Overlapping transcription initiation codes and promoter interpretation in vertebrate development and differentiation

Vanja Haberle

Dissertation for the degree philosophiae doctor (PhD)
at the University of Bergen

2015

Dissertation date: 05/01/2015

Scientific environment

This work has been performed in

Computational Biology Unit (CBU), Uni Computing

uni Computing

Sars International Centre for Marine Molecular Biology

and

Imperial College London and MRC Clinical Sciences Centre

Imperial College London

MRC Clinical Sciences Centre

Vanja Haberle is affiliated with

Department of Biology

UNIVERSITAS BERGENSIS

Acknowledgements

Throughout the course of my doctoral studies I have been involved in many collaborations and had the chance to work with many great people. I would like to thank everyone for creating a stimulating and pleasant working environment and for supporting my work during the past four years.

I am truly grateful to my supervisor, Boris Lenhard, for taking me as a PhD student and giving me the opportunity to work on many challenging and leading edge projects. Thank you for all the creative ideas and scientific guidance, for constantly supporting my work and encouraging me to share my results with the scientific community, and for being a demanding, but at the same time an understanding supervisor.

I am also thankful to Piero Carninci for taking the responsibility as my co-supervisor, for believing in my work and encouraging me to actively participate in the FANTOM5 project.

I would like to thank my collaborators, Ferenc Müller and his group members, Yavor and Nan, for pleasant communication between the "wet" and "dry" lab and for fruitful collaboration. I am also thankful to FANTOM consortium, especially Alistair Forrest and Yoshihide Hayashizaki, for providing the opportunity to share and discuss ideas during conferences in Yokohama and for coordinating the work for the main FANTOM5 publication. I thank all other collaborators for a chance to work on various projects, broadening my scientific interests and improving my communication skills.

During the last four years I was lucky enough to be a part of the research groups both at University of Bergen and at MRC Clinical Sciences Centre in London. Many thanks to Supat, Chirag, Christopher, Jan Christian, Sara, Altuna, Xianjun, Vedran, Gemma, Yogita, Reidar and Chandu for welcoming me to the group in Bergen and creating a pleasant working environment, for all the scientific advice and fun times after work. Thanks to all group members in London, Nathan, Anja, Liz, Ge, Piotr, Sam, Malcolm, Slavica and Amit for stimulating discussions, enjoyable lunch breaks and numerous pub nights. Special thanks to Liz and Nathan for proofreading the thesis. I would also like to thank all SARS members, for having me as a guest in their office during my last year and for involving me in all their social activities. Thanks to my officemate Wei for putting up with my habits and for valuable ping pong lessons, to all ping pong teammates and to movie club members.

I would like to thank my family, my parents and sister, for unconditional love and support in all my decisions. Without you guiding me along the right path from the beginnings, I would have never achieved this.

Finally, I thank Petar, my love and my friend, for endless patience, continuous support and infinite Skype calls. Thank you for always being there and for constantly reminding me about beautiful things in life outside of science.

Abstract

A core promoter is a minimal region sufficient to direct the accurate initiation of transcription. Various core promoter elements have been discovered that recruit and position transcriptional machinery, which then initiates transcription at individual transcription start sites (TSS); however, no universal promoter code has been established. The methods and results presented in this thesis focus on innovative analysis of precise transcription initiation data to reveal sequence and chromatin features underlying core promoters and their dynamic usage in development and differentiation.

Cap analysis of gene expression (CAGE) provides a single base-pair resolution map of TSSs and their relative usage, and it is a powerful tool for studying promoter structure and function. It has led to the discovery of major promoter classes that differ in transcription initiation patterns: “sharp” promoters in which the majority of transcription starts at one clearly dominant TSS, and “broad” promoters with multiple equally used TSS positions distributed along a wider region. By applying CAGE to a developmental time-course of zebrafish (*Danio rerio*) we created a first comprehensive map of transcription initiation during vertebrate embryogenesis and revealed widespread dynamics in promoter usage at all levels, from alternative promoters to individual TSSs. We found that thousands of promoters are utilized differently by the oocyte and the embryo, uncovering two independent codes that drive dynamic changes in TSS usage and promoter shape. Maternal TSS selection is guided by an A/T-rich W-box motif positioned at a fixed spacing from the TSS producing a sharp promoter architecture, whereas zygotic selection is restricted by the position of the first downstream nucleosome and produces broad promoter architecture with the dominant TSS aligned to inter- and intranucleosomal sequence positioning signals. The two grammars co-exist in close proximity or in physical overlap at promoters genome-wide.

We further showed that a tight association between dominant TSS in broad promoters and nucleosome positioning exists in human and mouse transcription.

Alignment of the intranucleosomal dinucleotide frequency patterns downstream of the TSS revealed that a well-positioned +1 nucleosome is a key determinant of TSS preference in broad promoters. Its presence in both zebrafish and mammals suggests the evolutionary conservation of the underlying nucleosome-associated TSS selection mechanism.

Precise TSS localisation is crucial for promoter-centred analyses of any genome-wide data. To facilitate the reuse of high-resolution and context-specific TSSs derived from a growing resource of CAGE data, we developed *CAGEr*, an R/Bioconductor software package for promoterome mining. *CAGEr* provides easy access to the majority of published CAGE datasets and presents a comprehensive workflow for processing, visualisation and analysis of precise promoter data, and allows its integration with other genome data types.

Taken together, the work presented in this thesis reveals unexpected dynamics in core promoter usage at TSS level and demonstrates that promoter type is not an inherent property of the genomic locus, but is rather dependent on the regulatory context. The existence of overlapping transcription initiation codes has important implications for future analyses of promoter content and function.

List of publications included in the thesis

The thesis includes the following papers, which will be referred to in the text using their roman numerals:

I. Nepal C, Hadzhiev Y, Previti C, Haberle V, Li N, Takahashi H, Suzuki AMM, Sheng Y, Abdelhamid RF, Anand S, Gehrig J, Akalin A, Kockx CEM, van der Sloot AAJ, van IJcken WFJ, Armant O, Rastegar S, Watson C, Strahle U, Stupka E, Carninci P, Lenhard B, Muller F: **Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis**. *Genome Research* 2013, **23**:1938–1950.

II. Haberle V, Li N, Hadzhiev Y, Plessy C, Previti C, Nepal C, Gehrig J, Dong X, Akalin A, Suzuki AM, van IJcken WFJ, Armant O, Ferg M, Strähle U, Carninci P, Müller F, Lenhard B: **Two independent transcription initiation codes overlap on vertebrate core promoters**. *Nature* 2014, **507**:381–385.

III. The FANTOM Consortium: Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jorgensen M, Dimont E, Arner E, Schmidl C, Schaefer U, Medvedeva YA, Plessy C, Vitezic M, Severin J, Semple CA, Ishizu Y, Young RS, Francescatto M, et al.: **A promoter-level mammalian expression atlas**. *Nature* 2014, **507**:462–470.

IV. Haberle V, Forrest ARR, Hayashizaki Y, Carninci P, Lenhard B: **CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses**. *Manuscript submitted* (September 2014).

List of other publications

During the course of my doctoral studies I have also contributed to the following publications:

- Hughes Carvalho R, Haberle V, Hou J, van Gent T, Thongjuea S, van Ijcken W, Kockx C, Brouwer R, Rijkers E, Sieuwerts A, Foekens J, van Vroonhoven M, Aerts J, Grosveld F, Lenhard B, Philipsen S: **Genome-wide DNA methylation profiling of non-small cell lung carcinomas**. *Epigenetics & Chromatin* 2012, **5**:9.

- Haberle V, Lenhard B: **Dissecting genomic regulatory elements in vivo**. *Nature Biotechnology* 2012, **30**:504–506.

- Hughes Carvalho R, Hou J, Haberle V, Aerts J, Grosveld F, Lenhard B, Philipsen S: **Genomewide DNA Methylation Analysis Identifies Novel Methylated Genes in Non–Small-Cell Lung Carcinomas**. *Journal of Thoracic Oncology* 2013, **8**:562–573.

- Frangini A, Sjöberg M, Roman-Trufero M, Dharmalingam G, Haberle V, Bartke T, Lenhard B, Malumbres M, Vidal M, Dillon N**: The Aurora B Kinase and the Polycomb Protein Ring1B Combine to Regulate Active Promoters in Quiescent Lymphocytes**. *Molecular Cell* 2013, **51**:647–661.

Table of contents

List of abbreviations

bp	Base-pairs
BRE	TFIIB recognition element
CAGE	Cap analysis of gene expression
cDNA	Complementary DNA
CGI	CpG island
ChIP	Chromatin immunoprecipitation
CRM	*Cis*-regulatory module
CTSS	CAGE-detected transcription start site
DCE	Downstream core element
DNA	Deoxyribonucleic acid
DPE	Downstream promoter element
ESC	Embryonic stem cell
FANTOM	Functional annotation of mammalian genome
GTF	General transcription factor
H3K27ac	Acetylation of lysine 27 on histone 3
H3K27me3	Tri-methylation of lysine 27 on histone 3
H3K36me3	Tri-methylation of lysine 36 on histone 3
H3K4me3	Tri-methylation of lysine 4 on histone 3
H3K9me3	Tri-methylation of lysine 9 on histone 3
H4K20me3	Tri-methylation of lysine 20 on histone 4
HCNE	Highly conserved non-coding element
hpf	Hours past fertilisation
Inr	Initiator
lncRNA	Long non-coding RNA
MBT	Mid-blastula transition
mRNA	Messenger RNA
MTE	Motif ten element
ncRNA	Non-coding RNA
NFR	Nucleosome-free region

PcG	Polycomb group proteins
PET	Paired-end ditag technology
PGC	Primordial germ cell
PIC	Pre-initiation complex
PRC	Polycomb-repressive complex
RNA	Ribonucleic acid
RNAPII	RNA polymerase II
rRNA	Ribosomal RNA
SAGE	Serial analysis of gene expression
TAF	TBP-associated factor
TBP	TATA-box binding protein
TC	Tag cluster / transcriptional cluster
TRF	TBP-related factor
TF	Transcription factor
TFBS	Transcription factor binding site
tRNA	Transfer RNA
TSS	Transcription start site

1 Introduction

1.1 Transcriptional regulation of gene expression

1.1.1 DNA, genes and the transmission of genetic information

Living organisms encode the instructions for the development of their body plan and interaction with the environment in deoxyribonucleic acid (DNA), a double-stranded polymeric molecule that consists of four kinds of nitrogenous bases sequentially ordered on a sugar-phosphate backbone. The total DNA sequence of an organism is referred to as its genome. In eukaryotic organisms, each cell contains in its nucleus a copy of the genome, with individual DNA molecules wrapped around histone proteins and organized into chromosomes. The structure and organisation of DNA enables efficient storage, replication and transmission of the information for creating an entire multicellular organism from a single cell. Discrete segments of the genome that encode for functional products are known as genes. They serve as templates for production of ribonucleic acid (RNA) molecules, many of which act as messengers transmitting information for the production of proteins, the principal functional entities in the cell. However, RNAs themselves can be final products performing structural, catalytic or regulatory functions.

1.1.2 Genes and gene expression

In the broadest sense, a gene is a region of the genome that encodes for a functional protein or RNA molecule [1]. In both cases the DNA sequence information is first converted into RNA in the process known as transcription [2]. If the gene is protein-coding, the transcribed RNA is called messenger RNA (mRNA) [3], and is further converted in the process of translation into a sequence of amino acids, which folds into a functional protein. Non-protein-coding genes give rise to non-coding RNAs (ncRNA), which are never translated, but carry out various functions in the cell. These include ribosomal RNAs (rRNA), the structural components of ribosomes believed to catalyse mRNA translation [4], transport RNAs (tRNA), which serve as adaptor molecules carrying amino acids and specifying which sequence in the mRNA corresponds to which amino acid during translation [5, 6], and finally various classes of ncRNAs with regulatory functions, such as micro RNAs [7, 8], short interfering RNAs [9] and long non-coding RNAs (lncRNA) [10-12].

The entire process of converting sequence information encoded within a gene into a precise amount of functional product is referred to as gene expression. This process is influenced by both internal and external stimuli and is tightly regulated by various mechanisms, acting at different levels from transcriptional to post-translational control, to ensure the correct amount of gene product is present in a particular cell at a particular point in time.

1.1.3 Transcriptional machinery and core promoters

Protein-coding genes and several classes of ncRNA genes in eukaryotes are transcribed by RNA polymerase II (RNAPII), a large multi-subunit enzyme that uses DNA as a template to produce complementary RNA molecule [13, 14]. RNAPII initiates transcription at individual nucleotides at the beginning of the gene called transcription start sites (TSS). The region surrounding a TSS is known as the core promoter and it is defined as a minimal region that is sufficient to direct the accurate initiation of transcription. A eukaryotic core promoter typically extends ~40 bp upstream and downstream of the TSS, and it is a place of the assembly of the transcriptional machinery [15]. This process requires general transcription factors (GTFs), which recognize and bind core promoter elements and recruit RNAPII. There are six general transcription factors: TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, which assemble on the core promoter in a stepwise manner and form the pre-initiation complex (PIC) [13, 16]. TFIID plays a central role in recognising and binding specific core promoter elements and creates an environment that facilitates transcription initiation [17]. Various core promoter elements have been identified in eukaryotic promoters and include a TATA-box, an Initiator (Inr), a Downstream Promoter Element (DPE), a Downstream Core Element (DCE), a TFIIB-Recognition

Element (BRE), and a Motif Ten Element (MTE) [18]. However, none of these elements are universal, since they are found only in a fraction of core promoters in various combinations and there are many promoters that lack any of these elements [19]. In addition, some core promoter elements are associated with specific biological functions, for instance the TCT motif, which is found exclusively in promoters of genes that encode the components of translational machinery [20].

Many core promoters in vertebrates overlap with CpG islands (CGI), which are genomic regions characterised by elevated C+G content and frequency of CG dinucleotides compared to the bulk genome [21, 22]. The current estimate is that ~70% of human promoters are associated with a CGI [23], with similar percentages observed for mouse and chicken [24]. The proportion of CGI-overlapping promoters seems substantially lower in amphibians and fish [24]. However, this is likely due to the fact that the definition of the CpG island relies on arbitrary thresholds set upon C+G content, observed over expected ratio of CpG dinucleotide counts and region length [25], which have been optimised for mammalian genomes and do not perform well for genomes with very different nucleotide composition. Nevertheless, association with CpG islands distinguishes two main classes of vertebrate promoters, high-CpG promoters and low-CpG promoters [23, 26], which are additionally characterised by distinct promoter features and functions of associated genes [26].

The complexity of the core promoter is further seen in the relation among specificity of expression, transcription initiation patterns, motif composition and the organization of the chromatin structure in the promoter region, as discussed further below. All this suggests that core promoters are not passive elements that serve only to direct the proper placement of the RNA polymerase II transcriptional machinery. They receive and integrate various regulatory inputs and convert them into the rate of transcription initiation. Core promoter elements can determine the responsiveness of the promoter to transcriptional regulation by *cis*-regulatory elements and *trans*-acting factors [27], and are hence central, active components of transcriptional regulation.

1.1.4 *Cis*-regulatory elements and *trans*-acting factors

The formation of the PIC and recruitment of RNAPII can direct only low levels of transcription, known as basal transcription [28]. In contrast, gene expression is characterised by a high dynamic range and is often context-specific. This is achieved by modulating the basal level of transcription by the action of *cis*-regulatory elements and *trans*-acting factors.

Transcriptional activity is greatly stimulated by factors known as activators, which are sequence-specific DNA-binding proteins that recognise and bind sites often located upstream of the core promoter [29]. There are many classes of activators characterised by different DNA-binding domains, each associating with their own class of specific DNA sequences [30]. They interact with components of the basal transcription machinery via their activation domain and stimulate PIC assembly [29]. Activators also function by recruiting transcriptional co-activators, a diverse class of non-DNA-binding factors, which act either directly or indirectly to regulate the activity of the RNAPII transcriptional machinery [31]. Transcriptional co-activators can serve as bridging molecules between activators and GTFs directly enhancing activator-facilitated entry of RNAPII to the PIC [13, 31]. On the other hand, the chromatin-related co-regulators affect transcription indirectly by remodelling nucleosomes or by covalent modifications of histones, creating a chromatin environment that facilitates GTF binding [32].

Transcription can also be inhibited by various mechanisms. *Trans*-acting repressors bind directly or indirectly to DNA and negatively regulate transcription by interacting with the components of the basal transcription machinery, by blocking transcriptional activation mediated by activators or by disrupting communication between promoters and distal regulatory elements [33]. They can also act by directly influencing chromatin structure or by recruiting chromatin-remodelling co-repressor complexes to establish repressive environment at specific loci [34]. In addition, lncRNAs have also been shown to act in *trans* to repress transcription from specific loci [35] and to mediate X chromosome inactivation in mammals [36].

Both transcriptional activators and repressors bind to *cis*-regulatory elements, which can be located proximally to, or at a distance from, the core promoter. These include proximal promoters, locus control regions (LCRs), enhancers, silencers and insulators (Figure 1), which all harbour specific sequence motifs known as transcription factor binding sites (TFBSs).

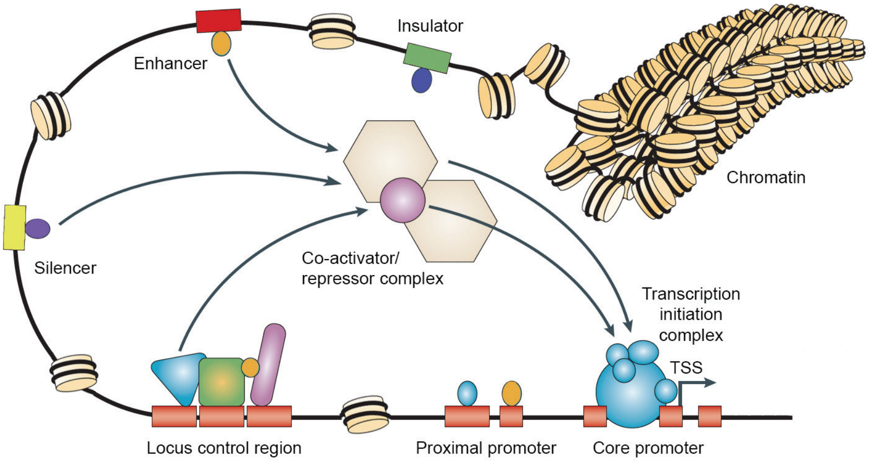

Figure 1. *Cis*-regulatory elements in eukaryotic genome. Typical localisation of each type of regulatory element is shown relative to TSS (arrow). Various transcription factors bind to DNA binding sites within proximal and distal elements and regulate the activity of transcriptional machinery through interactions with co-regulatory complexes. The figure is adapted from [37].

The proximal promoter is a region located immediately upstream of the core promoter and typically spans up to 250 bp upstream of the TSS [38]. It contains sequence motifs for binding of transcriptional activators and repressors, which are often organised into *cis*-regulatory modules (CRMs) [39]. Activating TFBSs tend to be located closer to TSS than repressing TFBSs [40, 41], and at least some of them seem to be positionally constrained with respect to the TSS [42] or to each other [43]. The combinatorial effect of transcription factor binding allows the proximal promoter to integrate the activity of multiple TFs and mediate context-specific gene expression [44]. Genes that contain proximal promoter motifs in a position-specific or distance-specific manner are often related, both in function and/or in expression pattern [43].

Enhancers and silencers are regulatory sequences located further away at a variable distance from the promoter; the distance can range from several hundred bp to one megabase [45]. Unlike the proximal promoter, which is positionally restricted with respect to the core promoter and TSS, these distal regulatory elements can be found upstream, within or downstream of the target gene [45, 46] and their activity seems to be independent of their orientation [47]. Enhancers have been identified by their ability to drive expression from a minimal promoter in transgenic assays [47-49]. They activate transcription, often in a temporally and spatially restricted manner, driving specific expression patterns during development and differentiation. Enhancer activity is mediated by interactions between sequence-specific DNA-binding proteins and sequence elements, which tend to cluster within the enhancer region forming CRMs [50]. In contrast, silencers are negative regulatory elements composed of binding sites for various factors that act collectively to establish a repressive higher-order chromatin structure at distal target promoters [51].

DNA elements that restrict long-range interactions between neighbouring genome domains are called insulators. They can be located between the distant regulatory element, such as an enhancer or silencer, and the target promoter, where they act by disrupting their communication and preventing the promoter from receiving regulatory input [52]. This direction-dependent enhancer-blocking activity requires binding of the ubiquitously expressed and highly conserved DNA-binding protein CTCF [53]. Another type of insulator acts through the formation of a barrier that prevents the spread of heterochromatin, thus restricting repressive chromatin environment to specific loci [52]. The demarcation of active and repressive domains is also dependent on CTCF binding [54], further confirming the tight relationship between insulator activity and CTCF.

Locus control regions (LCR) are clusters of *cis*-regulatory elements involved in transcriptional regulation of a specific genomic locus containing one or a set of related genes. An LCR can comprise various distal regulatory elements including enhancers, silencers and insulators, and is composed of arrays of multiple ubiquitous and lineage-specific TFBSs, which mediate tissue-specific expression of linked genes [55]. Studies on the well-characterised human β–globin LCR demonstrated that its activity is position-independent but orientation-dependent [56], distinguishing LCRs from simple enhancers.

1.1.5 Enhancers and long-range gene regulation

Most of the regulatory content of a metazoan genome lies outside of proximal promoters [57] and tends to be contained within enhancers, which seem to be a predominant type of functional elements in the non-coding portion of the genome. They are characterised by clusters of binding sites for many different TFs and chromatin regulators [49, 58]. The level of restriction on the arrangement of TFBSs distinguishes different types of enhancer architectures [50]. The enhanceosome is characterised by extensive overlap of individual TFBSs creating a composite element that operates as a single unit of regulation [59]. Cooperative interactions, both between neighbouring TFs and the bound chromatin, are essential for the activity of such enhancers. When an appropriate occupancy of TFBSs is achieved, recruitment of transcriptional co-activators and chromatin-remodelling proteins occurs and the formed complex promotes promoter-mediated gene activation [60]. Such enhancers seem to receive inputs from multiple activators and repressors and resolve them into a single output, thus operating as on/off switches for transcriptional activation [61]. On the other hand, the billboard model of enhancer function suggests independent recruitment of TFs, which does not require strict spacing and orientation of TFBSs within the enhancer [62]. This arrangement allows the enhancer to display both the active and repressed states, which are then interpreted by the transcriptional machinery at the target promoter through multiple interactions with the enhancer [62].

Transcriptional activation by enhancers is temporally and spatially restricted and can produce highly specific expression patterns during development. Many genes involved in development and establishment of the metazoan body plan are regulated by complex arrays of enhancers, each driving distinct aspects of the final expression pattern [63]. Mutations in distal-acting enhancers were shown to cause serious developmental defects [45], implicating the importance of long-range regulation in human disease [64].

Although it is generally thought that enhancers are located outside of the gene promoter region, a recent study has shown that regulatory elements within a promoter of one gene can act as enhancers to activate transcription from a remote promoter through long-range regulation [65]. Considering the fact that enhancers do not necessarily act on the closest promoter but can bypass neighbouring genes to regulate genes located more distantly along a chromosome, this further increases the complexity of the distal regulatory interactions within the genome.

Many non-coding sequences that are highly conserved between different vertebrate and mammalian species were found to be enriched for enhancers [49]. These highly conserved non-coding elements (HCNEs) are non-randomly distributed through the genome and tend to cluster around developmental regulator genes [66, 67] suggesting their involvement in complex regulation of those genes. Their functional relevance is further corroborated by the constraints imposed on the organisation and evolution of the genome, which seem to keep those arrays of regulatory elements in synteny with their respective target [68]. The genomic-regulatory block (GRB) model was proposed to explain such arrangements in the genome, where a single target gene is flanked by HCNEs scattered around the locus, which is often a gene desert or sometimes contains other genes not responsive to regulation by HCNEs [69]. In many cases these bystander genes are also kept in synteny because regulatory elements important for regulation of the target gene are embedded within their introns [69, 70] or even overlap their functional parts [71].

In addition to HCNEs, some protein-coding exons have also been shown to act as enhancers for neighbouring genes [72]. Although evolutionary sequence conservation has proved to be useful for the identification of enhancers [49, 73], there are also functional enhancers that do not seem to be conserved at sequence level [74, 75], and some of them have been shown to drive similar expression patterns in different species, suggesting functional conservation without sequence conservation [76].

Despite the advances in detecting active enhancers genome-wide [77, 78] and dissecting their regulatory content [79-81], there are still fundamental questions that need be addressed. How do enhancers work across such large distances and how is the specificity between an enhancer and its target promoter achieved? Several models have been proposed to describe how enhancers may communicate with their target gene promoter [82]. Currently the most plausible model supported by both theoretical [83] and experimental [84, 85] observations is the "looping" model in which the remote enhancer "loops out" the intervening DNA to reach the target promoter. It was shown that the formation of these chromatin loops depends on sequence-specific TFs bound to the enhancer and the promoter [85]. Furthermore, it appears that the enhancer loops form prior to gene activation and stably associate with paused RNAPII at promoters, keeping this loop topology ready for rapid activation of transcription by recruitment of additional factors [86]. The formation of chromatin loops brings the enhancer and its target promoter into close physical proximity in the nucleus and this feature is utilised by chromatin conformation capture experimental approaches [87] to detect long-range interactions genome-wide [88] and to identify target promoters of specific regulatory elements. However, the knowledge about the specificity of promoter-enhancer interactions is still very limited. There is evidence that the features of the target promoter determine its responsiveness to distal regulatory elements within accessible chromosomal domain. For instance, it was shown that the presence of specific core promoter elements in *Drosophila* makes promoters responsive to distinct enhancers [89]. Furthermore, the target genes of GRBs in vertebrates were shown to differ in various sequence, chromatin and transcriptional promoter features from non-responsive bystander genes, which likely specify them as a target of regulation by surrounding HCNEs [90]. These observations highlight the important functional role of the core promoter as an active participant in the long-range gene regulation.

1.1.6 Chromatin structure and epigenetic modifications

Genetic information is encoded in DNA in a linear fashion. However, to enable efficient storage, organisation and control of the large amount of DNA within the nucleus, the linear DNA molecules are coupled with histone and other non-histone proteins into a macromolecular complex known as chromatin. Two copies of each of the core histones H2A, H2B, H3 and H4 assemble into a histone octamer, which is then wound by approximately 147 bp of DNA forming a nucleosome [91]. Nucleosomes are arranged as a linear array along the DNA polymer creating a "beads on a string" structure. The packaging of DNA creates both a problem and an opportunity, since wrapping of DNA around histones potentially obstructs access to functional elements in DNA. However, the ubiquity of nucleosomes at all regions of chromosomal DNA can be exploited to direct the enzymes that read, replicate and repair DNA to the appropriate entry sites.

Nucleosome positioning was most extensively studied in the compact yeast genome, and the first genome-wide mapping of nucleosome positions at high resolution showed that the nucleosomes at most genes are generally organized in the same way [92]. Around the beginning of a gene there is a nucleosome-free region (NFR) flanked by two well-positioned nucleosomes (the −1 and +1 nucleosomes), which is followed by an array of nucleosomes that package the gene body (Figure 2). The first, +1 nucleosome, displays the tightest positioning and is subject to various histone protein variants and modifications, implicating its involvement in regulation of gene transcription. Further downstream nucleosomes exhibit lower levels of phasing. This basic pattern was later shown to be present in metazoan genomes as well [93-95].

In contrast, the vast majority of nucleosomes throughout the rest of the genome seem to be statistically positioned and form arrays of phased nucleosomes mostly around barriers imposed by DNA binding proteins or the minority of well-positioned nucleosomes [95-98]. Despite controversy around the degree to which primary sequence determines nucleosome positioning *in vivo* [99-102], it is clear that nucleosomes have certain sequence preference for their positioning. The region occupied by the centre of the nucleosome both *in vivo* and *in vitro* was shown to exhibit a significant increase in G/C usage, whereas A/T usage increases towards the nucleosome flanking regions [97]. Elements with such nucleotide composition were proposed to act as "container" sites able to produce a strongly positioned nucleosome [97], which then serves as a barrier for phasing of adjacent nucleosomes. On the other hand, a finer-scale 10 bp periodicity in A/T and G/C containing dinucleotides was found along the nucleosome-bound DNA and was proposed to contribute to precise positioning and/or rotational setting of DNA on nucleosomes [99, 103].

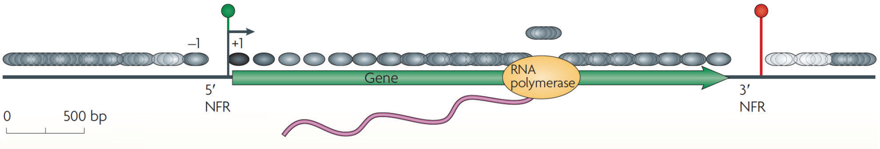

Figure 2. Nucleosomal landscape around yeast genes showing nucleosome-free region (NFR) upstream of the transcription start site (arrow) and downstream of transcription termination site. Array of well-positioned nucleosomes is present downstream of the TSS. The figure is adapted from [91].

How the nucleosome positioning pattern found around gene promoters is established and whether it requires active transcription by RNAPII machinery is still debated. There is evidence for both transcription-independent DNA sequence-driven [104], and transcriptional activity-aided nucleosome organisation [97], suggesting that there might not be a single mechanism responsible for nucleosome positioning at all promoters, but that the mechanism might be dependent on the type of the promoter itself.

Nucleosome positioning and formation of the "beads on the string" structure is just the first level of chromatin compaction. Further successive folding events lead to a higher level of organisation and formation of specific chromatin domains, involved in activation or repression of gene transcription [105, 106]. The organisation of the genome in the nucleus establishes the localisation of genes within those domains and also determines which parts of the genome will be in close proximity and potentially able to interact. Thus, dynamics at the chromatin level is an important factor in gene regulation.

Within the scope of gene regulation, the term *epigenetics* refers to functionally relevant changes to the genome or the chromatin that influence gene expression without altering the underlying DNA sequence (genetic information). These can be chemical modifications to either DNA or histone proteins, which mediate both heritable changes in gene activity and long-term alterations in the transcriptional potential that are not necessarily heritable.

The best-studied epigenetic modification acting directly on DNA is methylation of cytosine, which in vertebrates occurs mainly in the CpG dinucleotide context. DNA methylation is essential for normal development and is involved in several key processes including X-chromosome inactivation, genomic imprinting and suppression of repetitive elements [107]. *De novo* methylation occurs mainly during embryonic development, but it can also happen in adult cells due to aging or carcinogenesis. The majority of CpG dinucleotides in vertebrate genomes are methylated, except those located within CGIs. A small proportion of CGIs become methylated during development, causing permanent silencing of associated promoters and ensuring lineage-specific expression of developmental regulatory genes [108]. There are several mechanisms by which CpG methylation mediates gene silencing: 1) methylated cytosines can alter binding sites for transcriptional activators and exclude them from binding [109], 2) mCpG can serve as a marker for methyl-cytosine binding domain proteins, which recruit co-repressor protein complexes that induce chromatin compaction [110] and 3) methylation directly increases affinity of certain sequences for histone octamer, thus increasing nucleosome occupancy and stability at promoters [111].

Unlike DNA, histones are subject to hundreds of covalent modifications, including acetylation, methylation, phosphorylation, and ubiquitination. These occur primarily at specific positions within the amino-terminal histone “tails”, which emanate from the nucleosome core. Among various modifications, lysine acetylation and methylation are the most studied and best understood. Lysine acetylation almost always correlates with chromatin accessibility and transcriptional activity, and histone H3 lysine 27 acetylation (H3K27ac) was shown to mark active promoters and distal regulatory elements [112, 113]. Tri-methylation of histone H3 lysine 4 (H3K4me3) and H3 lysine 36 (H3K36me3) are both associated with transcribed chromatin; however, H3K4me3 marks promoter regions, whereas H3K36me3 is found along the body of transcribed genes [93, 114]. Unlike promoters, which are tri-methylated at H3 lysine 4, enhancers were shown to be mono-methylated [112, 115]. In contrast to these active marks, tri-methylation of H3 lysine 9 (H3K9me3), H3 lysine 27 (H3K27me3) and H4 lysine 20 (H4K20me3) generally correlate with repression. H3K9me3 and H4K20me3 are marks of constitutive heterochromatin, a tightly packed repressive form of chromatin at repetitive portions of chromosomes [114, 116]. Broad domains of H3K27me3 coincide with Polycomb-repressive complex 2 (PRC2), indicating the sites of Polycomb-mediated repression [117]. They mark loci of transcriptionally silent developmental regulator genes in embryonic stem cells (ESC) [118]. The same loci were shown to contain punctuated H3K4me3 marks localised at promoters even though they were not transcribed [118, 119], suggesting that these "bivalent" domains silence developmental genes in ESCs while keeping them poised for activation.

Even from the very limited set of modifications described above, it is evident that the possibilities of marking genomic loci with various histone modifications and their combinations are enormous. It was proposed that specific combinations of modifications at given locus form a so called "histone code", which is read by other proteins to bring about distinct downstream events [120, 121]. High-resolution mapping of numerous histone modifications in multiple cell types contributed to detection of most common combinations and associated functional genomic elements [122-124] and allowed segmentation of the genome into distinct domains based on the levels of various modifications [122, 123, 125, 126]. Although specific histone modification combinations generally reflect the identity of the underlying DNA element, a recent study has shown that actual levels of modification do not necessarily reflect the predicted regulatory activity [127].

1.2 Mapping genome-wide transcription start sites

1.2.1 Functional annotation of the genome in post-genomic era

The completion of the reference human genome sequence [128-130], as well as genomes of many other model organisms [131-134], together with the advancement in high-throughput sequencing technologies, opened the possibility to study transcription on a genome-wide scale. In this post-genomic era, the functional annotation of the genome proceeded through two complementary approaches. Experimental techniques relying on high-throughput sequencing technologies to map the transcriptome and regulatory elements have been developed. The most widely used techniques include RNA-sequencing (RNA-seq) for genome-wide quantitative mapping of transcribed regions [135, 136], chromatin immunoprecipitation followed by sequencing (ChIP-seq) for mapping transcription factor binding and histone modifications [93, 114, 137], and mapping of DNase I hypersensitive sites (DNase-seq) for identification of open chromatin and regulatory regions [138]. These sequencing-based technologies were preceded by their counterparts that utilised hybridisation to genome-wide tiling arrays [139-141]. On the other hand, computational approaches are used directly on genomic sequence to predict and model transcribed regions (*e.g.* open reading frames and coding sequences [142, 143]) and regulatory elements [37]. (*e.g.* transcription factor binding sites [144]).

1.2.2 Identification of transcriptional promoters

Mapping promoters genome-wide is the first step in deciphering the mechanisms of transcriptional regulation and different approaches have been used to detect promoters along the genome experimentally. The first kind of experiments uses features of active promoters such as presence of the PIC, various promoter-associated histone marks and accessible and open chromatin to localise promoters. For instance, Kim *et al.* [145] used ChIP with an antibody recognising a specific component of the PIC to produce the first genome-wide map of human promoters. Their results have shown widespread use of alternative promoters for many known genes and identified a substantial proportion of promoters that did not map to known genes, suggesting novel transcriptional units [145]. The same study also showed that promoters are associated with the H3K4me3 mark, which was subsequently confirmed in several other studies [93, 114], and that a large proportion of human promoters overlap with CpG islands, whereas other core promoter elements occur much more rarely [145]. However, these approaches can only identify loci that serve as promoters, but cannot map precise transcription start sites or quantify the level of transcription from the detected promoters.

Since RNA transcripts are produced from transcriptionally active promoters, an alternative approach is to use the RNA sequence data to derive positions of the promoters. However, the majority of the transcriptomic data maps transcribed portions of the genome but does not precisely reflect gene boundaries. For instance, a typical expressed sequence tag represents only a random short subsequence of the full complementary DNA (cDNA). Furthermore, RNA-seq, which is the most common technique for quantitative transcriptome profiling, produces uneven coverage of sequenced tags along the transcript, often not covering the 5' end [146]. In order to precisely map promoters, 5' end complete cDNAs are essential. The first genome-wide sequencing and annotation of full-length cDNAs was done for mouse by the FANTOM Consortium [147] and this collection was subsequently used to determined exact TSSs and characterise adjacent putative promoter regions [148]. Similarly, full-length human cDNAs were used to annotate and functionally analyse human promoters [149, 150]. More recently, several techniques that sequence short tags from the 5' end of cDNAs have been developed including 5' serial analysis of gene expression (5' SAGE) [151], oligo-capping [152, 153], cap analysis of gene expression (CAGE) [154] and paired-end ditag technology (PET) [155], which when combined with high-throughput sequencing achieve higher coverage producing more reliable and quantitative mapping of 5' ends. These techniques allow genome-wide precise TSS mapping at single nucleotide resolution and provide the means for analysing promoter-associated features at high resolution.

Different approaches for promoter mapping provide information on distinct aspects of promoter structure and function, making the integration of various datasets essential for understanding transcriptional regulation at promoter level.

1.2.3 Cap Analysis of Gene Expression (CAGE)

CAGE is a high-throughput method for transcriptome analysis [154] that utilizes "cap-trapping" [156], a technique based on the biotinylation of the 7-methylguanosine cap characteristic for RNAPII transcripts. After the biotinylated RNA is reverse transcribed, the resulting RNA/DNA heteroduplex is treated with RNase I to ensure that only 5'-complete cDNAs stay associated with the biotin tag, and pulled down by streptavidin-coated beads. A linker sequence containing a recognition site for a type III restriction endonuclease is ligated to the 5' end of the captured cDNA and the corresponding restriction enzyme is used to cleave off a short fragment (typically 27 bp) from the 5' end [157]. The resulting fragments are then amplified and sequenced using massive parallel high-throughput sequencing technology, which results in a large number of short sequenced tags that can be mapped back to the reference genome to infer the exact position of the TSSs used to initiate transcription of captured RNAs (Figure 3). The number of CAGE tags supporting each CAGE-detected TSS (CTSS) at a particular nucleotide position in the genome gives the information on the relative frequency of its usage and can be used as a measure of expression from that specific TSS [158]. Thus, CAGE provides information on two aspects of the capped transcriptome: 1) a genome-wide single base-pair resolution map of transcription start sites, and 2) relative levels of transcripts initiated at each CTSS. This information can be used for various analyses, from 5' end centred expression profiling [159, 160] to studying promoter architecture [26, 161].

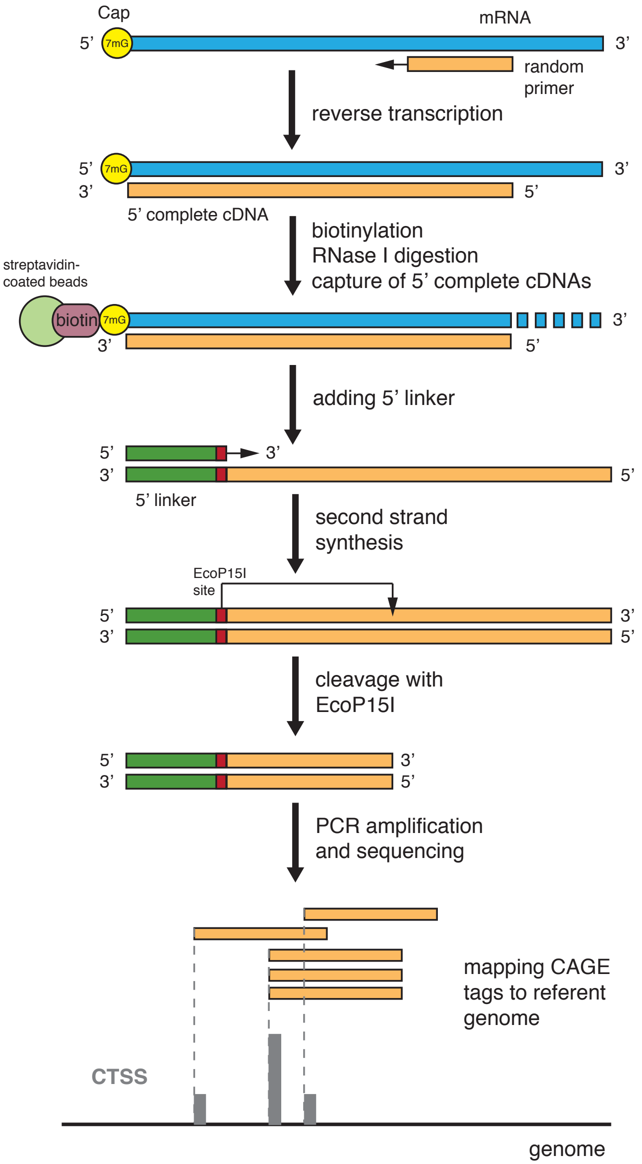

Figure 3. Schematic procedure of the CAGE experimental protocol for mapping transcription start sites at single bp resolution. The figure is redrawn based on [154] and [157].

Quantitative nature of CAGE has been used to model expression dynamics and to reconstruct the regulatory networks driving the differentiation [159] and maintaining identity of numerous human and mouse cell and tissue types (Paper III), by identifying key transcription factors binding at promoters. Moreover, CAGE signal has been shown to be enriched at enhancers [165] and has recently been used to construct an atlas of active enhancers over cells and tissues across the whole human body [166]. Thus, in addition to providing a valuable resource of genome-wide cell type-specific TSSs, which are a more precise alternative to TSS positions available in annotation databases, CAGE is also a powerful approach for studying various aspects of gene regulation.

However, not all genomic positions detected by CAGE seem to correspond to genuine RNAPII transcription initiation sites, as many CTSSs were found within internal exons with CAGE tags spanning exon-exon junctions [26]. A study profiling small RNAs and comparing them to distribution of CAGE tags concluded that processed coding and non-coding RNAs are metabolized into short RNAs that likely bear cap-like structures at their 5' ends and are captured by CAGE tags [167]. The function of these short and CAGE sensitive RNAs mapping to internal exons is still a mystery. However, these RNA species arise only from a discrete subset of genes and their abundance often does not correlate with the expression of the host gene, arguing against them being merely degradation intermediates [167].

1.2.4 Pervasive transcription and the landscape of transcription initiation

It is evident from a number of studies that the majority of the genome is transcribed in Metazoa [162, 165, 168-171]. A large number of non-coding transcripts arise from intronic and intergenic regions [162, 169] raising the questions of how and from which promoter they are expressed. In addition, many ncRNAs align with protein coding genes, either in the same orientation (sense) or in the opposite orientation (antisense) to the coding transcript [162, 172], further increasing the transcriptional complexity and providing regulatory potential. Recent studies have also identified a variety of ncRNAs transcribed from regulatory elements, such as promoter-associated RNAs [170] and enhancer RNAs [173]. The biological significance of this pervasive transcription and the function of various classes of ncRNAs are still largely unclear and controversial.

All of the above observations suggest that the current view of the genomic organisation into distinct gene units and associated regulatory elements that drive its expression might not account for the observed transcriptional complexity. Instead of being a product of regulation, the pervasive transcriptional activity itself might have a regulatory function [174], creating a complex transcription initiation landscape that yet remains to be deciphered.

1.3 Promoter structure and function

1.3.1 Core promoter elements and TSS selection

The "textbook" model of an RNAPII promoter has an A/T-rich DNA sequence (the TATA-box) approximately 30 bp upstream of the TSS, which in turn overlaps an initiator sequence (Inr) (Figure 4). The assembly of a PIC at such promoters is initiated by TFIID binding to the TATA-box, Inr sequence and/or other sites [15]. TFIID is a multi-protein complex comprising the TATA-box binding protein (TBP) and more than 10 distinct TBP-associated factors (TAFs) [13]. TBP is a crucial component that recognises and binds the TATA-box motif [175], initiating subsequent PIC assembly and RNAPII recruitment. Once the PIC has assembled, the region around the TSS melts to provide a template strand for RNAPII, which occurs 25–30 bp downstream of the TATA-box in all eukaryotes, except in budding yeast, where this distance can vary [176, 177]. Where present, the TATA-box seems to be the main determinant of the TSS, and initiation will occur at the suitable initiator-like sequence at an appropriate distance from the TATA-box [178].

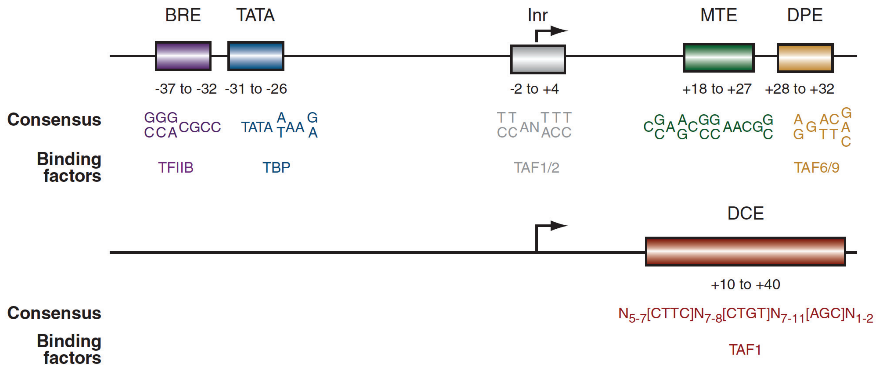

Figure 4. Metazoan core promoter elements. Position relative to TSS, human consensus sequence and transcription factors that bind each element are shown. The downstream core element (DCE) is shown on a separate promoter for illustration purpose only, although it can be present together with TATA-box and/or initiator element. A particular core promoter may contain some, all, or none of these elements. BRE: TFIIB-recognition element, Inr: initiator element, MTE: motif ten element, DPE: downstream promoter element, DCE: downstream core element. The figure is adapted from [18].

Although the TATA-box is a well known core promoter motif, it is present only in the minority (<15%) of mammalian promoters of protein-coding genes [19, 26, 41, 145]. A recent study mapped PIC components at high resolution in human genome and suggested that the TATA-box motif is more prevalent than previously thought and concluded that it is a general feature present in core promoters of both coding and non-coding transcripts [179]. However, analyses that led to this conclusion were not correctly designed and the prevalence of core promoter elements was not statistically validated [180], which led to the retraction of the reported results [181].

A more abundant, yet also not universal, metazoan core promoter element is the initiator (Inr), which directly overlaps the TSS [182]. The consensus sequence of *Drosophila* and mammalian Inr differs to some extent, however it is bound by the homologous TAFs within the TFIID complex, which include TAF1 and TAF2 [15]. The common characteristic of the Inr element is the pyrimidine (C or T) / purine (A or G) motif positioned -1/+1 bp relative to the TSS, so that the purine is the first transcribed nucleotide [15, 26]. In *Drosophila* the Inr element often occurs in combination with either a TATA-box [183], or with another core promoter element located downstream of the TSS, the downstream promoter element (DPE) [184]. They act synergistically to increase the efficiency of transcription by providing additional recognition sites for TFIID components and allowing cooperative TFIID binding.

The DPE was discovered in the analysis of TATA-less promoters in *Drosophila* [184] and was suggested to be conserved in humans [185]; however, its presence in mammalian genomes was never supported by high-resolution CAGE data. This element acts in conjunction with the Inr, and the core sequence of the DPE is located at precisely +28 to +32 bp relative to the +1 nucleotide in the Inr motif [186]. This strict requirement for Inr–DPE spacing is essential for cooperative binding of TFIID, thus DPE and Inr function together as a single core promoter unit. Transcription initiation from DPE-containing promoters is dependent on TAFs, specifically TAF6 and TAF9, which were shown to bind the DPE [13, 185].

The TFIIB recognition element (BRE) is the only well-characterized core promoter motif bound by a factor other than TFIID. It was initially identified as a sequence immediately upstream of a subset of TATA-box elements [187]; however, an additional TFIIB recognition site, the downstream BRE, was found immediately downstream of the TATA box [188]. Several studies have shown that TFIIB plays a central role in transcription start site selection in both yeast and human [189]. Multiple mutations in TFIIB were found to cause a shift in the TSS selection, suggesting its role in the precise positioning of RNAPII catalytic site at some core promoters [190]. BRE elements often occur in conjunction with the TATA-box and the observed spacing between TATA-box and TSS is a result of interaction between TBP, TFIIB and RNAPII, where TFIIB plays a central role in determining the spacing.

Despite the prevalence of CpG island-associated promoters, the precise mechanisms of their core promoter function are not well understood. One common feature of CGIs is the presence of multiple potential binding sites for transcription factor Sp1 [191]. Sp1 contributes to the maintenance of the hypomethylated state of CGIs and may work in concert with the general transcription machinery to support nucleation of the PIC [191]. TSSs are often located 40–80 bp downstream of the Sp1 sites, which suggests that Sp1 may direct the basal machinery to form a PIC within a loosely defined downstream window [192]. One possibility is that TFIID subunits capable of core promoter recognition then interact with the sequences within that window that are most compatible with their DNA recognition motifs, such as an Inr element, to specify the exact TSS.

Initial studies suggested that the basal transcription machinery is largely invariant across different cell types and conditions. However, an increasing number of tissue-specific isoforms of TAFs as well as additional members of the TBP protein family such as TBP-related factors (TRFs) have been identified in Metazoa and found to form distinct TFIID-related complexes that can function at distinct core promoters [193]. Interestingly, many of these factors are involved in germ cell development [194, 195]. The variability in basal transcription machinery composition might require different mechanisms for core promoter recognition leading to distinct patterns of TSS selection.

1.3.2 Nucleosome positioning and epigenetic features of promoters

Distinct chromatin structure and histone modifications have been associated with active promoters. Both in yeast and Metazoa, the region immediately upstream of the TSS is marked as a DNase I hypersensitive site, suggesting that it is a region of open chromatin depleted of nucleosomes [57, 196]. This nucleosome-free region makes core promoter elements more accessible and facilitates PIC assembly and RNAPII recruitment. The accessibility of the promoter was shown to correlate with mRNA abundance [196].

The NFR is flanked by two nucleosomes, the first upstream or -1 nucleosome and the first downstream or +1 nucleosome, whose positioning can be more or less precise depending on the type of the promoter [164, 197]. How the transcription initiation machinery contends with the +1 nucleosome seems to be different across different types of promoters. Precise mapping of PIC components in yeast showed that TFIID–TAF complex engages and is positioned by the +1 nucleosome at TATA-less promoters, whereas TATA-box containing promoters are largely depleted of TAFs and mediate PIC positioning through TBP and TFIIB interactions with the DNA [177]. Thus, in TATA-box promoters the +1 nucleosome can often overlap the TSS. Similarly, it was shown that at many promoters in *Drosophila* the +1 nucleosome resides >50 bp downstream of the TSS, where it engages with the paused RNAPII [94], further suggesting active role of the +1 nucleosome in transcriptional machinery positioning and RNAPII pausing.

Another important feature of nucleosomes flanking the TSS is the presence of specific histone variants. The H2A.Z variant was shown to be associated with promoters in both yeast and Metazoa [93, 94, 198]; however, in yeast both -1 and +1 nucleosomes incorporate H2A.Z, whereas in *Drosophila* this variant is found exclusively in the +1 and additional downstream nucleosomes [94]. Histone variant H3.3 was also found to be enriched at promoters, where it was present almost exclusively together with H2A.Z. These H3.3/H2A.Z double variant–containing nucleosomes mark promoters and other regulatory regions and are surprisingly found within NFRs [199], which should by definition be devoid of nucleosomes. However, it seems that they are very unstable and thus not detected under the conditions normally used in nucleosome preparation [199]. This instability might facilitate the access of transcription factors to promoters and other regulatory sites *in vivo*.

Promoter-associated nucleosomes are also subject to various histone modifications that were shown to correlate with promoter activity [93, 114, 122, 123]. The best-studied modifications associated with active promoters are H3K4me3 and H3K27ac, where H3K27ac level seems to be positively correlated with the level of expression, whereas H3K4me3 can be present on promoters that are not actively transcribing, but are poised for activation [118, 122, 123]. It was shown that basal transcription factor TFIID directly binds to the H3K4me3 mark via a specific domain of TAF3 [200], which suggests that H3K4me3 might play an important role in defining core promoters. TAF3-mediated binding of TFIID to H3K4me3-marked nucleosomes could serve either to anchor TFIID to already activated promoters or to recruit TFIID during promoter activation. Interestingly, TAF3-H3K4me3 interaction seems to be more important for activation of TATA-less promoters, implying the importance of this mechanism for activation of promoters lacking canonical core promoter DNA elements [200].

Because many PIC components, including TFIID, have nucleosome-binding subunits, positioned nucleosomes might define the location of the TSS by positioning the PIC. The conventional view is that most genes contain a predominant TSS, the location of which is defined by core promoter elements [28]. However, many promoters lack any of the known core promoter elements and the question remains how the transcription machinery establishes the location of the TSS at those promoters. A model has been proposed in which the TFIID complex binds to methylated (and acetylated) nucleosomes and recruits TBP to promoters [91]. TBP in turn binds TFIIB and places it immediately downstream towards the TSS. Since TFIIB was shown to dictate TSS selection [189], this model would explain how TSS positioning could be directed in part by TFIID bound to nucleosomes.

1.3.3 Promoter classes and modes of regulation

Early studies on individual promoters that had led to the discovery of various core promoter elements already suggested substantial promoter heterogeneity. Some combinations of core promoter elements were observed more often than others, defining different structural and functional types of promoters. For instance, the TATA-box and DPE are rarely found together, but each of them is often associated with an Inr element [19, 184, 186]. Furthermore, the TATA-box containing promoters appear to be functionally different from the DPE containing ones, and to respond to distinct distal regulatory elements [89].

Genome-wide mapping of promoters and promoter-associated features allowed comprehensive analysis of promoter structure and function and their classification based on underlying sequence, chromatin, transcription initiation and expression specificity characteristics. The underlying sequence composition analysis revealed that mammalian promoters segregate naturally into two classes by CpG dinucleotide content: high-CpG and low-CpG promoters [23]. The former class is characterised by the overlap with CpG islands, thus they are also referred to as CGI-associated promoters. High resolution mapping of TSSs by CAGE distinguished two major classes of promoters based on the TSS distribution [26]. “Sharp” (or “focused”) promoters have a single well-defined TSS and are often associated with a TATA-box precisely positioned ~30 bp upstream of the TSS [26, 178]. These classical “textbook” promoters represent only a minority of mammalian promoters and are commonly associated with tissue-specific genes and high conservation across species. Many TFs show distinct spatial biases with respect to TSS location and seem to be important contributors to the accurate prediction of single-peak TSSs [201]. The majority of mammalian promoters, however, comprise a second class of “broad” or “dispersed” promoters, characterised by multiple equally used TSSs distributed across a broader region [26], challenging the traditional definition of a gene and its precisely defined TSS. This class is strongly associated with CpG islands and ubiquitously expressed genes, however promoters of key developmental regulators were also found to belong to this class [90].

High resolution TSS mapping by PET in *Drosophila* revealed analogous transcription initiation patterns [202], separating promoters into “sharp” and “broad” classes. Unlike mammalian genomes, the fly genome does not contain CpG islands; however, the two promoter classes were shown to be associated with distinct core promoter elements. The positionally restricted canonical core promoter elements, including the TATA-box, Inr, DPE and MTE, were specifically enriched in sharp promoters [202, 203]. When comparing across other *Drosophila* genomes, elements in broad promoters had lower levels of conservation than those in sharp promoters [203]. Furthermore, the distinct promoter classes in fly were associated with the same functional categories of genes and showed similar expression specificity patterns as in mammals [26, 202, 203]. Together, this suggests functional conservation of the observed promoter classes across Metazoa.

Genome-wide analyses of various promoter-associated features provided further insight into structural and functional differences between CpG and non-CpG promoters in mammals. In pluripotent ES cells, a vast majority of CpG promoters are associated with H3K4me3 enrichment [114], suggesting that they are targets of trithorax-group proteins, which catalyse the deposition of this mark. These promoters have a potential to drive transcription, unless they are actively repressed by Polycomb group proteins (PcG), which deposits the repressive H3K27me3 mark and creates bivalent domains at key developmental genes and poises them for activation [118, 204]. The ones that are not repressed tend to be ubiquitously expressed. In contrast, CpG-poor promoters seem to be inactive by default, independent of repression by PcG proteins, and may instead be selectively activated by cell-type- or tissue-specific factors [114]. This is further corroborated by the observation that CpG promoters are associated with RNAPII across multiple cell types, whereas non-CpG promoters acquire active chromatin marks and RNAPII binding in a tissue-dependent way [205]. The two promoter classes also differ in nucleosome occupancy and the requirement for nucleosome remodelling complexes for their activation upon various external stimuli [206]. Taken together, this strongly suggests that CpG and non-CpG promoters in mammals are subject to distinct modes of regulation.

Unlike CpG and non-CpG promoter classification, which is vertebrate-specific, the corresponding sharp and broad promoter classes defined based on transcription initiation patterns are conserved across Metazoa [26, 202, 203]. These promoter classes are significantly differentiated by nucleosome organization and chromatin structure in both fly and mammals. Broad promoters display closer association with well-positioned nucleosomes and activating histone marks downstream of the TSS and have a more clearly defined NFR upstream, while sharp promoters have a less organized nucleosome structure and higher RNAPII presence [197].

Based on the configuration of promoter signals, TSS patterns, nucleosome positions and their epigenetic marks, and function of the associated gene, a unifying classification of metazoan promoters into three main classes was proposed (Figure 5) [164].

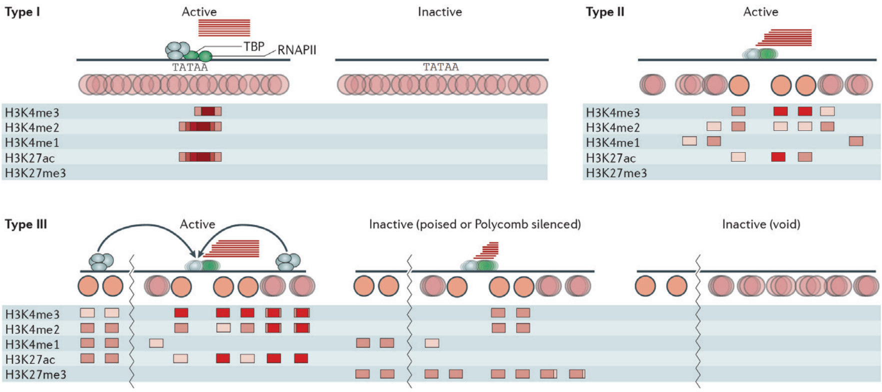

Figure 5. Transcriptional and chromatin features of the three main functional classes of metazoan promoters. Horizontal red lines represent 5' ends of transcripts reflecting the transcription initiation pattern. Nucleosomes are represented by red circles and the "fuzziness" reflects the precision of nucleosome positioning. The figure is adapted from [164].

Type I promoters are most often used for genes that are specifically expressed in terminally differentiated peripheral tissues of an adult. They are characterised by a sharp transcription initiation pattern and are often associated with a TATA-box or other core promoter elements positionally restricted to the well-defined TSS in both mammals and fly. In mammals they are characterised by low CpG content and tend to have key regulatory inputs close to their promoters [207]. At the chromatin level, Type I promoters are characterised by less-ordered nucleosomes [197], which can often cover the TSS; with H3K4me3 generally present downstream of the TSS when they are active and no RNAPII binding when they are not active [205]. Type II promoters are associated with ubiquitously active "housekeeping" genes and have broad promoter architecture with multiple TSSs spread across a wide region. In mammals, they tend to have a single CpG island covering the transcription initiation region, whereas in *Drosophila* they are associated with a distinct set of weaker core promoter elements [208]. The TSSs are located within a NFR and are flanked by two well-positioned nucleosomes that harbour active histone marks in all cell and tissue types. Type III promoters are characteristic of genes with expression that is developmentally regulated and coordinated across multiple cells. They share several characteristics with type II promoters, including a broad transcription initiation pattern and a well-defined NFR with positioned flanking nucleosomes, but also exhibit systematic differences that set them apart from the ubiquitously expressed class. The width of their transcription start region tends to be even broader than in Type II promoters [161]. Although their association with CGIs in mammals is similar to type II promoters, developmental genes have longer or multiple CGIs that often extend into the gene body [90]. The most prominent differences between type III and type II promoters are observed at the chromatin level. Developmental genes have a number of features that are associated with repression by PcG proteins, including wide distribution of PcG protein binding and both H3K27me3 and H3K4me3 marks, which create bivalent domains in ESCs [118]. Type III promoters are responsive to long-range regulation and can receive and integrate regulatory input from distal enhancers. They are often surrounded by arrays of HCNEs that act as enhancers ensuring precise spatial and temporal expression of those key developmental regulators [90].

1.3.4 Promoter usage dynamics

The traditional view of a gene with its precisely defined and fixed TSS has been first challenged by the findings that many genes can be transcribed from multiple promoters (alternative promoters) producing functionally diverse transcripts [209, 210]. Differential utilization of alternative promoters plays a critical role in regulating gene expression in a spatial, temporal or lineage-specific manner. This can be achieved by use of a distinct combination of core promoter elements in the alternative promoters [211, 212]. Moreover, studying 5’ ends of individual mRNAs by oligo-capping [152] and more recently genome-wide by CAGE, revealed that the transcription can start at multiple closely spaced TSSs within a single “broad” promoter [26], further increasing the diversity of produced transcripts. The closely spaced individual start sites can be associated with different core promoter elements and their activation can be dependent on distinct GTFs [213].

The complexity of transcription initiation in eukaryotic genomes is also seen in the bidirectional promoter arrangements, which in the human genome comprise more than 10% of promoters [214]. Bidirectional promoters are associated with broad transcription start regions overlapping a CGI and display a mirror sequence composition [215]. The transcription from bidirectional promoters can be differentially regulated in the two directions [177], suggesting that the promoter elements and features can overlap in the same locus and be differentially interpreted by the RNAPII complexes transcribing independently in the opposite directions. Thus, bidirectional promoters are a good example of overlapping transcription initiation codes, which are differentially interpreted in different regulatory contexts.

Differential utilisation of promoter types has been observed across various contexts. For instance, in *Drosophila* embryonic development promoters of maternally inherited transcripts showed differences in motif composition compared to zygotically active promoters [203]. In addition, many genes with maternally inherited transcripts were found to have alternative promoters utilized later in the development [203]. High-resolution quantitative mapping of TSSs across multiple human and mouse tissue types revealed substantial dynamics even at the level of individual TSSs within the same core promoter [216]. TSS selection within many CGI-associated broad promoters varies among tissues producing positional or regional bias in promoter usage [216]. This fine-scale regulation of transcription initiation events at the base pair level is likely related to epigenetic transcriptional regulation.

2 Aims of the study

The main aim of this thesis was to study the patterns of transcription initiation at high resolution, along the following principles:

- construct a genome-wide map of transcription initiation at single base-pair resolution during vertebrate embryogenesis using zebrafish (*Danio rerio*) as a model organism

- characterise the zebrafish maternal and embryonic promoterome in terms of different types of utilised promoters and associated promoter elements

- analyse dynamic changes in promoter usage throughout embryogenesis

- monitor the changes in promoter-associated nucleosome positioning and transcription-associated histone modifications during development

- infer the logic of transcription start site choice in maternal and zygotic transcription

- reveal sequence signals and nucleosome positioning underlying TSS choice at different promoters and in distinct regulatory environments

- expand the study of precise TSS-related sequence and nucleosome signals to mammalian genomes (human and mouse)

- develop a resource and tool for mining and visualisation of high-resolution TSSs derived from CAGE data, to facilitate the use of this precise and context-specific data in promoter-centred integrative analyses

- introduce CAGE data as a more precise and functionally relevant resource of TSSs than currently more widely used static TSS annotations available in common databases.

To address the listed points I have used computational and statistical approaches to analyse various types of genome-wide data produced experimentally by our collaborators. Since the publications included in this thesis are a result of collaborations and contain contributions from both experimental and computational collaborators, the following sections summarise the results to which I have contributed the most.

3 Summary of the results

3.1 Single nucleotide resolution map of transcription initiation during zebrafish embryogenesis (Paper I)

To characterise the promoter repertoire and its dynamic use during the development of a vertebrate embryo, we mapped transcription initiation events at single nucleotide resolution by CAGE in 12 stages of zebrafish (*Danio rerio*) development, spanning from unfertilised egg to 33 hours past fertilisation (hpf). This period includes the maternal to zygotic transition at mid-blastula transition (MBT), which represents the most dramatic change in transcription programme in vertebrate life cycle. Before the MBT, there is no transcriptional activity and the transcriptome of the early embryo reflects the transcriptional programme of the oocyte. During MBT, activation of the zygotic genome occurs in parallel with maternal mRNA degradation and the newly synthesised transcripts replace the inherited ones [217].

As expected, the majority of CAGE tag clusters (TC) were located in the vicinity of the 5’ ends of annotated genes. However, there was a substantial proportion of both inter- and intragenic CAGE signal, indicating potential unannotated promoters and post-transcriptionally processed RNA products, respectively. We discovered many novel and alternative promoters and showed that they are indeed functional by their association with activating H3K4me3 histone mark and downstream RNA-seq signal. A small subset was also tested in transgenic assay and shown to drive transcription.

High-resolution transcription initiation patterns derived from CAGE revealed the dichotomy of promoter width in zebrafish, separating sharp and broad promoter architectures, previously characterised in mammals [26] and fly [161]. This further corroborated the conservation of observed promoter classes across Metazoa. Maternal stages were characterised by a significantly higher proportion of sharp promoters compared to zygotic stages, whereas the usage of broad promoters increased after zygotic genome activation. We also detected widespread usage of multiple TCs within promoter regions in maternal stages, followed by a noticeable reduction during zygotic stages. In contrast, genes active in the embryo were more often associated with more than one promoter than those active in the oocyte, indicating prevalent usage of alternative promoters in the zygotic transcriptome.

To characterise non-promoter CAGE signal, we analysed the dynamics of exonic and intronic TCs separately. We provide several lines of evidence that exonic CAGE-detected RNAs are of post-transcriptional origin and not initiated from intragenic promoters. The exonic RNAs appear before zygotic genome activation and CAGE tags supporting them often map across splice junctions. Finally, the sequences underlying exonic TCs do not drive expression in transgenic assays. In contrast, intronic TCs are developmentally regulated and many of them suggest splice site-associated RNAs [218] arising during zygotic transcription. Interestingly, we found them to be enriched in introns of splicing-associated genes in both zebrafish and human, suggesting a potential mechanisms linking splicing activity with the regulation of expression of the splicing machinery. Both exonic and intronic TCs showed no sequence signatures found in conventional promoters and were not associated with H3K4me3, suggesting that they are not used as promoters. The associated RNAs are likely generated by post-transcriptional processing of full-length RNAs, which seems to utilise different mechanism for exonic and intronic RNAs.

We complemented the zebrafish promoter map with the first CAGE promoter map of *Tetraodon nigroviridis*, another teleost fish species, which allowed identification of conserved promoter features. We identified a novel GAAG core promoter motif that is used as an initiator by a small set of orthologous genes involved in vesicle transport and membrane-associated functions, and confirmed its presence and association with the same functional group of genes in human.

3.2 Overlapping transcription initiation codes drive dynamic promoter usage in zebrafish development (Paper II)

The high-resolution map of transcription initiation in zebrafish generated by CAGE (Paper I) gave us an opportunity to study developmental dynamics at individual TSS level. Expression profiling of individual TSSs revealed expected expression patterns, separating TSSs of inherited maternal transcripts that follow previously observed degradation pattern [217] from the TSSs of newly synthesised zygotic transcripts, which accumulate after MBT. Surprisingly, the two types of TSSs with antagonistic expression dynamics were often present within the same promoter region, creating a shifting pattern of promoter usage throughout development. Guided by this observation, we developed a novel method to systematically capture promoters with such shifting patterns, which we named "shifting" promoters. These promoters are expressed in both maternal and zygotic stages, keeping the expression of the associated genes constant; however, the differing expression dynamics and separation of individual TSSs within them allowed us to distinguish maternally inherited from newly synthesised transcripts initiated from the same promoter.

Using the defined set of shifting promoters, we further studied the sequence features underlying maternal and zygotic TSS selection. We found a sharp enrichment of TA, AT, AA and TT dinucleotides (WW dinucleotides) aligned precisely ~30 bp upstream of maternal TSS, indicating a presence of a functional TATA-box [178]. However, motif discovery *de novo* revealed a weaker and more degenerate motif than the canonical TATA-box, which we termed W-box. We also discovered a novel promoter architecture characteristic for maternal transcriptome, the multiple sharp architecture, in which every sharp TSS sub-cluster was associated with a W-box at the appropriate upstream position. Thus, the TSS selection in the oocyte seemed to be dependent on a precisely positioned W-box motif. We functionally validated this hypothesis in stable transgenic zebrafish lines carrying mutations in the W-box motifs, which disrupted the usage of associated downstream positions as TSSs.

In contrast, the zygotic TSS did not align with the observed W-box motif, but was characterised by a broader band of GC/CG dinucleotide enrichment around the TSS (mirrored by WW dinucleotide depletion) forming a sharp boundary ~50 bp downstream of the TSS. The boundary was followed by additional downstream alternating bands of GC/CG depletion and enrichment, which were exactly wide enough to accommodate a single nucleosome. This suggested that zygotic TSS selection is independent of the W-box motif and might be associated with nucleosome positioning. We expanded our analysis to the entire set of throughout-expressed promoters, including the ones that did not exhibit spatial separation in TSS usage, but rather had maternal and zygotic TSSs intertwined in the same region, and showed that preferred maternal and zygotic TSSs in all promoters follow the observed dinucleotide patterns. This confirmed a promoterome-wide distinction between determinants that govern TSS selection in the oocyte and the embryo and drive dynamic changes in promoter shape during development.

To study promoter-associated nucleosome positioning and its dynamics throughout the development, we mapped the positions of H3K4me3-marked nucleosomes by ChIP-seq coupled with micrococcal nuclease digestion in 4 developmental stages, both before and after MBT. This revealed precise positioning of H3K4me3-marked nucleosomes starting ~50 bp downstream of the zygotic TSS and aligning with dinucleotide enrichment patterns. In addition to broad bands of dinucleotide enrichments, we were able to detect the 10 bp periodicity in AA/TT dinucleotide frequency starting ~50 bp downstream of the TSS, which provides the intranucleosomal positioning signal [103] for the +1 nucleosome. Even before MBT, nucleosomes were roughly aligned with the zygotic TSSs that are yet to be activated, which together with the tight association between the TSS and nucleosome positioning signal suggests that +1 nucleosome guides TSS selection in the zygotic transcriptome. Furthermore, we provided several lines of evidence that H3K4me3-marked nucleosomes acquire their mark before zygotic genome activation and upon activation assume their final sequence-guided position downstream of the zygotic TSS independent of the transcriptional activity.

Our work revealed two independent codes guiding TSS selection in the oocyte and the developing embryo, and demonstrated that complex TSS patterns in constitutively expressed promoters represent readouts of two independent grammars intertwined in the same core promoter region.

3.3 Precise TSSs reveal underlying sequence and nucleosome positioning signals in mammalian promoters (Paper III)

As a member of the FANTOM consortium, a large collaborative initiative aimed at creating a comprehensive overview of mammalian gene expression at a promoter level, I was involved in analysis of CAGE datasets derived from numerous human and mouse primary cells, cell lines and tissues produced within FANTOM5 project. My work was focused on characterising high-resolution TSS patterns and promoter architectures, and their associated sequence and chromatin configuration features. Analysis of TSS distribution and promoter width confirmed the previously established separation into sharp and broad promoter architectures [26], which was detected across all cell and tissue types in both human and mouse. The number of various cell and tissue types allowed us to address the difference in the global expression specificity between the two promoter types. Sharp promoters were shown to have more restricted expression patterns, in line with the observation that they are more often associated with tissue-specific genes. A similar difference in expression specificity was observed for non-CpG versus CpG promoters.

The sequence underlying sharp and broad promoters displayed very different nucleotide patterns. Sharp promoters were associated with a narrow peak of WW dinucleotide enrichment positioned precisely ~30 bp upstream of the TSS, indicting the presence of a functional TATA-box or a TATA-like signal. In contrast, the dominant TSS in broad promoters aligned with the 10 bp periodic pattern in WW dinucleotide frequency starting ~50 bp downstream of the TSS. This precise phasing that provides intranucleosomal positioning signal [103] was shown to coincide perfectly with the position of the H2A.Z and H3K4me3-marked first downstream nucleosome in two different cell types. The tight association between dominant TSS in broad promoters and the nucleosome positioning signal indicated that the positioned +1 nucleosome is a key determinant of TSS preference in broad promoters. Finally, the presence of this association in both zebrafish (Paper II) and mammals suggests the evolutionary conservation of the underlying nucleosome-associated TSS selection mechanism.

3.4 Resource and tool for high resolution promoterome mining for integrative analyses (Paper IV)

All promoter-centred analyses of genome-wide data rely on TSS annotation and currently the widely used approach is to use static TSS annotations available in common databases such as RefSeq [219] and Ensembl [220]. Cap analysis of gene expression (CAGE) provides context-specific TSSs at single base-pair resolution. Despite their superior resolution and functional significance, published CAGE data are still underused in promoter analysis due to lack of tools that would enable easy access to available published datasets and its processing and integration with other genome-wide data. To address this, I developed *CAGEr*, a Bioconductor-compliant [221] software package for R statistical and computing environment [222].

The *CAGEr* package provides direct access to majority of published CAGE datasets including the large collection of ENCODE cell lines [165] and the recently published FANTOM5 collection for human and mouse primary cells and tissues (Paper III). It allows users to import and manipulate single bp resolution TSSs, cluster them into a context-specific promoterome and obtain various associated information such as position of the dominant TSS, promoter width and expression pattern across multiple contexts. This information provides additional layers in promoter-centred analyses of other types of genomic data, enabling separation of different classes of promoters. I have also implemented our novel method for detection of shifting promoter patterns (Paper II) alongside with the state-of-the-art methods for CAGE signal normalisation, TSS clustering and assessment of promoter width. Informative graphical outputs and track files for visualisation in the genome browser can be exported. All functionality is provided through well-documented high-level commands, which are organised into a comprehensive workflow and are accessible to users with no previous experience in CAGE data analysis.

We demonstrate the *CAGEr* workflow by applying it to a previously uncharacterised CAGE time-course of mouse testis development produced within FANTOM5 project (Paper III). The analysis revealed widespread differential TSS usage and promoter shifting between mouse embryonic and adult testis, suggesting significant changes in regulatory environment underlying mouse spermatogenesis, which drive differential TSS choice.

4 Discussion and perspectives

In Paper I we have provided the first quantitative mapping of single nucleotide resolution TSSs in zebrafish, an important vertebrate model organism [223]. This TSS data complements mammalian cell culture-based [165] and non-vertebrate animal models [161], and provides the first description of core promoter dynamics during vertebrate embryogenesis. Our results demonstrate global and pervasive changes in promoter utilisation during maternal to zygotic transition, which is characterised by a complete turnover of the transcriptome in the early stages of embryonic development. Widespread usage of alternative promoters during development suggest variability in transcripts 5' sequences and has implications for various aspects of genetic manipulations in zebrafish, from designing translation blocking knock-down reagents, such as morpholino antisense nucleotides, to introducing site-specific mutations.

Zebrafish is also an important system for transgenic assays designed to control cell-type specific expression or to detect and functionally characterise *cis*-regulatory elements (*e.g.* enhancer testing) [224]. Understanding core promoter architecture and regulation is essential in choosing appropriate core promoter sequences for transgenic assays; thus, the high-resolution TSS map and associated developmental dynamics provided in Paper I represent a valuable resource of relevant promoter information. Apart form being a significant contribution to characterising zebrafish as a model organism, the data presented in Paper I provides an opportunity for comparative analyses of transcription initiation during development and elucidation of features and mechanisms underlying transcription initiation dynamics.

Different TSS selection grammars deployed at separate promoters have been associated with different types of genes [177, 197], and only a handful of promoters were shown to switch between TATA-dependent and -independent initiation [225, 226]. In Paper II, we show for the first time that the two grammars co-exist in close proximity or physically overlap genome-wide, and are differentially used at thousands of promoters active in both the oocyte and the embryo. Our findings raise several important questions that provide directions for future studies.

Activation of the zygotic genome during MBT is characterised by the switch form maternal W-box guided TSS selection happening in the oocyte, to zygotic TSS selection, which is restricted by the position of the first downstream nucleosome and aligns dominant TSS with inter- and intranucleosomal positioning signals. However, the question remains: Why does the switch happen and when does the switch back occur? At some point during female germline development, the zygotic mode of transcription needs to be replaced by the maternal mode to produce the observed transcriptome of the differentiated oocyte. Germ-cell determinants are deposited early in zebrafish embryognesis and by the 24 hpf the primordial germ cells (PGC) have already migrated to their final location [227]. It is possible that these cells already utilise maternal mode of TSS selection, however our whole embryo data at differentiated stages inevitably masks cell type-specific promoter usage. Studies focusing on the promoterome of PGCs and its dynamics during female germline development are necessary to address this question.

Germline development in vertebrates is characterised by epigenetic reprogramming, including the demethylation of vast majority of CpG sites [228]. Since zygotic TSS selection is closely linked to CpG dinucleotide enrichment patterns and is likely promoted by the demethylated state of CpGs within CGIs, this mode of transcription initiation might be incompatible with the demethylated oocyte genome. Thus, the switch to maternal TSS selection, which is guided by the precisely positioned W-box motif, might be a mechanism to prevent unwanted transcription initiation at demethylated CpG sites throughout the genome during epigenetic reprogramming.

In our work, we have revealed and functionally validated two overlapping transcription initiation grammars. Further studies focusing on transcriptional machinery in the embryonic development and its interaction with uncovered sequence and chromatin features of core promoters should shed more light on how TSS selection in the two regulatory environments is mediated. Given the observation that the composition of the transcription machinery can be cell-type specific and is able to actively contribute to gene regulation [229, 230], one plausible mechanism involves oocyte- and early embryo-specific components of the basal transcription machinery. For instance, the TBP2 factor, a vertebrate-specific member of the TBP family, was shown to be a substitute for TBP in oocytes and is essential for germline development in mouse [231] and frog [195]. TBP2 is also highly expressed in zebrafish oocytes [232] and could mediate the observed maternal TSS selection through W-box binding. In contrast, the transcription machinery in the early embryo might preferentially interact with nucleosomes and mediate zygotic nucleosome-guided TSS selection through motif-independent TFIID recruitment by H3K4me3–TAF3 interactions [200].

The results presented in Paper II, including tight association of nucleosome positioning signal and dominant TSS, strongly suggest that precisely positioned +1 nucleosome plays a central role in TSS selection in the embryo. The absence of a nucleosome-positioning sequence signature, as well as of precise nucleosome positioning at promoters with a canonical TATA- box in other systems [197, 233], together with sharp promoter architecture, argues in favour of the W-box as the overriding determinant of maternal TSS selection. However, to validate the independence of maternal TSS selection on nucleosome positioning, it is necessary to map nucleosome positions in the oocyte, which is currently not plausible due to technical limitations on the number of cells required for a ChIP-seq experiment.

Finally, the extent of functional consequences of the switch in TSS selection remains to be addressed. Although the shift in TSS positions between maternal and zygotic transcriptome is restricted to a narrow region of several dozen bp and happens in both upstream and downstream direction, it still creates variability in the 5’ end sequences of produced transcripts. This variability does not impact the coding portion of the transcript and likely does not have a global functional effect on the transcriptome; however, it might interfere with the sequences in the 5’ untranslated region (UTR), such as micro RNA target sites or mRNA localisation signals in specific transcripts. Identification of such cases might shed light on how specific regulatory mechanisms interact within the global change in transcription initiation mode.

Overlapping transcription initiation codes and differential promoter usage might not be limited to embryonic development, and are possibly a widespread phenomenon occurring in other contexts, such as terminal differentiation. For instance, we have detected a large group of promoters with differential TSS usage during mouse testis development and maturation (Paper IV). Although not necessarily genome-wide and mediated by the switch in the basal transcription machinery, the differential usage of a specific group of promoters important in a certain system might be driven by overlapping codes, which enable their expression in very different regulatory environments. Regulatory contexts are defined by the availability of specific transcription factors and activation of distal-acting enhancers, and are highly dynamic during development and differentiation. Thus, core promoters might need to contain multiple independent determinants that allow them to remain active in the changing regulatory conditions.

The results presented in Papers II and III relied heavily on precisely defined TSSs. Mapping of TSSs at single bp resolution is essential for determining major promoter types, such as sharp and broad promoter architectures, in order to study their sequence and epigenetic features. For instance, precise quantitative mapping of transcription initiation events in Paper III revealed the tight association between dominant TSS in broad promoters and subtle dinucleotide frequency pattern providing the nucleosome positioning signal. Due to a lower resolution, this pattern is not visible using the 5' end positions available in annotation databases (Paper IV).

In Paper II we have shown that promoter usage can be highly dynamic at the TSS level and that promoter width and type is not an inherent property of a genomic locus, but rather a feature that depends on the regulatory context. This further emphasises the importance of using context-specific TSS information in analyses of genome-wide data. To address this, in Paper IV we introduced *CAGEr*, a resource and tool for precise TSS data mining and construction of context-specific promoteromes. It is aimed at facilitating the reuse of CAGE data and introducing it as a more precise and functionally relevant alternative to TSSs from annotation databases. *CAGEr* provides easy access to comprehensive TSS collections for majority of common model systems (cell lines [165] and organisms [161, 162]), directly from within R/Bioconductor environment [221], which is currently the most heavily used platform for genomic data analysis. This enables integration of precise TSS data with other genome-wide data types for promoter-centred analyses. In addition to high-resolution promoter positions, analyses performed with CAGEr provide additional layers of promoter-associated information, including promoter width and dynamics, which allow separation of different functional classes of promoters. The application of *CAGEr* is not limited to CAGE, but can be used with single bp resolution quantitative TSS data derived from other high-throughput technologies, such as oligo-capping or PET. The presented tool and resource should lead to widespread use of precise TSS data in regulatory genomics and help supersede the RefSeq- and Ensembl-based static 5’ end definitions.

5 References

1. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M: **What is a gene, post-ENCODE? History and updated definition.** *Genome Research* 2007, **17**:669–681.
2. White RJ: *Gene Transcription*. Oxford, UK: John Wiley & Sons; 2009.
3. Brenner S, Jacob F, Meselson M: **An unstable intermediate carrying information from genes to ribosomes for protein synthesis.** *Nature* 1961, **190**:576–581.
4. Nissen P, Hansen J, Ban N, Moore PB, Steitz TA: **The structural basis of ribosome activity in peptide bond synthesis.** *Science* 2000, **289**:920–930.
5. Hoagland MB, Stephenson ML, Scott JF, Hecht LI, Zamecnik PC: **A soluble ribonucleic acid intermediate in protein synthesis.** *J Biol Chem* 1958, **231**:241–257.
6. Crick FH: **The origin of the genetic code.** *Journal of molecular Biology* 1968, **38**:367–379.
7. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function.** *Cell* 2004, **116**:281–297.
8. Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP: **Vertebrate microRNA genes.** *Science* 2003, **299**:1540–1540.
9. Meister G, Tuschl T: **Mechanisms of gene silencing by double-stranded RNA.** *Nature* 2004, **431**:343–349.
10. Ernst C, Morton CC: **Identification and function of long non-coding RNA**. *Frontiers in Cellular Neuroscience* 2013, **7**:1–9.
11. Geisler S, Coller J: **RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts.** *Nat Rev Mol Cell Biol* 2013, **14**:699–712.
12. Rinn JL, Chang HY: **Genome regulation by long noncoding RNAs.** *Annual review of biochemistry* 2012, **81**:145–166.
13. Thomas MC, Chiang C-M: **The general transcription machinery and general cofactors.** *Crit Rev Biochem Mol Biol* 2006, **41**:105–178.
14. Lee Y, Kim M, Han J, Yeom K-H, Lee S, Baek SH, Kim VN: **MicroRNA genes are transcribed by RNA polymerase II**. *EMBO J* 2004, **23**:4051–4060.
15. Smale ST, Kadonaga JT: **The RNA polymerase II core promoter**. *Annual review of biochemistry* 2003, **72**:449–479.
16. Orphanides G, Lagrange T, Reinberg D: **The general transcription factors of RNA polymerase II.** *Genes & Development* 1996, **10**:2657–2683.
17. Kadonaga JT: **Perspectives on the RNA polymerase II core promoter.** *Wiley Interdiscip Rev Dev Biol* 2012, **1**:40–51.
18. Maston GA, Evans SK, Green MR: **Transcriptional Regulatory Elements in the Human Genome**. *Annu Rev Genom Human Genet* 2006, **7**:29–59.
19. Gershenzon NI, Ioshikhes IP: **Synergy of human Pol II core promoter elements revealed by statistical sequence analysis.** *Bioinformatics* 2005, **21**:1295–1300.

20. Parry TJ, Theisen JWM, Hsu JY, Wang YL, Corcoran DL, Eustice M, Ohler U, Kadonaga JT: **The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery**. *Genes & Development* 2010, **24**:2013–2018.
21. Gardiner-Garden M, Frommer M: **CpG islands in vertebrate genomes.** *Journal of molecular Biology* 1987, **196**:261–282.
22. Antequera F, Bird A: **Number of CpG islands and genes in human and mouse.** *Proc Natl Acad Sci USA* 1993, **90**:11995–11999.
23. Saxonov S, Berg P, Brutlag DL: **A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters.** *Proc Natl Acad Sci USA* 2006, **103**:1412–1417.
24. Long HK, Sims D, Heger A, Blackledge NP, Kutter C, Wright ML, Grützner F, Odom DT, Patient R, Ponting CP, Klose RJ: **Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates**. *eLife* 2013, **2**.
25. Takai D, Jones PA: **Comprehensive analysis of CpG islands in human chromosomes 21 and 22.** *Proc Natl Acad Sci USA* 2002, **99**:3740–3745.
26. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, Forrest ARR, Alkema WB, Tan SL, 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, et al.: **Genome-wide analysis of mammalian promoter architecture and evolution**. *Nature Genetics* 2006, **38**:626–635.
27. Butler JEF, Kadonaga JT: **The RNA polymerase II core promoter: a key component in the regulation of gene expression.** *Genes & Development* 2002, **16**:2583–2592.
28. Juven-Gershon T, Kadonaga JT: **Regulation of gene expression via the core promoter and the basal transcriptional machinery.** *Developmental Biology* 2010, **339**:225–229.
29. Kadonaga JT: **Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors.** *Cell* 2004, **116**:247–257.
30. Pabo CO, Sauer RT: **Transcription factors: structural families and principles of DNA recognition.** *Annual review of biochemistry* 1992, **61**:1053–1095.
31. Lemon B, Tjian R: **Orchestrated response: a symphony of transcription factors for gene control.** *Genes & Development* 2000, **14**:2551–2569.
32. Narlikar GJ, Fan H-Y, Kingston RE: **Cooperation between complexes that regulate chromatin structure and transcription.** *Cell* 2002, **108**:475–487.
33. Gaston K, Jayaraman PS: **Transcriptional repression in eukaryotes: repressors and repression mechanisms.** *Cell Mol Life Sci* 2003, **60**:721–741.
34. Perissi V, Jepsen K, Glass CK, Rosenfeld MG: **Deconstructing repression: evolving models of co-repressor action.** *Nat Rev Genet* 2010, **11**:109–123.
35. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY: **Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs.** *Cell* 2007, **129**:1311–1323.

36. Zhao J, Sun BK, Erwin JA, Song J-J, Lee JT: **Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome.** *Science* 2008, **322**:750–756.
37. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements**. *Nat Rev Genet* 2004, **5**:276–287.
38. Baumann M, Pontiller J, Ernst W: **Structure and basal transcription complex of RNA polymerase II core promoters in the mammalian genome: an overview.** *Mol Biotechnol* 2010, **45**:241–247.
39. Istrail S, Davidson EH: **Logic functions of the genomic cis-regulatory code.** *Proceedings of the National Academy of Sciences* 2005, **102**:4954–4959.
40. Whitfield TW, Wang J, Collins PJ, Partridge EC, Aldred SF, Trinklein ND, Myers RM, Weng Z: **Functional analysis of transcription factor binding sites in human promoters.** *Genome Biology* 2012, **13**:R50.
41. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM: **Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome.** *Genome Research* 2006, **16**:1–10.
42. Tharakaraman K, Bodenreider O, Landsman D, Spouge JL, Mariño-Ramírez L: **The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site.** *Nucleic Acids Research* 2008, **36**:2777–2786.
43. Vardhanabhuti S, Wang J, Hannenhalli S: **Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation.** *Nucleic Acids Research* 2007, **35**:3203–3213.
44. Smith AD, Sumazin P, Xuan Z, Zhang MQ: **DNA motifs in human and mouse proximal promoters predict tissue-specific expression.** *Proceedings of the National Academy of Sciences* 2006, **103**:6275–6280.
45. Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E: **A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly.** *Human Molecular Genetics* 2003, **12**:1725–1735.
46. Kleinjan DA, Seawright A, Mella S, Carr CB, Tyas DA, Simpson TI, Mason JO, Price DJ, van Heyningen V: **Long-range downstream enhancers are essential for Pax6 expression.** *Developmental Biology* 2006, **299**:563–581.
47. Banerji J, Rusconi S, Schaffner W: **Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences.** *Cell* 1981, **27**:299–308.
48. Ellingsen S, Laplante MA, König M, Kikuta H, Furmanek T, Hoivik EA, Becker TS: **Large-scale enhancer detection in the zebrafish genome.** *Development* 2005, **132**:3799–3811.
49. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM: **In vivo enhancer analysis of human conserved non-coding sequences**. *Nature* 2006, **444**:499–502.
50. Spitz F, Furlong EEM: **Transcription factors: from enhancer binding to developmental control.** *Nat Rev Genet* 2012, **13**:613–626.
51. Lanzuolo C, Roure V, Dekker J, Bantignies F, Orlando V: **Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex.** *Nat Cell Biol* 2007, **9**:1167–1174.

52. Gaszner M, Felsenfeld G: **Insulators: exploiting transcriptional and epigenetic mechanisms.** *Nat Rev Genet* 2006, **7**:703–713.
53. Bell AC, West AG, Felsenfeld G: **The protein CTCF is required for the enhancer blocking activity of vertebrate insulators.** *Cell* 1999, **98**:387–396.
54. Cuddapah S, Jothi R, Schones DE, Roh T-Y, Cui K, Zhao K: **Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains.** *Genome Research* 2009, **19**:24–32.
55. Li Q, Peterson KR, Fang X, Stamatoyannopoulos G: **Locus control regions.** *Blood* 2002, **100**:3077–3086.
56. Tanimoto K, Liu Q, Bungert J, Engel JD: **Effects of altered gene order or orientation of the locus control region on human beta-globin gene expression in mice.** *Nature* 1999, **398**:344–348.
57. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee B-K, Lee K, London D, Lotakis D, Neph S, et al.: **The accessible chromatin landscape of the human genome.** *Nature* 2012, **489**:75–82.
58. Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I: **Predicting tissue-specific enhancers in the human genome.** *Genome Research* 2007, **17**:201–211.
59. Panne D: **The enhanceosome.** *Current Opinion in Structural Biology* 2008, **18**:236–242.
60. Struhl K: **Gene regulation. A paradigm for precision.** *Science* 2001, **293**:1054–1055.
61. Visel A, Rubin EM, Pennacchio LA: **Genomic views of distant-acting enhancers.** *Nature* 2009, **461**:199–205.
62. Kulkarni MM, Arnosti DN: **Information display by transcriptional enhancers.** *Development* 2003, **130**:6569–6575.
63. Visel A, Akiyama JA, Shoukry M, Afzal V, Rubin EM, Pennacchio LA: **Functional autonomy of distant-acting human enhancers**. *Genomics* 2009, **93**:509–513.
64. Sakabe NJ, Savic D, Nobrega MA: **Transcriptional enhancers in development and disease.** *Genome Biology* 2012, **13**:238.
65. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A: **Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq**. *Science* 2013, **339**:1074–1077.
66. Sandelin A, Bailey P, Bruce S, Engström PG, Klos JM, Wasserman WW, Ericson J, Lenhard B: **Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes.** *BMC Genomics* 2004, **5**:99.
67. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJK, Cooke JE, Elgar G: **Highly conserved non-coding sequences are associated with vertebrate development.** *Plos Biol* 2005, **3**:e7.

68. Irimia M, Tena JJ, Alexis MS, Fernandez-Miñan A, Maeso I, Bogdanovic O, la Calle Mustienes de E, Roy SW, Gómez-Skarmeta JL, Fraser HB: **Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints.** *Genome Research* 2012, **22**:2356–2367.
69. Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engstrom PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K, Ghislain J, Pezeron G, Mourrain P, Ellingsen S, Oates AC, Thisse C, Thisse B, Foucher I, Adolf B, Geling A, Lenhard B, Becker TS: **Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates**. *Genome Research* 2007, **17**:545–555.
70. Engström PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B: **Genomic regulatory blocks underlie extensive microsynteny conservation in insects**. *Genome Research* 2007, **17**:1898–1908.
71. Dong X, Navratilova P, Fredman D, Drivenes Ø, Becker TS, Lenhard B: **Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons.** *Nucleic Acids Research* 2010, **38**:1071–1085.
72. Birnbaum RY, Clowney EJ, Agamy O, Kim MJ, Zhao J, Yamanaka T, Pappalardo Z, Clarke SL, Wenger AM, Nguyen L, Gurrieri F, Everman DB, Schwartz CE, Birk OS, Bejerano G, Lomvardas S, Ahituv N: **Coding exons function as tissue-specific enhancers of nearby genes.** *Genome Research* 2012, **22**:1059–1068.
73. Visel A, Bristow J, Pennacchio LA: **Enhancer identification through comparative genomics.** *Semin Cell Dev Biol* 2007, **18**:140–152.
74. Roh T-Y, Wei G, Farrell CM, Zhao K: **Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns.** *Genome Research* 2007, **17**:74–81.
75. Chatterjee S, Bourque G, Lufkin T: **Conserved and non-conserved enhancers direct tissue specific transcription in ancient germ layer specific developmental control genes.** *BMC Dev Biol* 2011, **11**:63.
76. Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS: **Conservation of RET regulatory function from human to zebrafish without sequence similarity.** *Science* 2006, **312**:276–279.
77. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA: **ChIP-seq accurately predicts tissue-specific activity of enhancers**. *Nature* 2009, **457**:854–858.
78. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J: **A unique chromatin signature uncovers early developmental enhancers in humans.** *Nature* 2011, **470**:279–283.
79. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S-I, Cooper GM, Ahituv N, Pennacchio LA, Shendure J: **Massively parallel functional dissection of mammalian enhancers in vivo.** *Nat Biotechnol* 2012, **30**:265–270.
80. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Kinney JB, Kellis M, Lander ES, Mikkelsen TS: **Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay**. *Nat Biotechnol* 2012:1–9.

81. Erceg J, Saunders TE, Girardot C, Devos DP, Hufnagel L, Furlong EEM: **Subtle Changes in Motif Positioning Cause Tissue-Specific Effects on Robustness of an Enhancer's Activity**. *PLoS Genet* 2014, **10**:e1004060.
82. Kolovos P, Knoch TA, Grosveld FG, Cook PR, Papantonis A: **Enhancers and silencers: an integrated and simple model for their function.** *Epigenetics Chromatin* 2012, **5**:1.
83. Mukhopadhyay S, Schedl P, Studitsky VM, Sengupta AM: **Theoretical analysis of the role of chromatin interactions in long-range action of enhancers and insulators**. *Proc Natl Acad Sci* 2011, **19919-19924**:19919–19924.
84. Tolhuis B, Palstra R-J, Splinter E, Grosveld F, de Laat W: **Looping and interaction between hypersensitive sites in the active beta-globin locus.** *Molecular Cell* 2002, **10**:1453–1465.
85. Nolis IK, McKay DJ, Mantouvalou E, Lomvardas S, Merika M, Thanos D: **Transcription factors mediate long-range enhancer-promoter interactions.** *Proc Natl Acad Sci USA* 2009, **106**:20222–20227.
86. Ghavi-Helm Y, Klein FA, Pakozdi T, Ciglar L, Noordermeer D, Huber W, Furlong EEM: **Enhancer loops appear stable during development and are associated with paused polymerase**. *Nature* 2014:1–22.
87. de Wit E, de Laat W: **A decade of 3C technologies: insights into nuclear organization.** *Genes & Development* 2012, **26**:11–24.
88. Sanyal A, Lajoie BR, Jain G, Dekker J: **The long-range interaction landscape of gene promoters.** *Nature* 2012, **489**:109–113.
89. Butler JEF: **Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs**. *Genes & Development* 2001, **15**:2515–2519.
90. Akalin A, Fredman D, Arner E, Dong X, Bryne JC, Suzuki H, Daub CO, Hayashizaki Y, Lenhard B: **Transcriptional features of genomic regulatory blocks.** *Genome Biology* 2009, **10**:R38.
91. Jiang C, Pugh BF: **Nucleosome positioning and gene regulation: advances through genomics.** *Nat Rev Genet* 2009, **10**:161–172.
92. Yuan G-C, Liu Y-J, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ: **Genome-scale identification of nucleosome positions in S. cerevisiae.** *Science* 2005, **309**:626–630.
93. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-Resolution Profiling of Histone Methylations in the Human Genome**. *Cell* 2007, **129**:823–837.
94. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, Tomsho LP, Qi J, Glaser RL, Schuster SC, Gilmour DS, Albert I, Pugh BF: **Nucleosome organization in the Drosophila genome.** *Nature* 2008, **453**:358–362.
95. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, Sidow A, Fire A, Johnson SM: **A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning.** *Genome Research* 2008, **18**:1051–1063.

96. Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF: **A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome.** *Genome Research* 2008, **18**:1073–1083.

97. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A: **Determinants of nucleosome organization in primary human cells.** *Nature* 2011, **474**:516–520.

98. Gaffney DJ, McVicker G, Pai AA, Fondufe-Mittendorf YN, Lewellen N, Michelini K, Widom J, Gilad Y, Pritchard JK: **Controls of Nucleosome Positioning in the Human Genome**. *PLoS Genet* 2012, **8**:e1003036.

99. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, Segal E: **The DNA-encoded nucleosome organization of a eukaryotic genome**. *Nature* 2009, **458**:362–366.

100. Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, Kadonaga JT, Liu XS, Struhl K: **Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo**. *Nat Struct Mol Biol* 2009, **16**:847–852.

101. Kaplan N, Moore I, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, Hughes TR, Lieb JD, Widom J, Segal E: **Nucleosome sequence preferences influence in vivo nucleosome organization**. *Nat Struct Mol Biol* 2010, **17**:918–920.

102. Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, Kadonaga JT, Liu XS, Struhl K: **Evidence against a genomic code for nucleosome positioning**. *Nature Publishing Group* 2010, **17**:920–923.

103. Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang J-PZ, Widom J: **A genomic code for nucleosome positioning**. *Nature* 2006, **442**:772–778.

104. Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, Ferrier P, Spicuglia S, Gut M, Gut I, Andrau JC: **CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters**. *Genome Research* 2012, **22**:2399–2408.

105. de Wit E, Braunschweig U, Greil F, Bussemaker HJ, van Steensel B: **Global chromatin domain organization of the Drosophila genome.** *PLoS Genet* 2008, **4**:e1000045.

106. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *Science* 2009, **326**:289–293.

107. Bird A: **DNA methylation patterns and epigenetic memory.** *Genes & Development* 2002, **16**:6–21.

108. Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker JW, Tian S, Hawkins RD, Leung D, Yang H, Wang T, Lee AY, Swanson SA, Zhang J, Zhu Y, Kim A, Nery JR, Urich MA, Kuan S, Yen C-A, Klugman S, Yu P, Suknuntha K, Propson NE, Chen H, Edsall LE, Wagner U, Li Y, Ye Z, et al.: **Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells**. *Cell* 2013, **153**:1134–1148.

109. Hu S, Wan J, Su Y, Song Q, Zeng Y, Nguyen HN, Shin J, Cox E, Rho HS, Woodard C, Xia S, Liu S, Lyu H, Ming G-L, Wade H, Song H, Qian J, Zhu H: **DNA methylation presents distinct binding sites for human transcription factors.** *eLife* 2013, **2**:e00726.

110. Lande-Diner L, Zhang J, Ben-Porath I, Amariglio N, Keshet I, Hecht M, Azuara V, Fisher AG, Rechavi G, Cedar H: **Role of DNA methylation in stable gene repression.** *J Biol Chem* 2007, **282**:12194–12200.

111. Collings CK, Waddell PJ, Anderson JN: **Effects of DNA methylation on nucleosome stability**. *Nucleic Acids Research* 2013, **41**:2918–2931.

112. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B: **Histone modifications at human enhancers reflect global cell-type-specific gene expression**. *Nature* 2009, **459**:108–112.

113. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R: **Histone H3K27ac separates active from poised enhancers and predicts developmental state.** *Proceedings of the National Academy of Sciences* 2010, **107**:21931–21936.

114. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T-K, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.** *Nature* 2007, **448**:553–560.

115. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B: **Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome**. *Nature Genetics* 2007, **39**:311–318.

116. Schotta G, Lachner M, Sarma K, Ebert A, Sengupta R, Reuter G, Reinberg D, Jenuwein T: **A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin.** *Genes & Development* 2004, **18**:1251–1262.

117. Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, Presser A, Nusbaum C, Xie X, Chi AS, Adli M, Kasif S, Ptaszek LM, Cowan CA, Lander ES, Koseki H, Bernstein BE: **Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains.** *PLoS Genet* 2008, **4**:e1000242.

118. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES: **A bivalent chromatin structure marks key developmental genes in embryonic stem cells.** *Cell* 2006, **125**:315–326.
119. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA: **A chromatin landmark and transcription initiation at most promoters in human cells.** *Cell* 2007, **130**:77–88.
120. Strahl BD, Allis CD: **The language of covalent histone modifications.** *Nature* 2000, **403**:41–45.
121. Rando OJ: **Combinatorial complexity in chromatin structure and function: revisiting the histone code.** *Current Opinion in Genetics & Development* 2012, **22**:148–155.
122. Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, Linder-Basso D, Plachetka A, Shanower G, Tolstorukov MY, Luquette LJ, Xi R, Jung YL, Park RW, Bishop EP, Canfield TK, Sandstrom R, Thurman RE, MacAlpine DM, Stamatoyannopoulos JA, Kellis M, Elgin SCR, Kuroda MI, Pirrotta V, Karpen GH, Park PJ: **Comprehensive analysis of the chromatin landscape in Drosophila melanogaster.** *Nature* 2011, **471**:480–485.
123. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE: **Mapping and analysis of chromatin state dynamics in nine human cell types.** *Nature* 2011, **473**:43–49.
124. The ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **488**:57–74.
125. Ernst J, Kellis M: **ChromHMM: automating chromatin-state discovery and characterization.** *Nature Methods* 2012, **9**:215–216.
126. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, Hardison RC, Dunham I, Kellis M, Noble WS: **Integrative annotation of chromatin elements from ENCODE data.** *Nucleic Acids Research* 2013, **41**:827–841.
127. Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA: **High-throughput functional testing of ENCODE segmentation predictions.** *Genome Research* 2014:gr.173518.114.
128. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, et al.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860–921.
129. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, et al.: **The sequence of the human genome.** *Science* 2001, **291**:1304–1351.

130. Schmutz J, Wheeler J, Grimwood J, Dickson M, Yang J, Caoile C, Bajorek E, Black S, Chan YM, Denys M, Escobar J, Flowers D, Fotopulos D, Garcia C, Gomez M, Gonzales E, Haydu L, Lopez F, Ramirez L, Retterer J, Rodriguez A, Rogers S, Salazar A, Tsai M, Myers RM: **Quality assessment of the human genome sequence.** *Nature* 2004, **429**:365–368.
131. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, et al.: **The genome sequence of Drosophila melanogaster.** *Science* 2000, **287**:2185–2195.
132. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, et al.: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420**:520–562.
133. International Chicken Genome Sequencing Consortium: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432**:695–716.
134. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, McLaren S, Sealy I, Caccamo M, Churcher C, Scott C, Barrett JC, Koch R, Rauch G-J, White S, Chow W, Kilian B, Quintais LT, Guerra-Assunção JA, Zhou Y, Gu Y, Yen J, Vogel J-H, Eyre T, Redmond S, Banerjee R, et al.: **The zebrafish reference genome sequence and its relationship to the human genome.** *Nature* 2013, **496**:498–503.
135. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nature Methods* 2008, **5**:621–628.
136. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing.** *Science* 2008, **320**:1344–1349.
137. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-Wide Mapping of in Vivo Protein-DNA Interactions**. *Science* 2007, **316**:1497–1502.
138. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D, Zhou D, Luo S, Vasicek TJ, Daly MJ, Wolfsberg TG, Collins FS: **Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS).** *Genome Research* 2006, **16**:123–131.
139. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.** *Nature* 2001, **409**:533–538.
140. Horak CE, Snyder M: **ChIP-chip: a genomic approach for identifying transcription factor binding sites.** *Meth Enzymol* 2002, **350**:469–483.

141. Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, Green R, Meltzer PS, Wolfsberg TG, Collins FS: **DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays.** *Nature Methods* 2006, **3**:503–509.

142. Yin C, Yau SS-T: **Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence.** *J Theor Biol* 2007, **247**:687–694.

143. Shakya DK, Saxena R, Sharma SN: **An adaptive window length strategy for eukaryotic CDS prediction.** *IEEE/ACM Trans Comput Biol Bioinform* 2013, **10**:1241–1252.

144. Hannenhalli S: **Eukaryotic transcription factor binding sites – modeling and integrative search methods.** *Bioinformatics* 2008, **24**:1325–1331.

145. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B: **A high-resolution map of active promoters in the human genome.** *Nature* 2005, **436**:876–880.

146. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10**:57–63.

147. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, Yamanaka I, Kiyosawa H, Yagi K, Tomaru Y, Hasegawa Y, Nogami A, Schönbach C, Gojobori T, Baldarelli R, Hill DP, Bult C, Hume DA, Quackenbush J, Schriml LM, Kanapin A, Matsuda H, Batalov S, Beisel KW, Blake JA, Bradt D, et al.: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420**:563–573.

148. Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Kel AE, Arakawa T, Carninci P, Kawai J, Hayashizaki Y, Takagi T, Nakai K, Sugano S: **Large-scale collection and characterization of promoters of human and mouse genes.** *In Silico Biol* 2004, **4**:429–444.

149. Suzuki Y, Tsunoda T, Sese J, Taira H, Mizushima-Sugano J, Hata H, Ota T, Isogai T, Tanaka T, Nakamura Y, Suyama A, Sakaki Y, Morishita S, Okubo K, Sugano S: **Identification and characterization of the potential promoter regions of 1031 kinds of human genes.** *Genome Research* 2001, **11**:677–684.

150. Trinklein ND, Aldred SJF, Saldanha AJ, Myers RM: **Identification and functional analysis of human transcriptional promoters.** *Genome Research* 2003, **13**:308–312.

151. Zhang Z, Dietrich FS: **Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE.** *Nucleic Acids Research* 2005, **33**:2838–2851.

152. Suzuki Y, Taira H, Tsunoda T, Mizushima-Sugano J, Sese J, Hata H, Ota T, Isogai T, Tanaka T, Morishita S, Okubo K, Sakaki Y, Nakamura Y, Suyama A, Sugano S: **Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites.** *EMBO Rep* 2001, **2**:388–393.

153. van Heeringen SJ, Akhtar W, Jacobi UG, Akkers RC, Suzuki Y, Veenstra GJC: **Nucleotide composition-linked divergence of vertebrate core promoter architecture.** *Genome Research* 2011, **21**:410–421.

154. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y: **Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.** *Proc Natl Acad Sci USA* 2003, **100**:15776–15781.

155. Ng P, Wei C-L, Sung W-K, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH, Liu ET, Ruan Y: **Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation.** *Nature Methods* 2005, **2**:105–111.

156. Carninci P, Kvam C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Izawa M, Muramatsu M, Hayashizaki Y, Schneider C: **High-efficiency full-length cDNA cloning by biotinylated CAP trapper.** *Genomics* 1996, **37**:327–336.

157. Takahashi H, Lassmann T, Murata M, Carninci P: **5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing**. *Nature Protocols* 2012, **7**:542–561.

158. de Hoon M, Hayashizaki Y: **Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference.** *BioTechniques* 2008, **44**:627–8– 630– 632.

159. FANTOM Consortium, Suzuki H, Forrest ARR, van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM, Lassmann T, Ravasi T, Hasegawa Y, de Hoon MJL, Katayama S, Schroder K, Carninci P, Tomaru Y, Kanamori-Katayama M, Kubosaki A, Akalin A, Ando Y, Arner E, Asada M, Asahara H, Bailey T, Bajic VB, Bauer D, Beckhouse AG, Bertin N, Björkegren J, Brombacher F, Bulger E, et al.: **The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line.** *Nature Genetics* 2009, **41**:553–562.

160. Valen E, Pascarella G, Chalk A, Maeda N, Kojima M, Kawazu C, Murata M, Nishiyori H, Lazarevic D, Motti D, Marstrand TT, Tang MHE, Zhao X, Krogh A, Winther O, Arakawa T, Kawai J, Wells C, Daub C, Harbers M, Hayashizaki Y, Gustincich S, Sandelin A, Carninci P: **Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE**. *Genome Research* 2009, **19**:255–265.

161. Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C, Booth BW, Zhang D, Wan KH, Yang L, Boley N, Andrews J, Kaufman TC, Graveley BR, Bickel PJ, Carninci P, Carlson JW, Celniker SE: **Genome-wide analysis of promoter architecture in Drosophila melanogaster**. *Genome Research* 2011, **21**:182–192.

162. The FANTOM Consortium and RIKEN Genome Exploration Research Group and Genome Science Group: **The Transcriptional Landscape of the Mammalian Genome**. *Science* 2005, **309**:1559–1563.

163. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, Waki K, Hornig N, Arakawa T, Takahashi H, Kawai J, Forrest ARR, Suzuki H, Hayashizaki Y, Hume DA, Orlando V, Grimmond SM, Carninci P: **The regulated retrotransposon transcriptome of mammalian cells**. *Nature Genetics* 2009, **41**:563–571.

164. Lenhard B, Sandelin A, Carninci P: **Metazoan promoters: emerging characteristics and insights into transcriptional regulation**. *Nat Rev Genet* 2012, **13**:233–245.

165. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, et al.: **Landscape of transcription in human cells**. *Nature* 2012, **488**:101–108.

166. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, Jørgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhata E, Maeda S, et al.: **An atlas of active enhancers across human cell types and tissues**. *Nature* 2014, **507**:455–461.

167. Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project: **Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs**. *Nature* 2009, **457**:1028–1032.

168. Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE, Bussemaker HJ, White KP: **A gene expression map for the euchromatic genome of Drosophila melanogaster.** *Science* 2004, **306**:655–660.

169. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS, Gingeras TR: **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution.** *Science* 2005, **308**:1149–1154.

170. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316**:1484–1488.

171. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, et al.: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447**:799–816.

172. RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) and the FANTOM Consortium: **Antisense Transcription in the Mammalian Transcriptome**. *Science* 2005, **309**:1564–1566.

173. Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME: **Widespread transcription at neuronal activity-regulated enhancers.** *Nature* 2010, **465**:182–187.

174. Berretta J, Morillon A: **Pervasive transcription constitutes a new level of eukaryotic genome regulation.** *EMBO Rep* 2009, **10**:973–982.

175. Burley SK: **The TATA box binding protein.** *Current Opinion in Structural Biology* 1996, **6**:69–75.

176. Hampsey M: **Molecular genetics of the RNA polymerase II general transcriptional machinery.** *Microbiol Mol Biol Rev* 1998, **62**:465–503.

177. Rhee HS, Pugh BF: **Genome-wide structure and organization of eukaryotic pre-initiation complexes**. *Nature* 2012, **483**:295–301.

178. Ponjavic J, Lenhard B, Kai C, Kawai J, Carninci P, Hayashizaki Y, Sandelin A: **Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters.** *Genome Biology* 2006, **7**:R78.

179. Venters BJ, Pugh BF: **Genomic organization of human transcription initiation complexes**. *Nature* 2013, **502**:53–58.

180. Siebert M, Söding J: **Universality of core promoter elements?** *Nature* 2014, **511**:E11–2.

181. Venters BJ, Pugh BF: **Retraction: Genomic organization of human transcription initiation complexes.** *Nature* 2014, **513**:444–444.

182. Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E: **Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters.** *Gene* 2007, **389**:52–65.

183. Smale ST, Baltimore D: **The "initiator" as a transcription control element.** *Cell* 1989, **57**:103–113.

184. Burke TW, Kadonaga JT: **Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters.** *Genes & Development* 1996, **10**:711–724.

185. Burke TW, Kadonaga JT: **The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila.** *Genes & Development* 1997, **11**:3020–3031.

186. Kutach AK, Kadonaga JT: **The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters.** *Molecular and Cellular Biology* 2000, **20**:4754–4764.

187. Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH: **New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB.** *Genes & Development* 1998, **12**:34–44.

188. Deng W, Roberts SGE: **A core promoter element downstream of the TATA box that is recognized by TFIIB.** *Genes & Development* 2005, **19**:2418–2423.

189. Hawkes NA, Roberts SG: **The role of human TFIIB in transcription start site selection in vitro and in vivo.** *J Biol Chem* 1999, **274**:14337–14343.

190. Deng W, Roberts SGE: **TFIIB and the regulation of transcription by RNA polymerase II.** *Chromosoma* 2007, **116**:417–429.

191. Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, Temper V, Razin A, Cedar H: **Sp1 elements protect a CpG island from de novo methylation.** *Nature* 1994, **371**:435–438.

192. Blake MC, Jambou RC, Swick AG, Kahn JW, Azizkhan JC: **Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter.** *Molecular and Cellular Biology* 1990, **10**:6632–6641.

193. Hochheimer A: **Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression**. *Genes & Development* 2003, **17**:1309–1320.

194. Veenstra GJ, Wolffe AP: **Gene-selective developmental roles of general transcription factors.** *Trends in Biochemical Sciences* 2001, **26**:665–671.

195. Akhtar W, Veenstra GJC: **TBP2 is a substitute for TBP in Xenopus oocyte transcription.** *BMC Biol* 2009, **7**:45.

196. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, Fields S, Stamatoyannopoulos JA: **Global mapping of protein-DNA interactions in vivo by digital genomic footprinting**. *Nature Methods* 2009, **6**:283–289.

197. Rach EA, Winter DR, Benjamin AM, Corcoran DL, Ni T, Zhu J, Ohler U: **Transcription Initiation Patterns Indicate Divergent Strategies for Gene Regulation at the Chromatin Level**. *PLoS Genet* 2011, **7**:e1001274.

198. Raisner RM, Hartley PD, Meneghini MD, Bao MZ, Liu CL, Schreiber SL, Rando OJ, Madhani HD: **Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin.** *Cell* 2005, **123**:233–248.

199. Jin C, Zang C, Wei G, Cui K, Peng W, Zhao K, Felsenfeld G: **H3.3/H2A.Z double variant-containing nucleosomes mark "nucleosome-free regions" of active promoters and other regulatory regions.** *Nature Genetics* 2009, **41**:941–945.

200. Vermeulen M, Mulder KW, Denissov S, Pijnappel WWMP, van Schaik FMA, Varier RA, Baltissen MPA, Stunnenberg HG, Mann M, Timmers HTM: **Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4**. *Cell* 2007, **131**:58–69.

201. Megraw M, Pereira F, Jensen ST, Ohler U, Hatzigeorgiou AG: **A transcription factor affinity-based code for mammalian transcription initiation.** *Genome Research* 2009, **19**:644–656.

202. Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao Y, Ohler U, Zhu J: **A paired-end sequencing strategy to map the complex landscape of transcription initiation**. *Nature Methods* 2010, **7**:521–527.

203. Rach EA, Yuan H-Y, Majoros WH, Tomancak P, Ohler U: **Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome**. *Genome Biology* 2009, **10**:R73.

204. Azuara V, Perry P, Sauer S, Spivakov M, Jørgensen HF, John RM, Gouti M, Casanova M, Warnes G, Merkenschlager M, Fisher AG: **Chromatin signatures of pluripotent cell lines.** *Nat Cell Biol* 2006, **8**:532–538.

205. Barrera LO, Li Z, Smith AD, Arden KC, Cavenee WK, Zhang MQ, Green RD, Ren B: **Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs.** *Genome Research* 2008, **18**:46–59.

206. Ramirez-Carrozzi VR, Braas D, Bhatt DM, Cheng CS, Hong C, Doty KR, Black JC, Hoffmann A, Carey M, Smale ST: **A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling.** *Cell* 2009, **138**:114–128.

207. Roider HG, Lenhard B, Kanhere A, Haas SA, Vingron M: **CpG-depleted promoters harbor tissue-specific transcription factor binding signals--implications for motif overrepresentation analyses.** *Nucleic Acids Research* 2009, **37**:6305–6315.

208. Ohler U: **Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction.** *Nucleic Acids Research* 2006, **34**:5943–5950.

209. Ayoubi TA, Van de Ven WJ: **Regulation of gene expression by alternative promoters**. *The FASEB Journal* 1996, **10**:1–8.

210. Davuluri RV, Suzuki Y, Sugano S, Plass C, Huang TH-M: **The functional consequences of alternative promoter use in mammalian genomes.** *Trends Genet* 2008, **24**:167–177.

211. Hansen SK, Tjian R: **TAFs and TFIIA mediate differential utilization of the tandem Adh promoters.** *Cell* 1995, **82**:565–575.

212. Ren B, Maniatis T: **Regulation of Drosophila Adh promoter switching by an initiator-targeted repression mechanism.** *EMBO J* 1998, **17**:1076–1086.

213. Mizuguchi G, Kanei-Ishii C, Sawazaki T, Horikoshi M, Roeder RG, Yamamoto T, Ishii S: **Independent control of transcription initiations from two sites by an initiator-like element and TATA box in the human c-erbB-2 promoter.** *FEBS Lett* 1994, **348**:80–88.

214. Trinklein ND: **An Abundance of Bidirectional Promoters in the Human Genome**. *Genome Research* 2004, **14**:62–66.

215. Engström PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, Lavorgna G, Brozzi A, Luzi L, Tan SL, Yang L, Kunarso G, Lian-Chong Ng E, Batalov S, Wahlestedt C, Kai C, Kawai J, Carninci P, Hayashizaki Y, Wells CA, Bajic VB, Orlando V, Reid JF, Lenhard B, Lipovich L: **Complex Loci in Human and Mouse Genomes**. *PLoS Genet* 2006, **2**:e47.

216. Kawaji H, Frith MC, Katayama S, Sandelin A, Kai C, Kawai J, Carninci P, Hayashizaki Y: **Dynamic usage of transcription start sites within core promoters.** *Genome Biology* 2006, **7**:R118.

217. Mathavan S, Lee SGP, Mak A, Miller LD, Murthy KRK, Govindarajan KR, Tong Y, Wu YL, Lam SH, Yang H, Ruan Y, Korzh V, Gong Z, Liu ET, Lufkin T: **Transcriptome analysis of zebrafish embryogenesis using microarrays.** *PLoS Genet* 2005, **1**:260–276.

218. Valen E, Preker P, Andersen PR, Zhao X, Chen Y, Ender C, Dueck A, Meister G, Sandelin A, Jensen TH: **Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes.** *Nature Publishing Group* 2011, **18**:1075–1082.

219. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, DiCuccio M, Kitts P, Maglott DR, Murphy TD, Ostell JM: **RefSeq: an update on mammalian reference**

sequences. *Nucleic Acids Research* 2014, **42**(Database issue):D756–63.

220. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kähäri AK, Keenan S, Kulesha E, Martin FJ, Maurel T, McLaren WM, Murphy DN, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, et al.: **Ensembl 2014.** *Nucleic Acids Research* 2014, **42**(Database issue):D749–55.
221. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biology* 2004, **5**:R80.
222. The R Development Core Team: **R: A Language and Environment for Statistical Computing**. *R Foundation for Statistical Computing* 2013:1–3079.
223. Dooley K, Zon LI: **Zebrafish: a model system for the study of human disease.** *Current Opinion in Genetics & Development* 2000, **10**:252–256.
224. Gehrig J, Reischl M, Kalmár É, Ferg M, Hadzhiev Y, Zaucker A, Song C, Schindler S, Liebel U, Müller F: **Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos.** *Nature Methods* 2009, **6**:911–916.
225. Majumder S, DePamphilis ML: **TATA-dependent enhancer stimulation of promoter activity in mice is developmentally acquired**. *Molecular and Cellular Biology* 1994, **14**:4258–4268.
226. Davis W, Schultz RM: **Developmental change in TATA-box utilization during preimplantation mouse development.** *Developmental Biology* 2000, **218**:275–283.
227. Raz E: **Primordial germ-cell development: the zebrafish perspective**. *Nat Rev Genet* 2003, **4**:690–700.
228. Rose CM, van den Driesche S, Meehan RR, Drake AJ: **Epigenetic reprogramming: preparing the epigenome for the next generation.** *Biochem Soc Trans* 2013, **41**:809–814.
229. Deato MDE, Tjian R: **Switching of the core transcription machinery during myogenesis**. *Genes & Development* 2007, **21**:2137–2149.
230. Müller F, Zaucker A, Tora L: **Developmental regulation of transcription initiation: more than just changing the actors**. *Current Opinion in Genetics & Development* 2010, **20**:533–540.
231. Gazdag E, Santenard A, Ziegler-Birling C, Altobelli G, Poch O, Tora L, Torres-Padilla ME: **TBP2 is essential for germ cell development by regulating transcription and chromatin condensation in the oocyte**. *Genes & Development* 2009, **23**:2210–2223.
232. Bártfai R, Balduf C, Hilton T, Rathmann Y, Hadzhiev Y, Tora L, Orbán L, Müller F: **TBP2, a Vertebrate-Specific Member of the TBP Family, Is Required in Embryonic Development of Zebrafish**. *Current Biology* 2004, **14**:593–598.
233. Nozaki T, Yachie N, Ogawa R, Kratz A, Saito R, Tomita M: **Tight associations between transcription promoter type and epigenetic variation in histone positioning and modification.** *BMC Genomics* 2011, **12**:416.