sites, discussed below, in conjunction with differences in rates of insertions and deletions, and point mutations, between DNA wrapped around nucleosomes and that in linker regions, act to leave a nucleosome-associated periodic pattern in genome sequences, ultimately moulding the DNA sequence over evolutionary time scales [76].

The movement of nucleosomes by a few bases along DNA can dramatically alter the accessibility of the genomic sequence. Variations in genome sequences subsequently impact on nucleosome affinities and promoter structure, resulting in distinct modes of gene regulation [72, 77]. Functional promoters in eukaryotes must attract RNAPII and also evade the effects of nucleosomal repression. Cryptic transcription may occur if the suppression induced by nucleosomes does not function [78]. Typically, transcription start sites are found in nucleosome-free regions [47] as a major mechanism for suppressing transcription is to wrap potential transcription start sites around nucleosomes [79].

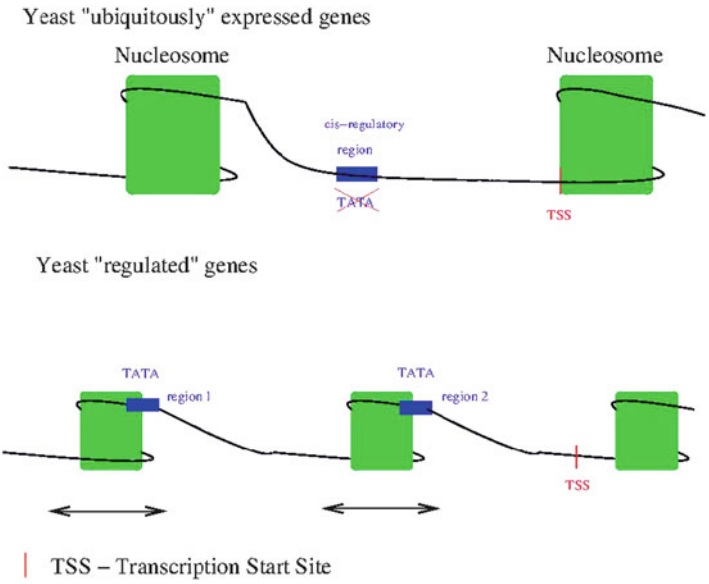

Fig. 2.7 A schematic diagram relating the positioning of cis-regulatory regions, the transcription start site (TSS) and nucleosomes for ubiquitously expressed and regulated yeast genes as described in Sect. 2.3.3. In the ubiquitous case, cis-regulatory regions tend to lie in the linker regions with the TSS at the start of a nucleosome. In addition such cis-regulatory regions do not have a TATA box. Regulated genes, on the other hand, have their cis-regulatory regions lie in the exposed regions of the nucleosomes and can be exposed or hidden more as individual nucleosomes slightly shift their position. These regions tend to have a TATA box

The nucleosome positioning signals are used by eukaryotes to regulate gene expression with distinct noise and activation kinetics through altering the architecture of promoters [72]. As outlined in Fig. 2.7, "ubiquitously" expressed genes in yeast have open promoters [80], characterised by a poly(dA:dT) tract resulting in a large nucleosome-depleted region (NDR) close to the transcription start site, in conjunction with accurately positioned nucleosomes further upstream. Associated cis-regulatory sequences reside within the NDR and the lack of nucleosomes means that transcription factors can bind to the regulatory DNA without competition. TATA-binding boxes are typically not found within these promoters.

As outlined in Fig. 2.7, "regulated" genes in yeast have covered promoters [80], with a more evenly distributed nucleosome positioning, resulting in transcription factors and nucleosomes competing for access to DNA. Transcription factor binding sites tend to be exposed on the nucleosome surface, near the border with a linker [81]. The nucleosome positioning sequences for these promoters result in high nucleosome occupancy close to the transcription start site. The regulation of chromatin, via subtle changes in nucleosome positioning and accessibility of DNA to transcription factors, enables large dynamic changes in expression [72, 77, 82]. TATA elements are frequently associated with this group of promoters [82].

Mammalian genes that have broad CpG-enriched promoters tend to produce multiple transcription start sites and are typically ubiquitously expressed [83]. The regulation of which start site is chosen is associated with the methylation state of the promoters [84]. Whereas mammalian genes with promoters containing a TATA box tend to produce a sharp single transcriptional start site and typically produce tissue-specific expression [83, 85]. In mammals promoters containing a TATA box evolve slower than promoters containing CpG islands [86]. Furthermore, the sequence of DNA at human promoters, enhancers and transcription factor binding sites, in contrast to yeast, typically encodes high intrinsic nucleosome occupancy [87], with these regions depleted in nucleosome-excluding poly(dA:dT) tracts.

The structure of DNA wrapped around nucleosomes details the tertiary structure of a gene within a sequence and structural variations at the chromatin level are likely to play a role in the regulation of the co-transcriptional processing of RNA. Long poly(dA:dT) tracts, which exclude nucleosomes, are avoided in exonic sequences, enabling an increased density of nucleosomes in exons [73]. Furthermore, differences in linker lengths between nucleosomes in exons and introns may result in different chromatin-packing arrangements [73]. The positioning of nucleosomes is also involved in exon definition events during co-transcriptional processing [49, 88] and nucleosome depletion has been associated with the regulation of polyadenylation [62]. Furthermore, [89] have identified peaks in the density of H2A.Z-containing nucleosomes just downstream of start codons and just upstream of stop codons in human T-cells.

2.3.4 Dynamic Nucleosomes and Gene Regulation

The regulation of the dynamics through which DNA alters its binding around nucleosomes is intimately involved in controlling gene expression [90] and the different mechanisms are outlined in Fig. 2.8. Models of such regulation are founded on the idea that the regulation results from a competition between nucleosomes and other DNA-binding proteins [91]. The affinities that these molecules have for the sequence (binding affinity landscape) dictate their competitive and cooperative interactions [91]. High nucleosome occupancy tends to reinforce cooperative interactions between transcription factors in displacing nucleosomes [87].

DNA accessibility and nucleosome positioning are also regulated through the action of ATP-dependent chromatin-remodelling complexes. Chromatin-remodelling complexes are present at many promoters [92] and the dynamic repositioning of nucleosomes has been associated with selecting the transcriptional start site [65] as well as other aspects of the initial stages of transcription [93]. Also, histone-devoid transcriptional start sites, in conjunction with the active cycling of factors on and off a promoter, permit formation of preinitiation complexes that are poised for transcription to be initiated [66, 90], a different state from a gene that is fully repressed. The evolution of chromatin-remodelling complexes is likely associated with changes in chromatin regulation during the evolution of vertebrates

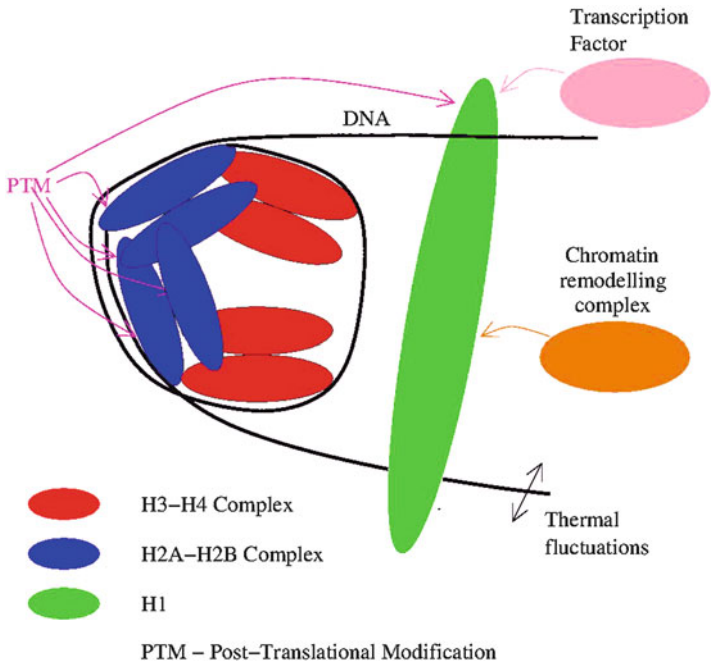

Fig. 2.8 A summary of the regulatory mechanisms that can be applied to the nucleosome as described in Sect. 2.3.4. PTMs can be applied to the H2 histones. The H1 histone can interact with transcription factors or chromatin-remodelling complexes. Furthermore, thermal fluctuations may result in the transient exposure of DNA regulatory sites to proteins

from unicellular eukaryotes [92]. Complexes, such as the ISWI (imitation switch; [94]) family, are involved in regulating higher-order chromatin structure [92], promoting regularly spaced nucleosomes and gene silencing [66]. Whereas the complex SWI/SNF (switching defective/sucrose non-fermentation; [94]) transiently exposes DNA regulatory sites through creating loops on a nucleosome's surface [65]. Some of these SWI/SNF complexes, such as BAF complexes in mammals, undergo progressive changes in subunit composition during the transition from a pluripotent stem cell to a multipotent progenitor cell [92]. At least four ATP-dependent remodelling complexes have nonredundant and specialised roles in maintaining pluripotent chromatin within stem cells. Tissue-specific complexes may enable matching between chromatin remodelling and transcription factors [92]. This can result in co-regulation of many genes or be restricted to the activation or repression of a single gene.

The disruption or displacement of nucleosomes will modify the rate at which polymerases pass over the DNA or the rate at which transcriptional factors will bind [65]. There are transcription-coupled changes in DNA topology or local chromatin structure, with histone and nucleosome removal during elongation of RNAPII [47]. The transit of RNAPII across the transcription unit is preceded by a

leading wave of histone posttranslational modifications that open the chromatin and transiently displace nucleosomes [95]. There are at least two processes by which this happens. The first results from the nucleosome within transcriptionally active genes having two components, a fluid H2A–H2B dimer and a stable H3–H4 tetramer – H3–H4 tetramers are ∼20 times more stable than H2A–H2B dimers [90]. The stability of the H2A–H2B dimer within the nucleosome will be further decreased by posttranslational modifications such as ubiquitylation, phosphorylation and acetylation. H2A–H2B dimers can also be exchanged through the actions of ATP-dependent chromatin-remodelling complexes. The movement of the H2A–H2B dimer could enable transcription factors and polymerases to access binding sites on DNA. The second process results from the linker histone H1 and its subtypes, associated with greater than 80 % of the nucleosomes in a mammalian nucleus, having residence times of a few minutes in interphase, consistent with dynamic interactions [90]. However, these residence times are variable and governed by the phase of the cell cycle, posttranslational modifications to H1, the subtype of H1 and competition for binding sites with other competing factors, such as transcription factors and chromatin-remodelling complexes, each of which themselves show dynamic interactions with chromatin [90]. Thus, alterations in residence times of H1 can result in changes to the residence time of a transcription factor, changing the balance between repression and activity. Thermal fluctuations of DNA wrapped around nucleosomes may also result in transient exposure of DNA regulatory sites to proteins. Such exposure is most energetically favourable towards the entrance and exit points of the DNA around the nucleosome and, indeed, transcription factor binding sites tend to be exposed near the border with a linker [81].

2.3.5 Histone Tails

Each histone has a tail which is targeted by a broad range of chemical moieties at multiple sites. Virtually all exposed polar residues (and some of the prolines) within the tails of histones are subject to covalent posttranslational modifications (PTMs). These include acetylation, methylation, phosphorylation, ubiquitylation, ADP-ribosylation, glycosylation [90] and SUMOylation [95], with lysine residues modified by up to three methyl groups. Acetylation and methylation results in only a small chemical group being added to the tail. However, ubiquitylation and SUMOylation are large appendages, almost the same size as the histone proteins, and their bulk could lead to more prominent changes in chromatin structure [95]. There is strong purifying selection among histone proteins and these targeted residues [66]. There has been considerable effort in establishing whether, and how, combinations of moieties on groups of histone tails act to produce a histone code that is used to regulate chromatin compaction and transcription [66].

A series of interlocking histone PTMs occurs during initiation, early elongation and mature elongation [95]. The transcriptional state of chromatin is correlated with several histone PTMs [66]. For example, hypoacetylation of H4K30me3 (trimethy-

lated lysine residue at position 30 of the H4 tail) and H3K27me3 is associated with silenced chromatin, whereas hyperacetylation of H3K4me3 and H3K36me3 is associated with actively transcribed chromatin. Moreover, the distribution of these marks can be distinctive, with H3K4me3 present at the beginning of genes whereas H3K36me3 accumulating within the body and towards the downstream region of genes [90]. However, single histone marks do not fully prescribe chromatin structure and its impact upon transcription and different marks works in combination when interacting with histone-binding proteins [66, 90]. Furthermore, experiments on transcriptionally synchronised genes are beginning to unravel a transcriptional clock controlled by dynamic nucleosomes [90]. Changes in the methylation and acetylation status of the histone pass through cycles, with particular combinations of histone modifications never coexisting on the same nucleosome at the same time. However, the sequence of events at a nucleosome appears to depend on many factors and there have been no simple rules describing the order of events [90]. In particular, different causes for why a gene is induced produce distinctive histone modifications [90]. Moreover, different sets of histone modifications act to regulate gene expression in high-CpG-content promoters and low-CpG-content promoters [23].

2.3.6 DNA Methylation

Cytosines within chromatin can be covalently modified so that they carry a methyl group at position 5 within their pyrimidine rings [96]. 5-Methylation of cytosine does not affect its base pairing with guanine, and cytosine is still replicated as cytosine. A consensus view has been that DNA methylation always appears in a CpG context (C followed by a phosphate and then a G, i.e. CG is on the same strand). Methylation of CpG has high mutagenic potential [96], as 5-methylcytosine can be deaminated to thymine. Such transitions accumulate over the course of evolution resulting in CpG dinucleotides being markedly unrepresented in genomes of vertebrates given the fraction of cytosines and guanines in the genome. However, there are islands of CpGs which are found at the expected frequency, and these tend to overlap with gene promoters [96].

Once a methyl is added to cytosine, it can be copied to newly synthesised DNA, resulting in an epigenetic memory that can be conserved during cell division. The DNA methylation pattern is maintained through mammalian development by DNMT1, a methyltransferase that is associated with the replication complex [97]. During cell division and DNA replication, DNMT1 is involved in recognising methylated CpG residues on hemimethylated DNA and methylates the opposite strand. However, epigenomic profiles also undergo targeted methylation and demethylation alterations during development, and these differential changes in methylation play a crucial role in cell lineage commitment [98]. For example, targeted repression and de novo methylation of genes responsible for pluripotency occur at gastrulation, whilst the embryo is beginning to separate into germ layers [97]. The importance of epigenetic alterations that affect tissue-specific

differentiation is such that their dysregulation could be the principal mechanism through which epigenetic changes cause cancer [84]. Different tissues show marked differences in DNA methylation [99], such that tissue-matched profiles from adult patients of different ages have more in common with each other than do disparate tissues from the same individual. Indeed, broad methylation patterns show tissue-specific conservation from humans to mouse [84], such that the methylation profiles of human and mouse brain cells, or human and mouse heart cells, have more in common than do the profile of a human brain and human heart cell.

Methylation acts to change the properties of chromatin. For example, methylation of DNA acts to modify nucleosome formation and positioning [96]. Biophysical and structural studies of DNA indicate that CpG methylation reduces backbone flexibility and dynamics, decreasing local DNA deformability. The position of 5-methyl group in the major groove increases steric hindrances on DNA wrapping around the nucleosome [96]. Only altering the conformation of a few nucleosomes through methylation may result in a significant impact upon the regular spacing arrays of nucleosomes expected to be involved in producing higher-order chromatin structure [96].

The methylation of cytosines affects how chromatin can subsequently bind to trans-acting factors and RNAPII. The binding of methylation to gene promoters can act to suppress transcription, and so any methylation associated with genes was believed to be indicative of transcriptional repression. However, this view is undergoing revision following the results from whole-genome epigenomic observations [99]. For example, a key function for differential methylation during differentiation is associated with changes in alternative transcription start sites [84]. Hypomethylation of promoters in conjunction with higher levels of gene-body methylation is positively correlated with transcription [50]. There is also recent evidence that DNA methylation acts to mark out aspects of gene structure within chromatin but shows cell-type-specific differences. DNA methylation peaks are found at the transcriptional start site in human T-cells [89]. Whereas DNA methylation shows a trough at the transcriptional start site in human embryonic stem cells and fibroblasts [50]. Both [50] and [89] find a drop in DNA methylation at the transcriptional termination site. Exons typically show higher CpG methylation fractions than do introns [50]. Interestingly, there is a sharp peak in CpG methylation at the exon–intron junction and a sharp dip in CpG methylation at the intron–exon junction [50], suggesting that transitions in DNA methylation are involved in splicing regulation. DNA methylation also peaks just downstream of the start codon and just upstream of the stop codon, suggesting that DNA methylation may be used as a signal for the addition, or removal, of co-transcriptional modifications that will only be utilised during translation at the ribosome [89]. Gene-body methylation may also inhibit incorrect choice of start sites for transcription [99].

The view that methylation is restricted to CpG sites is being questioned due to results from the first DNA methylomes that are now being sequenced at base resolution [100]. Almost 100 % of the methylcytosines in fully differentiated fibroblast cells are indeed in a CpG context, whereas pluripotent embryonic stem cells show almost 25 % of the methylcytosines in a non-CpG context (C followed

by a C, A or T; [100]). Moreover, 99 % of the methylation of CpG sites occurs on both strands (the opposite strand is also CpG), whereas methylation on CHG (where H = A, C or T) is highly asymmetrical, with 98 % of the cases being found only on one of the strands rather than both [100]. Moreover, within embryonic stem cells, non-CpG methylation is enriched within gene bodies but is depleted in DNA-protein-binding sites and enhancers [100].

At this time it is unclear whether gene-body methylation, and its marking out of gene structure, is restricted to subsets of genes in particular cell types. The biological implications of such methylation as well as methylation's interplay with transcriptional elongation and splicing are still uncertain. Indeed, the initial findings from whole-genome methylation profiles, from a small number of cell types, indicate that we are still someway from understanding the biology of gene-body DNA methylation. But the rate of discovery suggests that the next few years will lead to a significant illumination of the role of DNA methylation in gene regulation across the genome.

2.3.7 DNA Methylation Interactions with Histone Tails

There are regulatory interactions between enzymes involved in processing DNA methylation and histone modifications [97] and these interactions play a crucial role in mammalian development [101]. For example, G9a contains an SET domain which acts as a histone methyltransferase, and G9a also contains an ankyrin domain which recruits the DNA methyltransferases DNMT3A and DNMT3B. DNA methylation patterns are erased in the early embryo, resulting from passive demethylation caused by DNA (cytosine-5)-methyltransferase 1 (DNMT1) being excluded from the nucleus [98]. Methylation profiles across the genome are then re-established in each cell at approximately the time of implantation, through a wave of de novo methylation whilst ensuring the CpG islands remain unmethylated [97]. As shown in Fig. 2.9, the de novo DNA methylation template is written through histone modifications, with patterns of methylation of H3K4 across the genome being formed in the embryo before de novo DNA methylation. CpG islands in the early embryo have RNAPII bound to them and this acts to recruit H3K4 methyltransferases. Whereas the rest of the genome contains nucleosomes with unmethylated H3K4. Subsequently, de novo methylation occurs through the action of DNMT3A and DNMT3L (DNMT3-like, a paralogue of DNMT3A) complexed with DNMT3B. This recruits methyltransferases to DNA by binding to histone H3, whereas any form of methylation of H3K4 acts to inhibit this methylation. This results in de novo DNA methylation taking place at CpG sites throughout the genome but being prevented at CpG islands because of the presence of H3K4me. This model explains the strong anti-correlation between DNA methylation and H3K4me in a number of cell types [97]. Moreover, a DNMT3A–DNMT3L tetramer may oligomerise on DNA-containing histones without H3K4me and lead to the nearly global methylation of the mammalian genome [101].

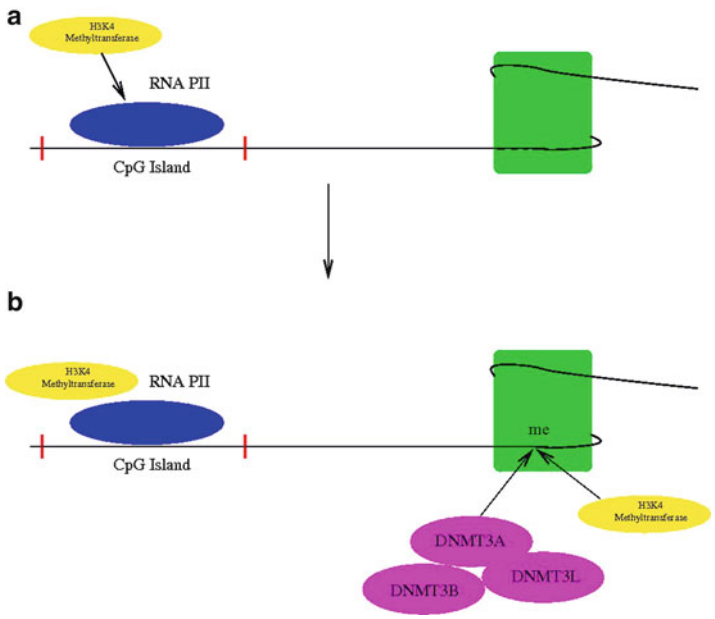

Fig. 2.9 A schematic diagram explaining the mechanism how non-CpG island sites are methylated in embryos as described in Sect. 2.3.7. In (**a**) CpG island sites are first bound by RNA PII sites which then recruit the H3K4 methyltransferase, sites that will be methylated are bound to nucleosomes. In (**b**) H3K4 methyltransferase in conjunction with DNMT3B, DNMT3A and DNMT3L to methylate the relevant CpG site. The CpG island sites are protected from methylation by the previously bound H3K4 methyltransferase

During post-implantation development, further epigenetic reprogramming occurs in primordial germ cells [101]. DNA methylation patterns are re-established by DNMT3A and the DNMT3B–DNMT3L complex at imprinted loci and transposable elements during gametogenesis. Targeting to transposable elements may involve Piwi-interacting RNAs, whilst targeting to imprinted genes involves the interactions of DNMT3L with unmethylated H3K4 tails [101].

DNA methylation also helps to maintain patterns of histone modifications through cell division [97]. During replication and cell division, regions that are methylated tend to be reassembled in a closed conformation, containing histones that are non-acetylated. Whereas unmethylated DNA gets repackaged in a conformation that is more open, containing nucleosomes whose histone tails are acetylated [97]. The mediation between DNA methylation and histone modifications likely results from methylcytosine-binding proteins such as MECP2 and MBD2, which are able to recruit histone deacetylases to methylated DNA. Enzymes such as G9a and DNMT1 may also interact with DNA methylation sites and direct H3K9 dimethylation [97].

2.4 The Spatial Organisation of the Genome in the Nucleus Acts to Regulate the Expression of Genes

2.4.1 Gene Expression Is Localised

It is increasingly clear that genomes and gene regulation are organised non-randomly in the nucleus [102]. Most nuclear events occur in spatially defined sites and in dedicated nuclear bodies, rather than occurring ubiquitously throughout the nucleus [103]. The formation of structures in the nucleus, such as Cajal bodies [104], results from stochastic assembly and self-organisation. Similarly, biological processes such as the formation of the DNA damage response may result from self-organising events [105]. Indeed, whole genomes may be considered as self-organising entities during mitosis, with networks of co-regulated gene expression and chromosomal association that are mutually related during differentiation resulting in self-organising lineage-specific chromosomal topologies [106]. The density of RNAPII may also act to regulate the colocalisation of gene expression [107].

Heterochromatin regions of the genome are usually found at the periphery of the nuclear membrane and are usually silent, whereas more open chromatin associated with active genes is typically found towards the centre. This is outlined in Fig. 2.10. Such a situation is consistent with biophysical models of the entropic organisation of self-avoiding polymers which suggest that long flexible polymers (associated with gene-rich chromosomes) will move to the centre of a confining sphere, whereas compact polymers (heterochromatin) will move to the periphery [108]. There is increasing evidence that gene activation or silencing is frequently associated with repositioning of the locus relative to nuclear compartments [109]. Active genes dynamically colocalise to shared sites of ongoing transcription [110] and genes such as Myc have been observed to preferentially relocate to regions in the nucleus at the same time as other genes with which they are co-regulated [111]. The movement of DNA into loops can result in proximal associations between co-regulated genes which are separated along the genome sequence [112]. The initiation step of transcription is required to tether genes to the same foci [113], but even in the absence of transcription, there are still localised concentrations of RNAPII [113]. A model consistent with much of the experimental data is that there are transcription zones within the nucleus in which RNAPII is concentrated locally through self-assembly processes [114]. These dense regions of nuclear RNAPII concentration have been termed factories, and it is possible that a transcriptional factory model may describe an aspect of the architecture of all genomes [115].

The synthesis of mRNA in mammalian cells is observed to be stochastic [116], with developmental genes exhibiting pulses of activity [117]. The stochastic nature of gene expression results from dynamic passage of genes through transcription factories [112]. Selection pressures will act upon groups of genes undergoing coordinated stochastic transcriptional regulation, and chromosome organisation shows the signature of selection for reduced gene expression noise [118]. Other

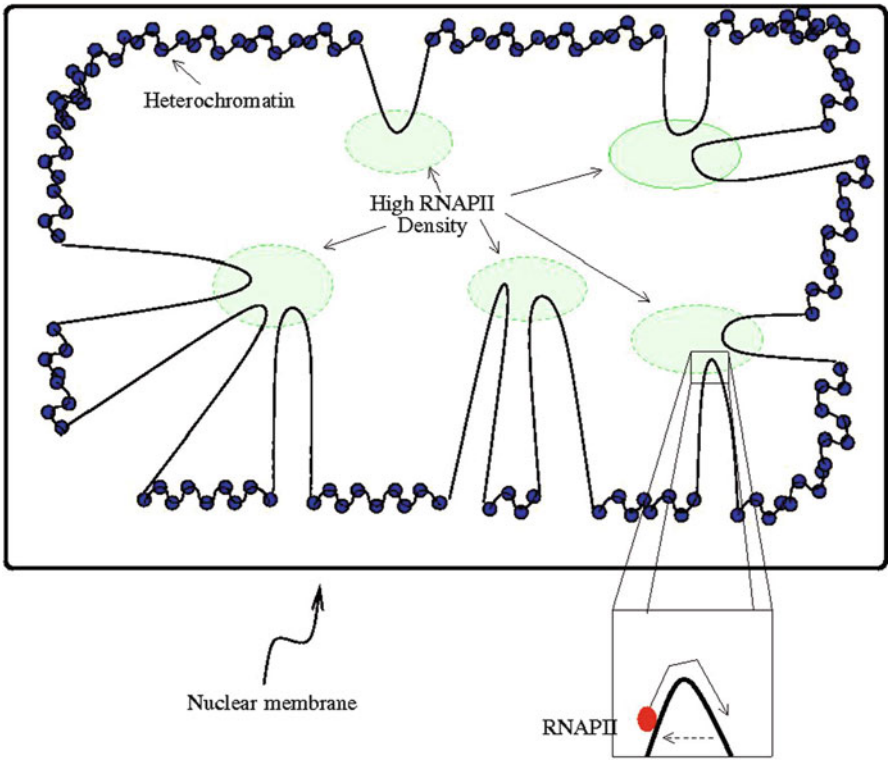

Fig. 2.10 A schematic diagram of chromatin organisation and transcription factories as outlined in Sects. 2.4.1 and 2.4.2. Heterochromatin (*blue* and *black circles* and *lines*) generally lies close to the nuclear membrane. Free DNA loops extend into the centre of the nucleus where it passes through regions of high RNA PII density referred to as transcription factories. Loops colocalise and hence are co-regulated exhibiting a similar noise structure in their expression. This is consistent with the stochastic nature of expression. The expanded region to the bottom of the diagram posits a hypothesis that as the start and end regions of the gene are physically close to each other, RNA PII can be re-recruited for transcription or be used for surveillance (Color figure online)

aspects of transcriptional regulation constrain the organisation of genes on eukaryotic chromosomes [119–121], with the 3D regulation of gene expression directly impacting upon genome evolution [122].

2.4.2 Loops and Networks of DNA Interactions Regulate Gene Expression

Genomes show tissue-specific spatial organisation [123] and cell nuclei frequently contain chromosome territories [124]. There is increasing evidence for three-dimensional networks of chromosomal interactions [125].

The topology of DNA around individual genes modifies gene regulation. As shown in the blown up region of Fig. 2.10, loop structures in which the promoter and terminator of a gene are in close proximity are associated with gene activity [22]. The role of the loop may be to increase the efficiency of recycling RNAP II back to a promoter after it has reached the end of a gene. The loop may also be involved in surveillance, with the results of an initial round of transcription being checked to ensure authentic signals are in place.

There are also cell-type-specific long-range looping interactions between enhancers and promoters which establish three-dimensional chromatin structures, such as for the CFTR gene [126, 127].

The conformational contacts between separate regions of chromatin change during cellular differentiation [128]. For example, extensive spatial chromatin remodelling accompanies gene repression during cellular differentiation [128], with repression of Hox A9, 10, 11 and 13 expression associated with the formation of distinct higher-order chromatin contacts between genes. Whereas, different chromatin conformations are associated with transcriptional activity. Major changes in higher-order structures of chromatin interactions are being associated with the regulation of transcriptional activity in increasing numbers of gene clusters, including the Bithorax complex in Drosophila [129] and the human apolipoprotein [130] and Hox A [128] gene clusters. The chromatin conformations may act as an epigenetic memory [129]. The conformation changes during differentiation may also be evolutionarily conserved [128].

DNA regulatory elements known as insulators mediate chromatin interactions, resulting in the formation of chromatin loops [131]. The name arises from the insulator's role in preventing inappropriate interactions between groups of enhancers and promoters. CCCTC-binding factor (CTCF) is one such insulator protein. CTCF contains three domains, one of which is a DNA-binding domain with 11 zinc fingers. It is evolutionally conserved from insects to mammals, and over 80 % amino acid residues are identical between human, chicken and frog and up to 100 % conservation within the zinc finger-containing region [132]. CTCF binds in tens of thousands of places across the genome, with the binding sites grouping into different classes [133], each of which exhibits distinct evolutionary, genomic, epigenomic and transcriptomic features. The chromatin architecture and form at CTCF-binding sites can result in cell-type-specific changes [134]. Understanding the code by which CTCF acts to coordinate the three-dimensional position and regulation of genes within a cell's nucleus is being actively sought [135]. CTCF is believed to fit tightly into the linker region between nucleosomes [135], which results in positioning of a nucleosome over a site acting to occlude the binding of CTCF [136]. Furthermore, CTCF is sensitive to the presence of a 5-methyl group in the major groove of DNA [96] and CTCF can only bind to unmethylated DNA [135]. Moreover, the binding of CTCF, possibly through the action of chromatin remodelers, acts to accurately position 20 nucleosomes, enriched in the histone variant H2A.Z, both upstream and downstream of the site [137]. CTCF also acts to demarcate cell-type-specific chromatin domains associated with active (H2AK5ac) and repressive (H3K27me3) histone modifications [136]. CTCF has also been found

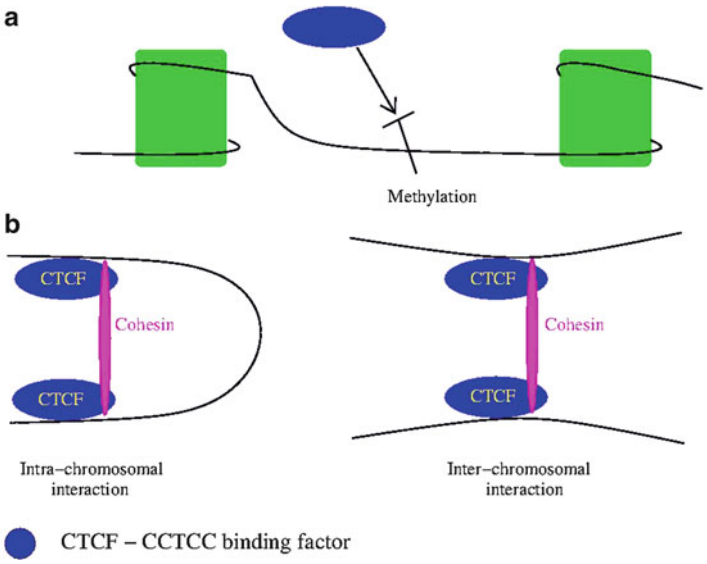

Fig. 2.11 A schematic diagram of how CTCF and cohesin can bind chromosomes as explained in Sect. 2.4.2. In (**a**) CTCF can bind between nucleosomes but cannot bind if its site is methylated. In (**b**) CTCF in conjunction with cohesin can bind between the same chromosome or between different chromosomes

to have a close relationship with the borders of lamina-associated domains [138], 0.1–10 megabase domains that are believed to anchor chromosomes to the nuclear envelope.

The important role that CTCF plays in establishing patterns of nuclear architecture and transcriptional control in vertebrates [139] is likely related to CTCF binding to cohesion [140–143], which creates intrachromosomal and interchromosomal links (shown in Fig. 2.11), resulting in a cell-type-specific network [134] that determines the three-dimensional structure of the genome [144]. Cohesin and CTCF are also involved in the maintenance of imprinting of loci such as the IGF2 (insulin-like growth factor 2)/H19 (H19 fetal liver mRNA) genes. A set of enhancers downstream of H19 play a role in regulating expression of both IFG2 and H19 – within developing embryos IGF2 is paternally expressed and H19 is maternally expressed [141]. On the maternal locus there are two unmethylated regions between IGF2 and H19 and where CTCF and a ring of cohesion can associate. The interaction between CTCF cohesion from this pair of regions results in a loop of chromatin-containing IGF2 which is then insulted from the action of the downstream enhancers, and only H19 is subsequently expressed. However, on the paternal locus, the region between IGF2 and H19 is methylated resulting in CTCF being unable to bind, leading to the H19 locus being bypassed and IGF2 being able to interact with the enhancers downstream of H19 [141].

Other elements may be involved in forming higher-order chromatin structures in the nucleus. For example, Polycomb response elements mediate the formation of chromosome higher-order structures in the Bithorax complex [129].

2.5 Regulation of Gene Expression by Non-coding RNA

2.5.1 Short Non-coding RNAs Associated with the Start, End and Enhancers of Genes

Short RNAs cluster at the 5′ and 3′ ends of genes [145]. A class of short transcripts close to transcription start sites of genes have been observed to be present at low abundance [79]. They have been named by several groups (promoter-associated sRNAs (PASRs, [145]), transcription start-site-associated RNAs (TSSa-RNAs, [146]) hereafter PASRs). PASRs are mostly derived from nucleosome-free DNA [79]. As shown in Fig. 2.12, they flank active promoters, with a peak in the abundance of short RNA antisense transcripts found ∼250 nucleotides upstream of a gene's transcription start site [146, 147] and a peak in the abundance of short sense transcripts found between approximately 50 nucleotides [146] and 2.5 kilobases [148] downstream of the transcription start site [147]. Such divergent transcription appears common for active promoters as most of them have engaged polymerases upstream, in an orientation opposite to the proximal gene [147]. There is a correlation in expression between PASRs and their proximal gene, suggesting they are both responding to a common inducement of expression, even though the transcripts are in opposite directions. The density of antisense termini-associated sRNAs (TASRs), found towards the 3′ ends of genes, is similarly correlated with the expression of the proximal gene [145].

A further source of gene-associated short RNAs is enhancers [149, 150]. Enhancer RNAs (eRNAs) have already been found in macrophages [149] and neurons [150] and it is likely that they will be identified in many, if not all, mammalian cell types. Enhancers overlap a sizeable fraction of extragenic transcription sites in higher eukaryotes [149]. In the studies of [149, 150], only a fraction of all enhancers were found to be associated with RNAPII and eRNA synthesis, suggesting that

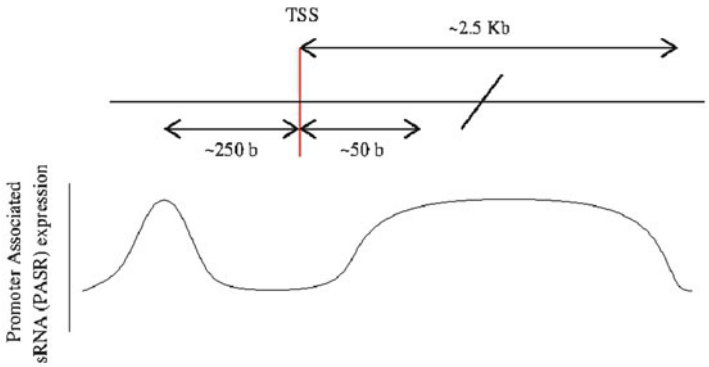

Fig. 2.12 A schematic diagram of the relationship between PASR expression levels and a TSS as explained in Sect. 2.5.1

there are a number of regulatory components involved with each enhancer. Changes in eRNA synthesis are correlated strongly with changes of mRNA expression at nearby genes [150], suggesting that eRNA synthesis may require a dynamic interaction between an enhancer and a promoter. Furthermore, upstream extragenic transcription frequently precedes the induction of an adjacent coding gene [149]. Transcripts from enhancers are not polyadenylated and they show little bias in transcribing both strands [150] as well as being very unstable [149].

The level of H3K4me at the enhancer and eRNA synthesis are tightly correlated, and so the process of eRNA synthesis may be to establish and maintain chromatin in a state required for enhancer function [149]. Indeed, it is likely that the function of many of these gene-associated short RNAs, including PASRs and TASRs, is to mediate transcription-coupled changes in chromatin structure [79]. Such changes may involve the prevention of nucleosomes obstructing transcription factor binding sites [147] or facilitating initiation through the impact of negative supercoiling [146] behind the passage of RNAP II. These will help promoter regions maintain a state poised for subsequent regulation [146]. Polymerase resides on approximately 30 % of human genes, with RNAP II observed to be pausing, appearing to wait for a signal to begin elongating [147]. Genes that are developmentally regulated or that respond to extracellular triggers are those that are likely to have pre-engaged RNAP II [22], so as to speed up the rate at which the gene is ready for transcription. It is likely that there is a rate-limiting step that stops RNAPII fully escaping into elongating [79]. It is presently unclear what this trigger is, but it is likely to be associated with pre-mRNA processing [22].

2.5.2 Long Non-coding RNAs

Significant numbers of long ncRNAs are regulated during development [151]. In particular, the binding of transcription factors, along with evidence of selection, conserved secondary structure, splicing patterns and subcellular localisation, suggests the explicit regulation of non-coding transcription [152]. Long ncRNAs can act as coactivators of transcription factors [153]. They can also act as "ligands" for RNA-binding proteins, causing an allosteric change from an inactive to active conformation, which in turn can inhibit transcription through modifying transcription factor and histone acetyltransferases [154]. Non-coding RNAs also modulate the subcellular localisation of some transcription factors [151]. Non-coding RNAs can also bind to, and regulate the action of, RNA polymerase II during heat shock [155]. Also, some of the transcripts labelled as non-coding may in fact be the source of functional small peptides [156].

The wide variety of regulatory roles ncRNAs can play are shown in Fig. 2.13. A number of chromatin-modifying enzymes contain RNA-binding motifs [157] and long non-coding RNAs recruit chromatin-remodelling complexes to genomic loci [152, 158–162]. Long non-coding RNAs act to direct genomic methylation [163]. They also provide a scaffold for histone-modifying enzyme recruitment

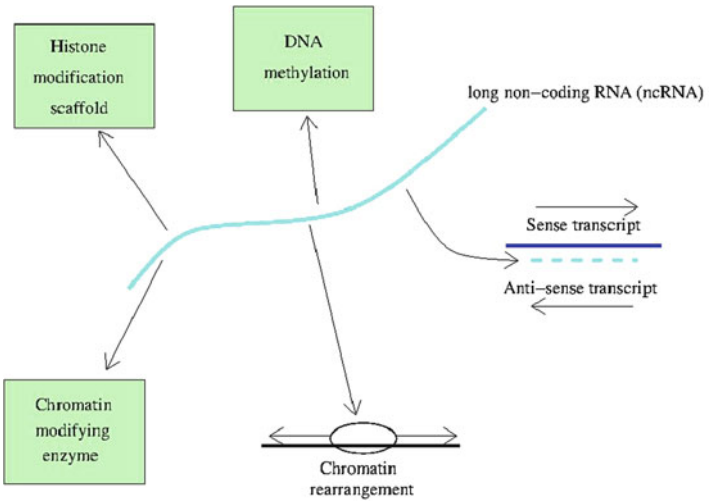

Fig. 2.13 An outline of the various roles long ncRNAs can play in regulation as explained in Sect. 2.5.2

[164], leading to heterochromatin formation [165]. The non-coding transcripts act as local modulators of chromatin structure, triggering chromatin modifications which then expand along the chromosome, even though the neighbouring regions are not complementary to the original transcript [164]. The expansion of the induced chromatin changes may just be restricted locally, or they can expand further and may underpin genomic imprinting [166] and X chromosome inactivation [167]. Another example is the expression of hundreds of long ncRNAs that are sequentially expressed along the Hox loci, defining chromatin domains of differential histone methylation and accessibility [168]. One of the ncRNAs in the Hox loci recruits the Polycomb chromatin-remodelling complex and silences transcription across 40 kb in trans through inducing chromatin to enter a repressive state.

Natural antisense transcripts can overlap part or all of another transcript [164] and many protein-coding genes can be regulated by their antisense transcript partners. The antisense transcripts can bind to their sense partners and enhance their stability, through modifying the binding of an HuR protein and suppressing deadenylation and decapping [169]. The binding of an antisense transcript can also induce changes in RNA secondary structure which act to expose AU-rich elements and make the sense transcript prone to degradation [170]. Interactions between sense and antisense transcripts can also block the binding sites of other regulatory factors such as microRNAs. This appears to be the case for β-secretase, a transcript regulated by its antisense partner and likely related to the pathogenesis of Alzheimer's disease [171].

Antisense RNAs typically undergo fewer splicing events than sense transcripts [172]. However, natural antisense transcripts can modify the alternative splicing isoforms of their sense partners [172, 173] and may also impact upon alternative polyadenylation [164]. Furthermore, long ncRNAs can be processed to yield small

RNAs and they can also modulate the efficiency by which other transcripts are cut into small RNAs and interact with the RNAi pathway [174]. Endogenous siRNAs have been observed to map to overlapping regions between sense and antisense RNAs, and the RNAi pathway could regulate both the sense and antisense transcripts in these cases [175]. However, the RNAi pathway is not responsible for antisense-mediated regulation of the expression of some genes [175]. Duplex formation of sense and antisense partners may also interact with the RNA-editing pathway [176].

In a number of cases, it appears to be the act of transcribing a non-coding transcript, rather than the transcript itself, which acts to regulate a nearby protein-coding gene. Transcriptional interference resulting from collisions between RNA polymerases producing the sense and overlapping antisense expression may occur [177], but this is likely not to be the predominant regulatory pathway mediated by antisense transcripts [164]. Transcription of an ncRNA can pass across the promoter of the protein-coding gene and interfere with transcription factor binding, preventing the expression of the protein-coding gene [151]. Transcriptional elongation induces the addition of histone marks that act to prevent transcription initiation from locations within the body of the transcript [151]. ncRNA transcription can induce histone modifications that repress the transcription of an overlapping protein-coding gene. Furthermore, continuous transcription of ncRNA can prevent silencing of genes by proteins such as Polycomb group proteins [178]. Non-coding RNAs can also help to recruit Trithorax group proteins [179] which help to main active transcription states by counteracting the effects of the Polycomb proteins.

2.5.3 Regulation by MicroRNAs

MicroRNAs (miRNAs) are short (∼22nt) non-coding single-stranded RNAs [180]. They function by usually repressing mRNAs post-transcriptionally through complementary binding to partial overlaps in target mRNAs. They play a central role in coordinating the activities of many thousands of transcripts and they play an integral role in the development and regulation of different cell types and tissues [181].

There are several good reviews of miRNA biogenesis, e.g. [182]. RNA polymerase II mediates the transcription of most miRNAs and is summarised here and Fig. 2.14. Pri-miRNAs are long primary transcripts which typically form a stem hairpin structure, a terminal loop and ssRNA flanking segments. The nuclear enzyme Drosha, assisted by DGCR8 (DiGeorge syndrome critical region gene 8), cleaves the RNA near the stem of the hairpin, about 11 bp away from the ssRNA–dsRNA junction [182]. This releases a pre-miRNA which is then exported from the nucleus by the protein exportin-5. In the cytoplasm, the enzyme Dicer further cleaves the pre-miRNA near the terminal loop to yield a duplex of ∼22nt. One of the strands is loaded into an Argonaute (AGO) protein, and this is used to guide complementary target mRNA sequences for repression.

miRNAs have played a significant role in the phenotypic evolution of metazoans [183] and there is a close coupling between miRNA evolution and the establishment

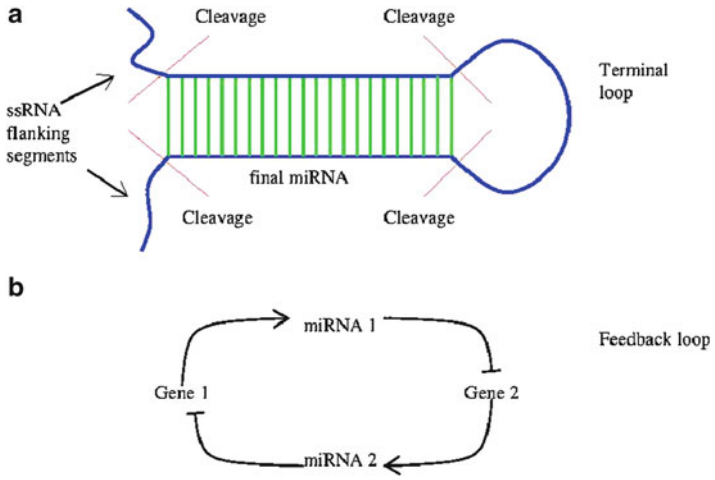

Fig. 2.14 (**a**) A schematic diagram of a pri-miRNA and the region that eventually forms the miRNA. In (**b**) we list one regulatory mechanism miRNAs can play in regulating genes explained in Sect. 2.5.3. This case ensures that either gene 1 or gene 2 is expressed

of tissue identities early in bilaterian evolution [184]. An expansion in the number of miRNAs has also been hypothesised to lie behind the origin of vertebrate complexity [185]. The increase in new miRNA families is likely due to the ease in which they are formed along with the wide impact they have on gene networks. The formation of a new miRNA is likely related to pervasive transcription of sequences containing hairpin loops, each of which is only a few mutations away from being a new miRNA [180, 186]. Once a miRNA is operational, and modifying the regulation of many genes, it undergoes very strong purifying selection meaning that their sequences are extremely well conserved [180, 186], making miRNAs excellent phylogenetic markers [187]. However, the targets to which miRNAs bind show little conservation in animals, indicating that miRNA regulatory networks have undergone extensive rewiring during metazoan evolution [180]. Unlike the continuous formation of new miRNA families, there has been a much smaller expansion and evolution of transcription factors during metazoan evolution [187]. Gene duplication is the dominant source of new transcription factors. There is a greater chance of evolutionary advantage for a duplicated transcription factor to undergo a few mutations and bind to new targets of DNA than it is for a non-transcription factor family member to mutate enough to be able to bind to DNA [187].

There appear to be two broad classes of miRNAs [188]. Class I miRNAs are regulated by large numbers of transcription factors and are likely to function within developmental programmes. Whereas class II miRNAs are regulated by small numbers of transcription factors and are likely to function in maintaining tissue identity in adults. The widespread regulation of genes by miRNAs leads to many pathologies resulting from disruptions in the regulation of miRNAs, and they are being increasingly identified as being involved in a range of diseases [189],

including many neurodegenerative diseases [190]. Because of concerns about off-target effects of new drugs, it is also being recognised increasingly that development in pharmacogenomics will require greater knowledge of miRNAs [191].

The expression of many transcription factors is subject to miRNA regulation. Feedback motifs are rare in pure transcription factor networks [192], and miRNAs provide the necessary post-transcriptional feedback [193, 194]. miRNAs usually repress gene expression, but not always [195]. One of the roles microRNAs might play is to tune expression at threshold points [183], such that stochastic gene expression programmes will have less noisy outcomes [196]. This type of regulation is required as noise can induce bimodality in positive transcriptional feedback loops [197]. The resulting robustness leads to stabilised developmental pathways, increasing phenotypic reproducibility [198]. The networks through which microRNAs act to regulate self-renewal in stem cells, as well as the transformation of stem cells into differentiated cells, are beginning to be mapped out [199].

There are differences between transcription factor and miRNA regulation related to biological processes in which they are involved. In animals, the repression of miRNAs is usually weak compared to TF-mediated repression [180] and it increasingly appears that miRNAs act to fine-tune the translational and transcriptional output of TFs [200]. miRNAs can act to quickly suppress or reactivate protein production at ribosomes [201], whereas changes in TF binding modifying transcription rates take longer before the information feeds through to protein production [195]. Furthermore, the actions of miRNAs, unlike TFs, can be localised to different parts of a cell. This compartmentalising can then be used in processes such as neurons requiring to regulate gene expression on a synapse-specific scale rather than across the cell [202].

2.6 Common Themes

2.6.1 Structural Considerations

2.6.1.1 The Shape of RNA Impact upon Gene Regulation

RNA molecules form stable secondary and tertiary structures in vitro and in vivo [203]. RNA secondary structures play an important role in binding splicing factors [38], and the search for novel RNA-binding targets for well-known proteins can be enhanced if secondary structure is taken into account [204]. Furthermore, the binding of microRNAs to target sequences depends upon the local tertiary structure of RNA [205]. Moreover, RNA editing also depends on the structure of the RNA, as ADAR converts adenosines to inosines (A to I) using double-stranded RNA substrates.

The reliable computational prediction of RNA structure would be very useful in understanding its underlying function; however, despite some progress in the area,

it remains a highly challenging problem [206]. Buratti and Baralle [37] cautioned against the use of *in silico* predictions of pre-mRNA structure such as those obtained by Mfold [207] and Pfold [208]. Buratti and Baralle [37] noted that existing computer algorithms provide a folding prediction (and usually more than one) for virtually any RNA sequence and are strongly biased by the length of the RNA examined. Buratti and Baralle [37] discussed the example of NF-1 gene transcripts, which are implicated in the generation of human tumours. Correlations between the in silico changes in secondary structure and splicing in NF-1 are heavily dependent on the RNA window. This makes it difficult to assign significance to them.

2.6.1.2 The Shape of DNA Impacts upon Gene Regulation

Gene regulation is related to the properties of chromatin in the nucleus. This ranges from posttranslational modifications of histone tails which alter their propensity to bind to each other or to allow transcription, to the movement of histones affecting accessibility of binding sites for transcriptional factors, to the looping of DNA that bring the 5′ and 3′ ends of active genes into proximity, to CTCF acting to regulate networks of binding between different chromosomes and to the movement of co-regulated genes in and out of transcriptional factories.

2.6.2 Gene Structure Is Written Out in Chromatin

The density of nucleosomes and lengths of linkers between nucleosomes differ in exons and introns [73]. The positioning of nucleosomes, as well as histone modifications, is involved in co-transcriptional splicing decisions [48, 49]. Moreover, nucleosome depletion has also been associated with the regulation of polyadenylation [62].

Gene-body DNA methylation also likely plays a role in splicing as exons show higher methylation fractions than do introns and there are sharp transitions in methylation states at exon–intron junctions [50]. Differences in DNA methylation also occur at transcriptional start sites and termination sites [50, 89]. DNA methylation acts to make DNA more rigid [96], and so the regulation of co-transcriptional events may involve a feedback between nucleosome positioning and DNA methylation. Interactions between DNA methylation and histone tail PTMs may also play a role in regulating these events.

There are peaks in DNA methylation as well as in the density of H2A.Z-containing nucleosomes just downstream of start codons and just upstream of stop codons in human T-cells [89]. Given that the use of start and stop codons is not required till translation, an exciting possibility arising from the observations of [89] is that the chromatin markings may be indicative of co-transcriptional modification events which act to label where a protein starts and finishes.

2.6.3 Interacting Codes

Gene regulation appears to be intimately controlled through the actions of several codes – namely, the modulation of a regulatory mechanism by the DNA or protein sequence it encounters. Within the DNA sequence, there is a nucleosome positioning code and this is increasingly well understood. There is likely a CTCF code which helps to regulate the three-dimensional positioning of genes within a cell's nucleus. The heptad repeats in the CTD of RNAPII can undergo different types of posttranslational modifications and these are intimately involved in regulating the binding of factors required for many of the steps in transcription, post-transcriptional processing and termination of transcription [24].

The tails of histones can undergo many types of posttranslational modifications. However, cracking this histone code is proving challenging [66]. This is further complicated by the interactions that occur between the CTD code and the histone code [24]. Moreover, other chromatin-associated proteins, such as HP1, are also posttranslationally modified resulting in the possibility of subcodes with the histone code [209]. Furthermore, H3 histone variants modify the properties of chromatin and their distribution along chromosomes is analogous to a barcode [67]. In addition, much of the impact of H1 variants on the histone code remains to be determined [68].

2.6.4 Kinetics and Competition Between Processes Underpin Gene Regulation

Self-organisation and assembly of structures such as Cajal bodies [104] depends upon the time-dependent concentrations of subcomponents. Furthermore, the movement of genes in and out of transcription factories will also result in changes to the rate of expression [112]. The form of chromatin also causes differences in elongation rates which in turn affect splice site selection [47]. There are a number of other steps available for regulation in splicing [38], encompassing a large number of kinetic events. The kinetic parameters may have a determining role in splice-site choice [36]. It is clear that in order to model how the changing form of recently transcribed RNA impacts on post-transcriptional processes such as splicing, we need to consider the implications of RNA secondary structure, the binding of ribonuclear proteins, the speed of transcription, the form of chromatin and any histone modifications and the dynamic interplay between all these processes.

There is binding competition between a number of processes. Many transcription factors and chromatin-associated proteins have highly transient interactions with chromatin, undergoing rapid cycles of binding and unbinding [103]. The high levels of molecular crowding in the nucleus help to increase the efficiency of binding resulting in local changes in density dramatically altering the rate at which nuclear structures form [103]. Nucleosomes and transcription factors each have affinities

for a DNA sequence and competitive and cooperative interactions between these proteins act to determine their occupancy [70]. The cycling of factors on and off promoters enables the formation of poised transcriptional complexes [90], which are typically observed in approximately 30 % of promoters [22].

There are also interactions between different types of regulators. miRNAs can bind to exon–exon junctions, suggesting that they can target splice isoforms [210]. An intron retention event can lead to transcripts containing miRNA-binding sites that they would not otherwise have [211]. Moreover, the biogenesis of miRNAs can result in crosstalk to pre-mRNA splicing [212]. The binding sites of RNA-binding proteins can overlap with microRNA target sites [213] and RNA structure also acts to modify microRNA binding [205]. RNA editing is also coordinated with splicing [54] and there is a close interplay between editing and miRNAs [52]. There is also crosstalk between RNA editing and RNA interference [214]. Next-generation sequencing will increasingly underpin experiments to map out these networks of interactions [43].

2.6.5 Gene Regulation Can Be Tissue Specific

Three-quarters of the mRNA in a cell are common across tissues, and about 8,000, or approximately one-third, of human protein-coding genes are ubiquitously expressed [12]. However, much of the rest of RNA appears to be tissue specific and likely underpins phenotypic complexity in mammals. Alternative splicing and alternative polyadenylation vary between tissues [33]. The majority of retrotransposon expression is tissue specific [215]. RNA editing is enhanced in the brain [52]. Long non-coding RNAs show developmental regulation [151]. miRNAs function in developmental programmes and maintain tissue identity [181]. The state of chromatin also changes as cells transform from pluripotent stem cells into multipotent progenitor cells and the composition of chromatin-remodelling complexes are tissue specific [92].

2.7 Putting It All Together: How Would You Cope if You Could Sequence Everything?

Despite the ferocious complexity of the different mechanisms involved in gene regulation, common themes are emerging as demonstrated in the previous section and summarised as a mind map in Fig. 2.15. It is not unreasonable to assume that in the near future next-generation sequencing techniques will allow the sequencing of all the DNA and expressed types of RNA involved in a given response or process [4]. Such a range of data will be necessary to unravel the complexity of the multilayered regulation of gene expression.

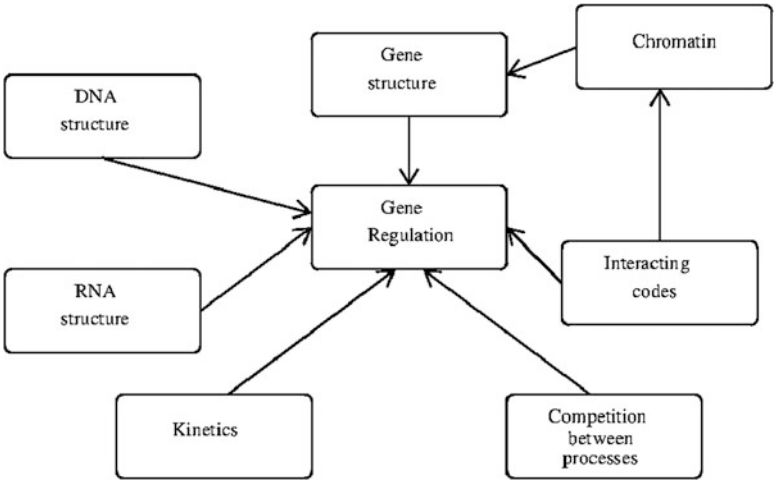

Fig. 2.15 Mind map of Sect. 2.6

A better understanding of chromatin and RNA biology will play a central role in how we use cross-species information reliably. For example, alternative splicing is likely to be one of the principal contributors to the evolution of phenotypic complexity in mammals [33]. The splicing patterns in mammalian model organisms, such as mice, are therefore likely to differ with humans in a number of ways, and so differing populations of isoforms may complicate the interpretation of the negative side effects of pharmaceuticals. RNA editing will also result in different transcript populations in humans compared to other mammals [55], again complicating studies to identify how drugs impact on gene expression systems. A further complication is that of miRNAs, which play a key role in regulating tissue-specific transcription. There are more than 100 extra miRNAs in humans compared to chimpanzees and more than 150 extra miRNAs in human compared to mouse [183], and these extra miRNAs are likely to result in gene expression patterns being found in humans that are not found in our nearest neighbours.

One of the fundamental goals of systems biology is to generate meaningful quantitative models of the regulation of gene expression. In order to do this, not only must there be a significant increase in the types of data being collated (as we have shown in this review), the amount of each type must also be increased considerably. This is necessary to circumvent the so-called curse of dimensionality where the output from all the genes is measured but only in a small number of conditions. It will be necessary to bring multiple studies together, so as to identify some of the subtle changes in gene expression that are biologically meaningful [17]. This indicates a huge increase in the amount of data being gathered, processed and analysed. Already, genomics is one of many fields facing a deluge of data [216]. Bioinformatics repositories are already at the petabyte scale [217] – the growth of sequencing data will result in the repositories transcending the exabyte scale within

the decade. The archiving of next-generation sequencing data has well-established resources such as the Short Read Archive [218]. In order to cope with the flow of data, the Short Read Archive is adopting high-speed file transfer protocols, at present `fasp` (Aspera Inc.). However, the transfer of data between external bioinformatics laboratories is already leading to increasing problems in keeping up-to-date [219] and things will only get worse. Moreover, the management of next-generation sequencing data within institutions is already leading to a number of bottlenecks, requiring increasing resources to be spent on systems administration and computers [220] rather than on personnel to make use of the data. A further cost which is only likely to escalate is that of power to run the facility. The computational and staffing issues being faced by the genomics community are likely to limit the democratisation of sequencing.

Genomics is not alone in facing a need for processing very large datasets. The state of the art has arisen from commercial use [216], with organisations such as Google efficiently processing searches and data mining on enormous datasets. Virtually all of these organisations are rapidly implementing data centres to cope with their data-processing requirements. The economies of scales associated with centres mean that they can sell the resources to external users, through the cloud computing model. Bioinformaticians have now begun to look at cloud computing as one feasible solution to cope with the rapid growth of data [221]. There are now increasing needs for large-scale biological data and computational infrastructure to be developed on international scales, such as ELIXIR in Europe [222].

All of this will result in ever larger datasets requiring ever larger computational and experimental infrastructure, as well as larger-sized teams to cope with the data and use it to discover new biology. We can sequence everything, we can afford to do so, we can learn huge amounts, and we will have to likely change some of our working practises to be able to fully utilise the technology.

References

1. Schloss J (2008) How to get genomes at one ten-thousandth the cost. Nat Biotechnol 26:1113–1115
2. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145
3. Branton D, Deamer D, Marziali A, Bayley H, Benner S, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich S, Krstic P, Lindsay S, Ling X, Mastrangelo C, Meller A, Oliver J, Pershin Y, Ramsey J, Riehn R, Soni G, Tabard-Cossa V, Wanunu M, Wiggin M, Schloss J (2008) The potential and challenges of nanopore sequencing. Nat Biotechnol 26:1146–1153
4. Kahvejian A, Quackenbush J, Thompson J (2008) What would you do if you could sequence everything? Nat Biotechnol 26:1125–1133
5. Pleasance E, Cheetham R, Stephens P, McBride D, Humphray S, Greenman C, Varela I, Lin M, Ordonez G, Bignell G, Ye K, Alipaz J, Bauer M, Beare D, Butler A, Carter R, Chen L, Cox A, Edkins S, Kokko-Gonzales P, Gormley N, Grocock R, Haudenschild C, Hims M, James T, Jia M, Kingsbury Z, Leroy C, Marshall J, Menzies A, Mudie L, Ning Z, Royce T, Schulz-Trieglaff O, Spiridou A, Stebbings L, Szajkowski L, Teague J, Williamson D, Chin L, Ross M, Campbell P, Bentley D, Futreal P, Stratton M (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463:191–196

6. Denoeud F, Aury J, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F (2008) Annotating genomes with massive-scale RNA sequencing. Genome Biol 9:R175
7. Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628
8. Yassour M, Kaplan T, Fraser H, Levin J, Pfiffner J, Adiconis X, Schroth G, Luo S, Khrebtukova I, Gnirke A, Nusbaum C, Thompson D, Friedman N, Regev A (2009) Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc Natl Acad Sci U S A 106:3264–3269
9. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63
10. Morozova O, Hirst M, Marra M (2009) Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet 10:135–151
11. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods Suppl 6:S22–S32
12. Ramsköld D, Wang E, Burge C, Sandberg R (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequenced data. PLoS Comput Biol 5:e1000598
13. Ma W, Wong W (2011) The analysis of Chip-Seq data. Methods Enzymol 497:51–73
14. Wall P, Leebens-Mack J, Chanderbali A, Barakat A, Wolcott E, Liang H, Landherr L, Tomsho L, Hu Y, Carlson J, Ma H, Schuster S, Soltis D, Soltis P, Altman N, dePamphilis C (2009) Comparison of next-generation sequencing technologies for transcriptome characterization. BMC Genomics 10:347
15. Oshlack A, Wakefield M (2009) Transcript length bias in RNA-seq data confounds systems biology. Biol Direct 4:14
16. Coombs A (2008) The sequencing shakeup. Nat Biotechnol 26:1109–1112
17. Needham C, Manfield I, Bulpitt A, Gilmartin P, Westhead D (2009) From gene expression to gene regulatory networks in Arabidopsis thaliana. BMC Syst Biol 3:85
18. Tsankov A, Brown C, Yu M, Win M, Silver P, Casolari J (2006) Communication between levels of transcriptional control improves robustness and adaptivity. Mol Syst Biol 2:65
19. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. Nat Genet 22(3):281–285
20. Scheinine A, Mentzen WI, Fotia G, Pieroni E, Maggio F, Mancosu G, De La Fuente A (2009) Inferring gene networks: dream or nightmare?: Part 2: challenges 4 and 5. Ann N Y Acad Sci 1158:287–301
21. Komili S, Silver P (2008) Coupling and coordination in gene expression processes: a systems biology view. Nat Rev Genet 9:38–48
22. Moore M, Proudfoot N (2009) Pre-mRNA processing reaches back to transcription and ahead to translation. Cell 136:688–700
23. Karlić R, Chung H-R, Lasserre J, Vlahoviček K, Vingron M (2010) Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A 107:2926–2931
24. Egloff S, Murphy S (2008) Cracking the RNA polymerase II CTD code. Trends Genet 24:280–288
25. Klaff P, Riesner D, Steger G (1996) RNA structure and the regulation of gene expression. Plant Mol Biol 32:89–106
26. Antson A (2000) Single stranded-RNA binding proteins. Curr Opin Struct Biol 10:87
27. Carlson C, Stephens O, Beal P (2003) Recognition of double-stranded RNA by proteins and small molecules. Biopolymers 70:86–102
28. Dreyfuss G, Matunis M, Pinol-Roma S, Burd C (1993) hnRNP proteins and the biogenesis of mRNA. Annu Rev Biochem 62:289–321
29. Eperon L, Graham I, Griffiths A, Eperon I (1988) Effects of RNA secondary structure on alternative splicing of pre-mRNA: is folding limited to a region behind the transcribing RNA polymerase? Cell 54:393–401
30. Aguilera A (2005) Cotranscriptional mRNP assembly: from the DNA to the nuclear pore. Curr Opin Cell Biol 17:242–250

31. Li X, Manley J (2006) Cotranscriptional processes and their influence on genome stability. Genes Dev 20:1838–1847
32. Lindahl T, Nyberg B (1974) Heat-induced deamination of cytosine residues in deoxyribonucleic acid. Biochemistry 13:3405–3410
33. Wang E, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore S, Schroth G, Burge C (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470–476
34. Melamud E, Moult J (2009) Stochastic noise in splicing machinery. Nucleic Acids Res 37:4873–4886
35. Hallegger M, Llorian M, Smith C (2010) Alternative splicing: global insights. FEBS J 277:856–866
36. Nilsen T, Graveley B (2010) Expansion of the eukaryotic proteome by alternative splicing. Nature 463:457–463
37. Buratti E, Baralle F (2004) Influence of RNA secondary structure on the pre-mRNA splicing process. Mol Cell Biol 24:10505–10514
38. Smith D, Query C, Konarska M (2008) Nought may endure but mutability: spliceosome dynamics and the regulation of splicing. Mol Cell 30:657–666
39. Nagel R, Lancaster A, Zahler A (1998) Specific binding of an exonic splicing enhancer by the pre-mRNA splicing factor SRp55. RNA 4:11–23
40. Damgaard C, Tange T, Kjems J (2002) hnRNP A1 controls HIV-1 mRNA splicing through cooperative binding to intro and exon splicing silencers in the context of a conserved secondary structure. RNA 8:1401–1415
41. Blencowe B (2000) Exonic splicing enhancers: mechanisms of action, diversity and role in human genetic diseases. Trends Biochem Sci 25:106–110
42. Matsuo M, Nishio H, Kitoh Y, Francke U, Nakamura H (1992) Partial deletion of a dystrophin gene leads to exon skipping and to loss of an intra-exon hairpin structure from the predicted mRNA precursor. Biochem Biophys Res Commun 182:495–500
43. Licatalosi D, Darnell R (2010) RNA processing and its regulation: global insights into biological networks. Nat Rev Genet 11:75–87
44. Barash Y, Calarco J, Gao W, Pan Q, Wang X, Shai O, Blencowe B, Frey B (2010) Deciphering the splicing code. Nature 465:53–59
45. Kornblihtt A (2006) Chromatin, transcription elongation and alternative splicing. Nat Struct Mol Biol 13:5–7
46. Proudfoot N (2003) Dawdling polymerases allow introns time to splice. Nat Struct Mol Biol 10:876–878
47. Li B, Carey M, Workman J (2007) The role of chromatin during transcription. Cell 128:707–719
48. Luco R, Pan Q, Tominaga K, Blencowe B, Pereira-Smith O, Misteli T (2010) Regulation of alternative splicing by histone modifications. Science 327:996–1000
49. Schwartz S, Meshorer E, Ast G (2009) Chromatin organization marks exon-intron structure. Nat Struct Mol Biol 16:990–996
50. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong C, Low H, Sung K, Rigoutsos I, Loring J, Wei C (2010) Dynamic changes in the human methylome during differentiation. Genome Res 20:320–331
51. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones D, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837
52. Nishikura K (2010) Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem 79:2.1–2.29
53. Barak M, Levanon E, Eisenberg E, Paz N, Rechavi G, Church G, Mehr R (2009) Evidence for large diversity in the human transcriptome created by Alu RNA editing. Nucleic Acids Res 37:6905–6915
54. Laurencikiene J, Källman A, Fong N, Bentley D, Öhman M (2006) RNA editing and alternative splicing: the importance of co-transcriptional coordination. EMBO Rep 7:303–307

55. Gommans W, Mullen S, Maas S (2009) RNA editing: a driving force for adaptive evolution? BioEssays 31:1137–1145
56. Maas S, Kawahara Y, Tamburro K, Nishikura K (2006) A-to-I RNA editing and human disease. RNA Biol 3:1–9
57. Niswender C, Herrick-Davis K, Dilley G, Meltzer H, Overholser J, Stockmeier C, Emeson R, Sanders-Bush E (2001) RNA editing of the Human Serotonin 5-HT_{2C} receptor: alterations in suicide and implications for serotonergic pharmacotherapy. Neuropsychopharmacology 24:478–491
58. Sebastiani P, Montano M, Puca A, Solovieff N, Kojima T, Wang M, Melista E, Meltzer M, Fischer S, Andersen S, Hartley S, Sedgewick A, Yasumichi A, Bergman A, Barzilai N, Terry D, Riva A, Anselmi C, Malovini A, Kitamoto A, Sawabe M, Arai T, Gondo Y, Steinberg M, Hirose N, Atzmon G, Ruvkun G, Bladwin C, Perls T (2009) RNA editing genes associated with extreme old age in humans and with lifespan in C. elegans. PLoS One 4:e8210
59. Li J, Levanon E, Yoon J-K, Aach J, Xie B, LeProust E, Zhang K, Gao Y, Church G (2009) Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science 324:1210–1213
60. Ebhardt H, Tsang H, Dai D, Liu Y, Bostan B, Fahlman R (2009) Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. Nucleic Acids Res 37:2461–2470
61. Millevoi S, Vagner S (2010) Molecular mechanisms of eukaryotic pre-mRNA 3′ end processing regulation. Nucleic Acids Res 38:2757–2774
62. Spies N, Nielsen C, Padgett R, Burge C (2009) Biased chromatin signatures around polyadenylation sites and exons. Mol Cell 36:245–254
63. Mignone F, Gissi C, Liuni S, Pesole G (2002) Untranslated regions of mRNAs. Genome Biol 3:reviews0004.1–0004.10
64. Chepelev I, Wei G, Tang Q, Zhao K (2009) Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic Acids Res 37:e106
65. Jiang C, Pugh B (2009) Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 10:161–172
66. Campos E, Reinberg D (2009) Histones: annotating chromatin. Annu Rev Genet 43:559–599
67. Hake S, Allis C (2006) Histone H3 variants and their potential role in indexing mammalian genomes: the H3 barcode hypothesis. Proc Natl Acad Sci U S A 103:6428–6435
68. Godde J, Ura K (2008) Cracking the enigmatic linker histone code. J Biochem 143:287–293
69. Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore I, Wang J-P, Widom J (2006) A genomic code for nucleosome positioning. Nature 442:772–778
70. Segal E, Widom J (2009) What controls nucleosome positions? Trends Genet 25:335–343
71. Kaplan N, Moore I, Fondufe-Mittendorf Y, Gossett A, Tillo D, Field Y, LeProust E, Hughes T, Lieb J, Widom J, Segal E (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 458:362–366
72. Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore I, Sharon E, Lubling Y, Widom J, Segal E (2008) Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Comput Biol 4:e1000216
73. Cohanim A, Haran T (2009) The coexistence of the nucleosome positioning code with the genetic code on eukaryotic genomes. Nucleic Acids Res 37:6466–6476
74. Warnecke T, Batada N, Hurst L (2008) The impact of nucleosome code on protein-coding sequence evolution in yeast. PLoS Genet 4:e1000250
75. Washietl S, Machné R, Goldman N (2008) Evolutionary footprints of nucleosome positions in yeast. Trends Genet 24:583–587
76. Sasaki S, Mello C, Shimada A, Nakatani Y, Hashimoto S, Ogawa M, Matsushima K, Gu S, Kashara M, Ahsan B, Sasaki A, Saito T, Suzuki Y, Sugano S, Kohara Y, Takeda H, Fire A, Morishita S (2009) Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. Science 323:401–404
77. Tirosh I, Barkai N (2008) Two strategies for gene regulation by promoter nucleosomes. Genome Res 18:1084–1091

78. Cheung V, Chua G, Batada N, Landry C, Michnick S, Hughes T, Winston F (2008) Chromatin- and transcription related factors repress transcription from within coding regions throughout the Saccharomyces cerevisiae genome. PLoS Biol 6:2550–2562
79. Buratowski S (2008) Gene expression – where to start? Science 322:1804–1805
80. Cairns B (2009) The logic of chromatin architecture and remodelling at promoters. Nature 461:193–198
81. Albert I, Mavrich T, Tomsho L, Qi J, Zanton S, Schuster S, Pugh B (2007) Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. Nature 446:572–576
82. Choi J, Kim Y-J (2009) Intrinsic variability of gene expression encoded in nucleosome positioning sequences. Nat Genet 41:498–503
83. Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume D (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat Rev Genet 8:424–436
84. Irizarry R, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, Ji H, Potash J, Sabunciyan S, Feinberg A (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 41:178–186
85. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple C, Taylor M, Engström P, Frith M, Forrest A, Alkema W, Tan S, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond S, Wells C, Orlando V, Wahlestedt C, Liu E, Harbers M, Kawai J, Bajic V, Hume D, Hayashizaki Y (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38:626–635
86. Taylor M, Kai C, Kawai J, Carninci P, Hayashizaki Y, Semple C (2006) Heterotachy in mammalian promoter evolution. PLoS Genet 2:627–639
87. Tillo D, Kaplan N, Moore I, Fondufe-Mittendorf Y, Gossett A, Field Y, Lieb J, Widom J, Segal E, Hughes T (2010) High nucleosome occupancy is encoded at human regulatory sequences. PLoS One 5:e9129
88. Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcárcel J, Guigó R (2009) Nucleosome positioning as a determinant of exon recognition. Nat Struct Mol Biol 16:996–1001
89. Choi J, Bae J-B, Lyu J, Kim T-K, Kim Y-J (2009) Nucleosome depletion and DNA methylation at coding region boundaries. Genome Biol 10:R89
90. Mellor J (2006) Dynamic nucleosomes and gene transcription. Trends Genet 22:320–329
91. Segal E, Widom J (2009) From DNA sequence to transcriptional behaviour: a quantitative approach. Nat Rev Genet 10:443–456
92. Ho L, Crabtree G (2010) Chromatin remodelling during development. Nature 463:474–484
93. Whitehouse I, Rando O, Delrow J, Tsukiyama T (2007) Chromatin remodelling at promoters suppresses antisense transcription. Nature 450:1031–1035
94. Clapier C, Cairns B (2009) The biology of chromatin remodelling complexes. Annu Rev Biochem 78:273–304
95. Berger S (2007) The complex language of chromatin regulation during transcription. Nature 447:407–412
96. Pennings S, Allan J, Davey C (2005) DNA methylation, nucleosome formation and positioning. Brief Funct Genomics Proteomics 3:351–361
97. Cedar H, Bergman Y (2009) Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet 10:295–304
98. Hemberger M, Dean W, Reik W (2009) Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. Nat Rev Mol Cell Biol 10:526–537
99. Suzuki M, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9:465–476

100. Lister R, Pelizzola M, Dowen R, Hawkins R, Hon G, Tonti-Filippini J, Nery J, Lee L, Ye Z, Ngo Q-M, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar A, Thomson J, Ren B, Ecker J (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462:315–322
101. Law JA, Jacobsen S (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11:204–220
102. Takizawa T, Meaburn K, Misteli T (2008) The meaning of gene positioning. Cell 135:9–13
103. Misteli T (2007) Beyond the sequence: cellular organization of genome function. Cell 128:787–800
104. Kaiser T, Intine R, Dundr M (2008) De Novo formation of a subnuclear body. Science 322:1713–1717
105. Soutoglou E, Misteli T (2008) Activation of the cellular DNA damage response in the absence of DNA lesions. Science 320:1507–1510
106. Rajapakse I, Perlman M, Scalzo D, Kooperberg C, Groudine M, Kosak S (2009) The emergence of lineage-specific chromosomal topologies from coordinate gene regulation. Proc Natl Acad Sci U S A 106:6679–6684
107. Junier I, Martin O, Képès F (2010) Spatial and topological organization of DNA chains induced by gene co-localization. PLoS Comput Biol 6:e1000678
108. Cook P, Marenduzzo D (2009) Entropic organization of interphase chromosomes. J Cell Biol 186:825–834
109. Lanctôt C, Cheutin T, Cremer M, Cavalli G, Cremer T (2007) Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. Nat Rev Genet 8:104–115
110. Osborne C, Chakalova L, Brown K, Carter D, Horton A, Debrand E, Goyenechea B, Mitchell J, Lopes S, Reik W, Fraser P (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet 36:1065–1071
111. Osborne C, Chakalova L, Mitchell J, Horton A, Wood A, Bolland D, Corcoran A, Fraser P (2007) Myc dynamically and preferentially relocates to a transcription factory occupied by Igh. PLoS Biol 5:e192
112. Schoenfelder S, Sexton T, Chakalova L, Cope N, Horton A, Andrews S, Kurukuti S, Mitchell J, Umlauf D, Dimitrova D, Eskiw C, Luo Y, Wei C-L, Ruan Y, Bieker J, Fraser P (2010) Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat Genet 42:53–62
113. Mitchell J, Fraser P (2010) Transcription factories are nuclear subcompartments that remain in the absence of transcription. Genes Dev 22:20–25
114. Sutherland H, Bickmore W (2009) Transcription factories: gene expression in unions? Nat Rev Genet 10:457–466
115. Cook P (2010) A model for all genomes: the role of transcription factories. J Mol Biol 395:1–10
116. Raj A, Peskin C, Tranchina D, Vargas D, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. PLoS Biol 4:e309
117. Chubb J, Trcek T, Shenoy S, Singer R (2006) Transcriptional pulsing of a developmental gene. Curr Biol 16:1018–1025
118. Batada N, Hurst L (2007) Evolution of chromosome organization driven by selection for reduced gene expression noise. Nat Genet 39:945–949
119. Hurst L, Pál C, Lercher M (2008) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5:299–310
120. Janga S-C, Collado-Vides J, Babu M (2008) Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. Proc Natl Acad Sci U S A 105:15761–15766
121. Kosak S, Scalzo D, Alworth S, Li F, Palmer S, Enver T, Lee J, Groudine M (2007) Coordinate gene regulation during hematopoiesis is related to genomic organization. PLoS Biol 5:e309
122. Babu M, Janga S, de Santiago I, Pombo A (2008) Eukaryotic gene regulation in three dimensions and its impact on genome evolution. Curr Opin Genet Dev 18:1–12

123. Parada L, McQueen P, Misteli T (2004) Tissue-specific spatial organization of genomes. Genome Biol 5:R44
124. Meaburn K, Misteli T (2007) Chromosome territories. Nature 445:379–381
125. Dekker J (2008) Gene regulation in the third dimension. Science 319:1793–1794
126. Gheldof N, Smith E, Tabuchi T, Koch C, Dunham I, Stamatoyannopoulos J, Dekker J (2010) Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. Nucleic Acids Res 38:4325–4336
127. Ott C, Blackledge N, Kerschner J, Leir S-H, Crawford G, Cotton C, Harris A (2009) Intronic enhancers coordinate epithelial-specific looping of the active CFTR locus. Proc Natl Acad Sci U S A 106:19934–19939
128. Fraser J, Rousseau M, Shenker S, Ferraiuolo M, Hayashizaki Y, Blanchette M, Dostie J (2009) Chromatin conformation signatures of cellular differentiation. Genome Biol 10:R37
129. Lanzuolo C, Roure V, Dekker J, Bantignies F, Orlando V (2007) Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. Nat Cell Biol 9:1167–1174
130. Mishiro T, Ishihara K, Hino S, Tsutsumi S, Aburatani H, Shirahige K, Kinoshita Y, Nakao M (2009) Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. EMBO J 28:1234–1245
131. Ong C-T, Corces V (2009) Insulators as mediators of intra- and inter-chromosomal interactions: a common evolutionary theme. J Biol 8:73
132. Nikolaev L, Akopov S, Didych D, Sverdlov E (2009) Vertebrate protein CTCF and its multiple roles in a large-scale regulation of genome activity. Curr Genomics 10:294–302
133. Essien K, Vigneau S, Apreleva S, Singh L, Bartolomei M, Hannenhalli S (2009) CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. Genome Biol 10:R131
134. Hou C, Dale R, Dean A (2010) Cell type specificity of chromatin organization mediated by CTCF and cohesion. Proc Natl Acad Sci U S A 107:3651–3656
135. Ohlsson R, Lobanenkov V, Klenova E (2010) Does CTCF mediate between nuclear organization and gene expression? BioEssays 32:37–50
136. Cuddapah S, Jothi R, Schones D, Roh T-Y, Cui K, Zhao K (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res 19:24–32
137. Fu Y, Sinha M, Peterson C, Weng Z (2008) The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet 4:e1000138
138. Guelen L, Pagie L, Brasset E, Meuleman W, Faza M, Talhout W, Eussen B, de Klein A, Wessels L, de Laat W, van Steensel B (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453:948–952
139. Hore T, Deakin J, Marshall-Graves J (2008) The evolution of epigenetic regulators CTCF and BORIS/CTCFL in amniotes. PLoS Genet 4:e1000169
140. Bose T, Gerton J (2010) Cohesinopathies, gene expression and chromatin organization. J Cell Biol 189:201–210
141. Feeney K, Wasson C, Parish J (2010) Cohesin: a regulator of genome integrity and gene expression. Biochem J 428:147–161
142. McNairn A, Gerton J (2008) The chromosome glue gets a little stickier. Trends Genet 24:382–389
143. Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson H, Jarmuz A, Canzonetta C, Webster Z, Nesterova T, Cobb B, Yokomori K, Dillon N, Aragon L, Fisher A, Merkenschlager M (2008) Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell 132:422–433
144. Williams A, Flavell R (2008) The role of CTCF in regulating nuclear organization. J Exp Med 205:747–750
145. Kapranov P, Cheung J, Dike S, Nix D, Duttagupta R, Willingham A, Stadler P, Hertel J, Hackermüller J, Hofacker I, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G,

Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras T (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316:1484–1488
146. Seila A, Calbrese J, Levine S, Yeo G, Rahl P, Flynn R, Young R, Sharp P (2008) Divergent transcription from active promoters. Science 322:1849–1851
147. Core L, Waterfall J, Lis J (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322:1845–1848
148. Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen M, Mapendano C, Schierup M, Jensen T (2008) RNA exosome depletion reveals transcription upstream of active promoters. Science 322:1851–1854
149. De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi B, Muller H, Ragoussis J, Wei C-L, Natoli G (2010) A large fraction of extragenic RNA PolII transcription sites overlap enhancers. PLoS Biol 8:e1000384
150. Kim T-K, Hemberg M, Gray J, Costa A, Bear D, Wu J, Harmin D, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley P, Kreiman G, Greenberg M (2010) Widespread transcription at neuronal activity-regulated enhancers. Nature 465:182–187
151. Wilusz J, Sunwoo H, Spector D (2009) Long noncoding RNAs: functional surprises from the RNA world. Genes Dev 23:1494–1504
152. Mercer T, Dinger M, Mattick J (2009) Long non-coding RNAs: insights into functions. Nat Rev Genet 10:155–159
153. Feng J, Bi C, Clark B, Mady R, Shah P, Kohtz J (2006) The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes Dev 20:1470–1484
154. Wang X, Arai S, Song X, Reichart D, Du K, Pascual G, Tempst P, Rosenfeld M, Glass C, Kurokawa R (2008) Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. Nature 454:126–130
155. Mariner P, Walters R, Espinoza C, Drullinger L, Wagner S, Kugel J, Goodrich J (2008) Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. Mol Cell 29:499–509
156. Kondo T, Plaza S, Zanet J, Benrabah E, Valenti P, Hashimoto Y, Kobayashi S, Payre F, Kageyama Y (2010) Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. Science 329:336–339
157. Bernstein E, Allis C (2005) RNA meets chromatin. Genes Dev 19:1635–1655
158. Dinger M, Amaral P, Mercer T, Pang K, Bruce S, Gardiner B, Askarian-Amiri M, Ru K, Soldà G, Simons C, Sunkin S, Crowe M, Grimmond S, Perkins A, Mattick J (2008) Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome Res 18:1433–1445
159. Morris K, Santoso S, Turner A-M, Pastori C, Hawkins P (2008) Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. PLoS Genet 4:e1000258
160. Nagano T, Mitchell J, Sanz L, Pauler F, Ferguson-Smith A, Feil R, Fraser P (2008) The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science 322:1717–1720
161. Pandey R, Mondal T, Mohammad F, Enroth S, Redrup L, Komorowski J, Nagano T, Mancini-DiNardo D, Kanduri C (2008) Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell 32:232–246
162. Redrup L, Branco M, Perdeaux E, Krueger C, Lewis A, Santos F, Nagano T, Cobb B, Fraser P, Reik W (2009) The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. Development 136:525–530
163. Tufarelli C, Sloane-Stanley J, Garrick D, Sharpe D, Ayyub H, Wood W, Higgs D (2003) Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. Nat Genet 34:157–165

164. Faghihi M, Wahlestedt C (2009) Regulatory roles of natural antisense transcripts. Nat Rev Mol Cell Biol 10:637–643
165. Yu W, Gius D, Onyango P, Muldoon-Jacobs K, Karp J, Feinberg A, Cui H (2007) Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. Nature 451:202–206
166. Kanduri C (2008) Functional insights into long antisense noncoding RNA Kncq1ot1 mediated bidirectional silencing. RNA Biol 5:208–211
167. Ohhata T, Hoki Y, Sasaki H, Sado T (2008) Crucial role of antisense transcription across the Xist promoter in Tsix-mediated Xist chromatin modification. Development 135:227–235
168. Rinn J, Kertesz M, Wang J, Squazzo S, Xu X, Brugmann S, Goodnough L, Helms J, Farnham P, Segal E, Chang H (2007) Functional demarcation of active and silent chromatin domains in human Hox loci by noncoding RNAs. Cell 129:1311–1323
169. Matsui K, Nishizawa M, Ozaki T, Kimura T, Hashimoto I, Yamada M, Kaibori M, Kamiyama Y, Ito S, Okumura T (2008) Natural antisense transcript stabilizes inducible nitric oxide synthase messenger RNA in rat hepatocytes. Hepatology 47:686–697
170. Rossignol F, Vaché C, Clottes E (2002) Natural antisense transcripts of hypoxia-inducible factor 1alpha are detected in different normal and tumour human tissues. Gene 299:135–140
171. Faghihi M, Modarresi F, Khalil A, Wood D, Sahagan B, Morgan T, Finch C, St. Laurent G III, Kenny P, Wahlestedt C (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase. Nat Med 14:723–730
172. He Y, Vogelstein B, Velculescu V, Papadopoulos N, Kinzler K (2008) The antisense transcriptomes of human cells. Science 322:1855–1857
173. Hastings M, Milcarek C, Martincic K, Peterson M, Munroe S (1997) Expression of the thyroid hormone receptor gene, erbAα, in B lymphocytes: alternative mRNA processing is independent of differentiation but correlates with antisense RNA levels. Nucleic Acids Res 25:4296–4300
174. Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, Surani M, Sakaki Y, Sasaki H (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature 453:539–544
175. Faghihi M, Wahlestedt C (2006) RNA interference is not involved in natural antisense mediated regulation of gene expression in mammals. Genome Biol 7:R38
176. Peters N, Rohrbach J, Zalewski B, Byrkett C, Vaughn J (2003) RNA editing and regulation of Drosophila 4f-rnp expression by sas-10 antisense readthrough mRNA transcripts. RNA 9:698–710
177. Osato N, Suzuki Y, Ikeo K, Gojobori T (2007) Transcriptional interferences in cis natural antisense transcripts of human and mice. Genetics 176:1299–1306
178. Schmitt S, Prestel M, Paro R (2005) Intergenic transcription through a polycomb group response element counteracts silencing. Genes Dev 19:697–708
179. Sanchez-Elsner T, Gou D, Kremmer E, Sauer F (2006) Noncoding RNAs of trithorax response elements recruit Drosophila Ash1 to Ultrabithorax. Science 311:1118–1123
180. Chen K, Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet 8:93–103
181. Yu Z, Jian Z, Shen S, Purisima E, Wang E (2007) Global analysis of microRNA target gene expression reveals that miRNA targets are lower expressed in mature mouse and Drosophila tissues than in the embryos. Nucleic Acids Res 35:152–164
182. Kim V, Han J, Siomi M (2008) Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol 10:126–139
183. Kosik K (2009) MicroRNAs tell an evo-devo story. Nat Rev Neurosci 10:754–759
184. Christodoulou F, Raible F, Tomer R, Simakov O, Trachana K, Klaus S, Snyman H, Hannon G, Bork P, Arendt D (2010) Ancient animal microRNAs and the evolution of tissue identity. Nature 463:1084–1088
185. Heimberg A, Sempere L, Moy V, Donoghue P, Peterson K (2008) MicroRNAs and the advent of vertebrate morphological complexity. Proc Natl Acad Sci U S A 105:2946–2950
186. Liu N, Okamura K, Tyler D, Phillips M, Chung W-J, Lai E (2008) The evolution and functional diversification of animal microRNA genes. Cell Res 18:985–996

187. Wheeler B, Heimberg A, Moy V, Sperling E, Holstein T, Heber S, Peterson K (2009) The deep evolution of metazoan microRNAs. Evol Dev 11:50–68
188. Yu X, Lin J, Zack D, Mendell J, Qian J (2008) Analysis of regulatory network topology reveals functionally distinct classes of microRNAs. Nucleic Acids Res 36:6494–6503
189. Couzin J (2008) MicroRNAs make big impression in disease after disease. Science 319:1782–1784
190. Hébert S, de Strooper B (2009) Alterations of the microRNA network cause neurodegenerative disease. Trends Neurosci 32:199–206
191. Passetti F, Ferreira C, Costa F (2009) The impact of microRNAs and alternative splicing in pharmacogenomics. Pharmacogenomics J 9:1–13
192. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovski D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298:824–827
193. Johnston R, Chang S, Etchberger J, Ortiz C, Hobert O (2005) MicroRNAs acting in a double-negative feedback loop to control a neuronal cell fate decision. Proc Natl Acad Sci U S A 102:12449–12454
194. Martinez N, Ow M, Barrasa M, Hammell M, Sequerra R, Doucette-Stamm L, Roth F, Ambros V, Walhout A (2008) A C. elegans genome-scale microRNA network contains composite feedback motifs with high flux capacity. Genes Dev 22:2535–2549
195. Hobert O (2008) Gene regulation by transcription factors and microRNAs. Science 319:1785–1786
196. Cohen S, Brennecke J, Stark A (2006) Denoising feedback loops by thresholding – a new role for microRNAs. Genes Dev 20:2769–2772
197. To T, Maheshri N (2010) Noise can induce bimodality in positive transcriptional feedback loops without bistability. Science 327:1142–1145
198. Hornstein E, Shomron N (2006) Canalization of development by microRNAs. Nat Genet Suppl 38:S20–S24
199. Melton C, Judson R, Blelloch R (2010) Opposing microRNA families regulate self-renewal in mouse embryonic stem cells. Nature 463:621–626
200. Muddashetty R, Bassell G (2009) A boost in microRNAs shapes up the neuron. EMBO J 28:617–618
201. Schratt G, Tuebing F, Nigh E, Kane C, Sabatini M, Kiebler M, Greenberg M (2006) A brain-specific microRNA regulates dendritic spine development. Nature 439:283–289
202. Schratt G (2009) microRNAs at the synapse. Nat Rev Neurosci 10:842–849
203. Brion P, Westhof E (1997) Hierarchy and dynamics of RNA folding. Annu Rev Biophys Biomol Struct 26:113–137
204. Lopez de Silanes I, Zhan M, Lal A, Yang X, Gorospe M (2004) Identification of a target RNA motif for RNA-binding protein HuR. Proc Natl Acad Sci U S A 101:2987–2992
205. Long D, Lee R, Williams P, Chan C, Ambros V, Ding Y (2007) Potent effect of target structure on microRNA function. Nat Struct Mol Biol 14:287–294
206. Gorodkin J, Hofacker IL, Torarinsson E, Yao Z, Havgaard JH, Ruzzo WL (2010) De novo prediction of structured RNAs from genomic sequences. Trends Biotechnol 28(1):9–19
207. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406–3415
208. Knudsen B, Hein J (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics 15:446–454
209. Lombark G, Bensi D, Fernandez-Zapico M, Urrutia R (2006) Evidence for the existence of an HP1-mediated subcode within the histone code. Nat Cell Biol 8:407–415
210. Tay Y, Zhang J, Thomson A, Lim B, Rigoutsos I (2008) MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. Nature 455:1124–1128
211. Tan S, Guo J, Huang Q, Chen X, Li-Ling J, Li Q, Ma F (2007) Retained introns increase putative microRNA targets within 3′ UTRs of human mRNA. FEBS Lett 581:1081–1086
212. Shomron N, Levy C (2009) MicroRNA-biogenesis and pre-mRNA splicing crosstalk. J Biomed Biotechnol 2009:594678

213. Kedde M, Strasser M, Boldajipour B, Oude Vrielink J, Slanchev K, le Sage C, Nagel R, Voorhoeve P, van Duijse J, Ørom U, Lund A, Perrakis A, Raz E, Agami R (2007) RNA-binding protein Dnd1 inhibits microRNA access to target mRNA. Cell 131:1273–1286
214. Nishikura K (2006) Editor meets silencer: crosstalk between RNA editing and RNA interference. Nat Rev Mol Cell Biol 7:919–931
215. Faulkner G, Kimura Y, Daub C, Wani S, Plessy C, Irvine K, Schroder K, Cloonan N, Steptoe A, Lassmann T, Waki K, Hornig N, Arakawa T, Takahashi H, Kawai J, Forrest A, Suzuki H, Hayashizaki Y, Hume D, Orlando V, Grimmond S, Carninci P (2009) The regulated retrotransposon transcriptome of mammalian cells. Nat Genet 41:563–571
216. Bell G, Hey T, Szalay A (2009) Beyond the data deluge. Science 323:1297–1298
217. Cochrane G, Akhtar R, Bonfield J, Bower L, Demiralp F, Faruque N, Gibson R, Hoad G, Hubbard T, Hunter C, Jang M, Juhos S, Leinonen R, Leonard S, Lin Q, Lopez R, Lorenc D, McWilliam H, Mukherjee G, Plaister S, Radhakroshnan R, Robinson S, Sonhany S, Hoopen P, Vaughan R, Zalunin V, Birney E (2009) Petabyte-scale innovations at the European Nucleotide Archive. Nucleic Acids Res 37(Database Issue):D19–D25
218. Shumway M, Cochrane G, Sugawara H (2010) Archiving next generation sequencing data. Nucleic Acids Res 38(Database Issue):D870–D871
219. Sangket U, Phongdarra A, Chotigeat W, Nathan D, Kim WY, Bhak J, Ngamphiw C, Tongsima S, Khan A, Lin H, Tan T (2008) Automatic synchronization and distribution of biological databases and software over low-bandwidth networks among developing countries. Bioinformatics 24:299–301
220. Richter B, Sexton P (2009) Managing and analyzing next-generation sequence data. PLoS Comput Biol 5:e1000369
221. Bateman A, Wood M (2009) Cloud computing. Bioinformatics 25:1475
222. Brooksbank C, Cameron G, Thornton J (2010) The European Bioinformatics Institute's data resources. Nucleic Acids Res 38(Database Issue):D17–D25
223. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389(6648):251–260