Analysis of transcription factor complexes and their associated chromatin interactions during erythropoiesis

SUPAT THONGJUEA

Dissertation for the degree philosophiae doctor (PhD)

at the University of Bergen

2014

Dissertation date: 17.01.2014

Scientific environment

This work has been performed in Computational Biology Unit (CBU), Uni Computing

and

SARS International Centre for Marine Molecular Biology.

Supat Thongjuea is affiliated with

Department of Molecular Biology

Acknowledgements

I have had the great honor to work with fantastic people during the past four years of my doctoral program. I would like to express my sincere gratitude to all those who provided me the guidance, support, and wishes for the completion of my study.

A special gratitude I give to my supervisor, Dr. Boris Lenhard. Boris gave me a chance to firstly work as a scientific programmer for a year before taking me as his PhD student with challenging projects. I would like to thank for his patience, motivation, support, brilliant ideas, and excellent guidance.

I would like to thank Professor Rein Aasland for being my co-supervisor and his interest in my work. I would also like to thank Professor Inge Jonassen for his support.

Many thanks to our past group members: Jan Christian, Chirag, Xianjun, Vanja, Sara, Yogita, Altuna, Christopher, Gemma, Reidar, Chandu, Ying, Vedran, and David. Thanks for all of you for creating such a great environment in the group for not just only work but also for fun and friendship. I would also like to thank for all new group members in Imperial College London and special thanks to Nathan Harmston for guidance and edit my thesis.

I would like to thank to all CBU and SARS members for creating pleasant working environment and special thanks to the computer system administrators who always help us to fix a problem of our servers. I would also like to thank administration people especially Monika Voit for her help in everything.

I would like to thank Grosveld's lab members for the nice collaboration: Frank Grosveld, Eric Soler, Charlotte Andrieu-Soler, Athina Mylona, and Ralph Stadhouders.

Many thanks to past and present Thai friends in Bergen: Bass, Aom, Yiam, May, Yong, Tangmo, Angsuma, Issie, Jun, Tidty, Olarn, Puriphat, Mine, Chalermchart,

Nasun, and Lek Samrit. I also would like to thank Professor Pongsri Brudvik for her support. It has been difficult to live so far away from home but all of you made such pleasant Thai environment in Bergen as stay at home.

Finally, thanks to my parents, brothers, and all family members for unconditional love and support. I would like to thank my love, Kib, for always being beside me. Thanks for your endless patience, support, and love. I love you all very much.

Abstract

Cis-regulatory elements can control gene transcription over large distances. Studying the association between *cis*-regulatory elements and the mechanism of gene regulation is important since the disruption in the transcriptional control can lead to developmental defects and many complex diseases. This thesis describes the identification and characterization of transcription factor complexes and their associated *cis*-regulatory regions during erythropoiesis. These include the dynamics of protein complexes and their associated binding patterns, the chromatin interactions between them and their target promoter at specific loci, and the involvement of identified *cis*-regulatory complexes in gene regulation during erythroid cell differentiation.

Transcription factor occupancies of multiple key factors were profiled using ChIP-seq during erythroid development, in proerythroblast-like cells and in fully differentiated erythrocyte-like cells. These factors consist of proteins involved in the LDB1 complex (LDB1, GATA1, SCL/TAL1, ETO2, and MTGR1) and several other factors that are critical for gene regulation such as RUNX1, GFI1B, FOG1, LSD1, LMO2, LMO4, P300, TIF1γ, CTCF, and RNAPII. To analyze ChIP-seq data, a set of bioinformatics tools were developed that expanded on the functionality of existing R/Bioconductor packages. These tools provide functions for data processing, manipulation, mining and visualization, thus greatly facilitating hypothesis generation and interpretation of experimental results. Integrative analysis enabled the identification of how protein complexes change during differentiation and how identified complexes affect gene regulation. We show that the dynamics of the LDB1 complex compositions, in particular, the binding of the co-repressor ETO2 and MTGR1 significantly decreases during differentiation. This dynamic LDB1 complex occurs at a very specific subset of genes that are induced late during erythroid differentiation, suggesting the role of LDB1 as an activation complex. Using computational techniques we discovered twelve distinct patterns of P300 binding complexes. These binding patterns showed distinct characteristics and could be classified into enhancer, promoter, and insulator-associated classes. Integrating the

identified binding patterns with associated gene expression profiles demonstrated the central roles of P300 TF complexes in both activating and repressing target genes.

Finally, an integrative approach using multiple ChIP-seq data sets together with 3C-seq has been used to demonstrate the long-range regulation of regulatory elements and their associated chromatin contacts at the β-globin and *Myb* loci. To analyze 3C-seq data, I developed a R/Bioconductor package called *r3Cseq* to facilitate 3C-seq data analyses. Using *r3Cseq,* we demonstrated the importance of long-range chromatin contacts. We showed that the dynamics of long-range chromatin interactions between TF binding sites and the associated target promoters is involved in the transcriptional control of β-globin and *Myb* genes.

In summary, this thesis work is primarily concerned with the development of computational methods and tools for the analysis of large-scale experimental data, with an emphasis on data generated by ChIP-seq and 3C-seq technologies, and the application of these tools to generating new hypotheses and acquiring new biological knowledge on mammalian gene regulation.

List of publications included in the thesis

The thesis contains the following articles (* denotes my equal first author contribution). They will be referred using their roman numerals in the text.

I. Soler, E., Andrieu-Soler, C., de Boer, E., Bryne, J.C., **Thongjuea, S.**, Stadhouders, R., Palstra, R.J., Stevens, M., Kockx, C., van Ijcken, W., Hou, J., Steinhoff, C., Rijkers, E., Lenhard, B. and Grosveld, F. (2010) **The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation**. *Genes & development*, **24**, 277-289.

II. Portales-Casamar, E.*, **Thongjuea, S.***, Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W. and Sandelin, A. (2010) **JASPAR 2010: The greatly expanded open-access database of transcription factor binding profiles**. *Nucleic acids research*, **38**, D105-110.

III. Stadhouders, R.*, **Thongjuea, S.***, Andrieu-Soler, C., Palstra, R.J., Bryne, J.C., van den Heuvel, A., Stevens, M., de Boer, E., Kockx, C., van der Sloot, A., van den Hout, M., van Ijcken, W., Eick, D., Lenhard, B., Grosveld, F. and Soler, E. (2012) **Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development**. *The EMBO journal*, **31**, 986-999.

IV. **Thongjuea, S.***, Stadhouders, R.*, Grosveld, F.G., Soler, E. and Lenhard, B. (2013) **r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data**. *Nucleic acids research*, **41**, e132.

V. **Thongjuea, S.**, Sumić, S., Andrieu-Soler, C., de Boer, E., Stadhouders, R., van Ijcken, W., Soler, E., Grosveld, F. and Lenhard, B. (2013) **Genome-wide dynamics of P300 transcription factor complexes during erythroid differentiation**. (Manuscript)

List of other publications

During my doctoral studies, I contributed to the following publications:

I. Ribeiro de Almeida, C., Stadhouders, R., de Bruijn, M.J., Bergen, I.M., **Thongjuea, S.**, Lenhard, B., van Ijcken, W., Grosveld, F., Galjart, N., Soler, E. et al. (2011) **The DNA-binding protein CTCF limits proximal Vkappa recombination and restricts kappa enhancer interactions to the immunoglobulin kappa light chain locus**. *Immunity*, **35**, 501-513.

II. Ribeiro de Almeida, C., Stadhouders, R., **Thongjuea, S**., Soler, E. and Hendriks, R.W. (2012) **DNA-binding factor CTCF and long-range gene interactions in V(D)J recombination and oncogene activation**. *Blood*, **119**, 6209-6218.

III. Ghamari, A., van de Corput, M.P., **Thongjuea, S**., van Cappellen, W.A., van Ijcken, W., van Haren, J., Soler, E., Eick, D., Lenhard, B. and Grosveld, F.G. (2013) **In vivo live imaging of RNA polymerase II transcription factories in primary cells**. *Genes & development*, **27**, 767-777.

IV. Mylona, A.*, Andrieu-Soler, C.*, **Thongjuea, S.***, Martella, A., Soler, E., Jorna, R., Hou, J., Kockx, C., van Ijcken, W., Lenhard, B. and Grosveld, F. (2013) **Genome-wide analysis shows that Ldb1 controls essential hematopoietic genes/pathways in mouse early development and reveals novel players in hematopoiesis**. *Blood*, **121**, 2902-2913.

Table of Contents

List of abbreviations

3C	Chromosome conformation capture
4C	Chromosome conformation capture on chip or circular chromosome conformation capture
5C	Chromosome conformation carbon-copy
ACH	Active chromatin hub
BRE	TFIIB recognition element
bp	Base pairs
CAGE	Cap analysis of gene expression
CGI	CpG island
ChIP	Chromatin immunoprecipitation
CNE	Conserved noncoding element
CRM	*Cis*-regulatory module
CTCF	CCCTC-binding factor
DCE	Downstream core element
DHS	DNaseI hypersensitive site
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
DNA-FISH	DNA fluorescence *in situ* hybridization
DPE	Downstream promoter element
ENCODE	Encyclopedia of DNA elements
EuTRACC	European transcriptome, regulome and cellular commitment consortium
FAIRE	Formaldehyde-assisted isolation of regulatory element
FANTOM	Functional annotation of the mammalian genome
GEO	Gene expression omnibus
GTF	General transcription factor
INR	Initiator element
ICR	Imprinting control region
Kb	Kilo base pairs
LCR	Locus control region
lincRNA or lncRNA	long (intergenic) non-coding RNA

NCBI	National center for biotechnology information
MEL	Mouse erythroleukaemic
MTE	Motif ten element
PBM	Protein-binding microarray
PcG	Polycomb group
PCR	Polymerase chain reaction
PIC	Transcription pre-initiation complex
PRC	Polycomb-repressive complex
PRE	Polycomb response element
PWM	Position weight matrix
RNAPII:	RNA polymerase II
SELEX	Systematic evolution of ligands by exponential enrichment
SNP	Single nucleotide polymorphism
TF	Transcription factor
TFBS	Transcription factor binding site
TIC	Transcription initiation complex
TSS	Transcription start site
XCI	X-chromosome inactivation

1. General introduction

1.1 Transcriptional regulation in mammalian genomes

Genes are transcribed to protein-coding and non-coding RNAs. The process of gene transcription is usually referred to as gene expression. Regulated gene expression is essential for the development and differentiation from a single fertilized cell to numerous different cell types, which are required for the development of tissues and the whole organism [1-3].

In eukaryotes, transcription takes place within the nucleus of a cell, and is initiated by one of the three RNA polymerase enzyme complexes. RNA Polymerase II (RNAPII) synthesizes many protein-coding and many non-coding RNAs (including most microRNAs) [4,5]. Transcription is initiated when general transcription factors (GTFs), such as TFIIA-H, assemble on DNA at a region known as a core promoter in order to recruit RNAPII onto sequence-specific DNA binding sites. The core promoter contains a transcription start site (TSS), and other crucial core DNA promoter elements. These elements may include: a TATA box, the downstream promoter element (DPE), TFIIB recognition elements (BREs), downstream core elements (DCEs), the motif ten elements (MTE), and the initiator element (INR). Importantly, each of these elements is found in only a fraction of core promoters with different combinations and is rarely present all together in one particular promoter [6-10].

GTFs form a complex with the core promoter, termed the transcription pre-initiation complex (PIC). The formation of PIC leads to the recruitment of RNAPII and the subsequent formation of a transcription initiation complex (TIC). At this stage, the TIC is sufficient to direct low levels of transcription, generally referred to as basal transcription [11] – at least at a subset of promoters [12].

Transcription is highly context-specific and modulated by a set of *cis*-acting elements and *trans*-acting factors that either stimulate or repress expression. Several *trans*-

acting factors, termed activators, interact with components of RNAPII transcriptional machinery to stimulate transcription [13,14]. Activators stimulate PIC assembly by recruiting other non-DNA binding proteins, known as co-activators, via protein-protein interactions. There are several types of co-activators, which are classified depending on their distinct biological and biochemical functions. Some types of co-activators promote accessibility between GTFs and activators [15]. Other types are involved in the formation of chromatin remodeling complexes, which modify chromatin structure and restructure nucleosomes around the core promoter (as well as in other regulatory regions). These co-activators facilitate the binding GTFs to the core-promoter and other essential DNA elements that are otherwise protected by nucleosomes and the compactness of heterochromatin [16].

Another type of *trans*-acting factors, termed *repressors*, inhibit transcription through one of several different mechanisms (as reviewed in [17-19]): (1) they may directly inhibit transcription of the basal transcription machinery by interacting with core subunits of GTFs protein complex, either resulting in modifying the TIC or inhibiting the binding of GTFs to DNA; (2) they may regulate activators by competing to bind DNA at the same specific DNA-binding regions as activators; or they may bind to activators via protein-protein interactions, thus preventing activators from interacting with their targets; (3) they may bind to specific binding sites called insulators located between promoters and *cis*-activating elements known as enhancers, thus preventing communication between them; or (4) they may recruit other non-DNA binding proteins termed co-repressors, forming chromatin-remodeling complexes to deacetylate histone tails, to methylate nucleotides, and to recruit other repressive nucleosome and chromatin-remodeling complexes, such as a Polycomb-repressive complexes (PRC1 and PRC2), to certain regions. In addition, other gene repression mechanisms have recently been discovered. For instance, long (intergenic) non-coding RNAs (lincRNAs or lncRNAs) have been shown to play critical roles in transcriptional repression of *HOX* gene clusters and X chromosome inactivation [20,21]. Further more, nucleosomes restrict the accessibility of the protein complexes to DNA to ensure that transcription can be activated in the correct context [22].

Activators and repressors bind to several types of *cis*-acting DNA elements with specific sequence constraints, called transcription factor binding sites (TFBSs). Since

TFBSs have critical roles in the regulation of gene expression, most of the work presented in this thesis concentrates on the identification and characterization of these elements. *Cis*-acting elements can be found in different classes of regulatory regions, including promoters, enhancers, silencers, insulators, and locus control regions (LCRs) (Figure 1). The following sections will provide basic descriptions of these regulatory regions, their structure and function in gene regulation.

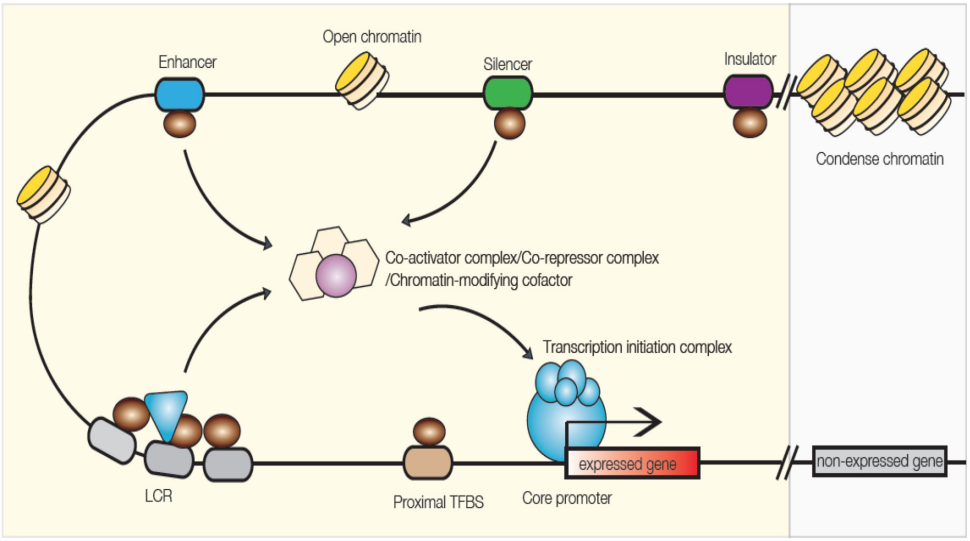

Figure 1. The typical location for each type of regulatory elements (proximal or distal), relative to the transcription start site. These elements are proximal or distal by distinct TFs that may recruit other cofactors to stabilize or destabilize the transcription-initiation machinery or to modify chromatin structure. Insulators prevent the spread of condensed chromatin through the formation of a 'barrier' to allow the accessible DNA for the binding of TFs. The condensed chromatin restricts the accessibility of the transcriptional activation complexes to DNA to ensure that transcription can only be initiated at the correct context. The figure is modified from the original model of Figure1 in [23].

1.1.1 Promoters

Promoters are regulatory regions located in front of the transcribed part of genes. There are two types of promoter sub-regions which can distinguished based on the distance from a TSS and their motif content: (1) core promoters, which are regions

located within ~100 bp around TSSs containing essential DNA elements (i.e. TATA box), and are targeted by GTFs to form the PIC complex; and (2) proximal promoter regions, which are further away from the TSS, but generally limited to a few hundred bp upstream, and contain essential TFBSs that mediate regulatory inputs by binding activators or repressors that play critical roles in context-specific gene expression [24,25]. In vertebrates, many core promoters are known to be associated with CpG islands (CGIs), which are genomic regions that contain a higher frequency of CG dinucleotides [26-28] than most of the rest of the genome. The association between core promoters and CGI frequency can be used to classify core promoters into two groups according to their GC content and CpG frequency: (1) high-CG promoters, associated with CGIs and are generally associated with housekeeping and developmental genes; (2) low-CG promoters, associated with low CG content, many of which have a TATA box, and are associated with genes involved in tissue-specific transcription [28-31].

A genome-wide analysis of mouse and human promoters using cap analysis of gene expression (CAGE) classified mammalian promoters into “sharp” and “broad” promoters [9]. Sharp promoters have a precisely defined TSS position, with the dominant peak located at a restricted distance from TATA box, and is overrepresented in low-CG promoters. In contrast, broad promoters have multiple TSSs or broad TSS clusters and are generally associated with high-CG promoters. CGIs seem to be tightly associated with broad promoters in the initial mammalian promoter studies. However, a recent analysis of promoters looking at non-vertebrate promoters, in *Drosophila*, which does have TATA box promoters but does not have CGIs, found both sharp and broad TSS clusters, which were associated with the same functional classes of genes as in mammalian genomes. These findings suggest that high-CG promoters are not a requirement for broad TSS. Instead, distinct patterns of nucleosome positioning and histone modifications were found to be associated with sharp and broad promoters in both human and *Drosophila* [32], implying that these epigenetic marks may be a more precise means to distinguish sharp and broad promoters than the presence of CGIs (reviewed in [10]).

Proximal promoters are genomic regions located upstream of the core promoters and are known to harbor sets of multiple TFBSs known as *cis*-regulatory modules (CRMs)

[33]. CRMs integrate the activity of many transcription factors (TFs) involved in cellular communication during tissue development as well as allowing the preservation of specific expression patterns in terminally differentiated tissues [24,25,34]. In mammals, TFBSs experimentally identified in proximal promoter regions are highly enriched in low-CG promoters, whereas high-CG promoters have a lower density of TFBSs [25]. Several studies demonstrated that DNA motifs found in proximal promoters in mammalian genomes were capable of successfully predicting tissue-specific gene expression patterns [24,35]. Computational and chromatin immunoprecipitation (ChIP)-based approaches have been developed to identify and to characterize these elements, in order to help understand gene regulation in higher organisms.

1.1.2 Enhancers and silencers

Enhancers [36-38] and silencers [39,40] are the *cis*-regulatory elements located further away from the TSS compared to those in the proximal promoter. The distance between these elements and their target promoters is extremely variable, ranging from a few hundred bp up to megabase distances. One of the best known enhancers that acts from an extreme distance is the enhancer for the mouse sonic hedgehog (*Shh*) gene, located within the intron of *Lmbr1* gene, which is positioned several hundreds kb away from the *Shh* promoter [41]. The location of enhancers and silencers can be observed upstream, downstream, within exons and introns of the target gene, and even within the gene body of neighboring genes [36-38,41,42] and they are often, but not always, conserved across species [43-46].

Traditionally, enhancers and silencers are detected by their ability to drive or reduce the expression of a reporter gene after transferring a DNA construct containing them next to a reporter gene into transgenic organisms or into cells in culture [47]. In gene transfer assays, they typically regulate their target genes in a distance- and orientation-independent manner [47]. Similar to proximal promoters, enhancers and silencers may contain either individual or multiple TFBSs for tissue- or cell-specific binding of activators and repressors [48-50]. It is thought that distant enhancers regulate their target promoters via chromatin looping; it has been shown that distant regulatory elements of the active β-globin locus can be positioned in close spatial

proximity to gene promoters while intervening sequences are looped out [51,52]. Similar to enhancers, it has been shown that the Polycomb response elements (PREs), which are silencers and are the binding target of PRC1/2, interact with promoters to initiate a higher-order chromatin structure to maintain the compactness of chromatin at repressed genes [53,54].

The advent of chromosome conformation capture (3C)-based technologies [54], has led to the generation of interaction profiles between regulatory elements and their corresponding promoters, which supports the chromatin loop hypothesis and its involvement in gene regulation. By using 3C-based technologies, several genes, described in Section 1.4, have been demonstrated to be under long-range regulation in different cell types over various stages of development and cell differentiation.

1.1.3 Insulators

Insulators are *cis*-regulatory elements, typically ~0.5-3 kb in length [55,56], located throughout the genome [57]. Characterization of insulators using transgenic constructs has revealed two roles of insulators: (1) the ability to block communication between adjacent regulatory elements, for example between an enhancer and a promoter, in a position-dependent and orientation-independent manner known as "enhancer-blocking" [58-61]; and (2) the ability to prevent the spread of heterochromatin through the formation of a "barrier" [62,63]. One of the best characterized insulator elements is a 42 bp fragment located at the 5' end of the chicken β-globin locus and its orthologous element located at the 5' end of the human β-globin gene [61,64]. Insulators have also been identified within an imprinting control region (ICR) located ~5 kb upstream of the noncoding *IGR2/H19* locus and found to be involved in its regulation [65,66]. These elements are bound by CCCTC-binding factor (CTCF), an ubiquitously expressed protein with 11 zinc finger motifs, which is highly conserved from *Drosophila* to human [67,68]. Studies of vertebrate insulators using enhancer-blocking transgenic assays has revealed that CTCF is universally required for the enhancer-blocking activity of insulators [69].

The first genome-wide analysis of CTCF binding sites in human genome identified ~14,000 CTCF binding sites in primary human fibroblast cells [70]. This number of CTCF binding sites is similar to that observed when using computational approaches involving the identification of conserved noncoding elements CNEs [71]. These two studies showed that: (1) the distribution of CTCF binding sites in human is nonrandom and strongly correlates with gene locations; (2) the identified human consensus motif of CTCF binding sites, as well as many individual CTCF binding sites, are highly conserved across mammals; and (3) the binding location of CTCF is largely invariant across different cell types, setting CTCF apart from many well-characterized factors that bind DNA in a cell type-specific manner. A recent analysis of human CTCF binding sites in various cell types using ChIP-seq also showed that CTCF binding sites overlap extensively between cell types; it further showed that regions occupied by the repressive histone mark H3K27me3 and active regions marked by H2AK5ac are separated by CTCF, suggesting the importance of the chromatin barrier activity of insulators [72]. Using 3C-based technologies, CTCF binding sites have been shown to participate in mediating chromatin interactions in several selected gene regions; for instance: (1) CTCF mediates the formation of an active chromatin hub (ACH) to promote coordinated transcription of β-globin genes throughout differentiation [73]; (2) the contacts between DNA-bound CTCF proteins play a critical role in mediating the formation of chromatin loops necessary for regulation at the imprinted *IGF2/H19 locus* [74]; and (3) CTCF binding sites restrict the interaction specificity of κ enhancer elements to Igκ locus, which is essential for the regulation of V(D)J recombination [75]. Furthermore, a study of the CTCF-associated chromatin interactome in mouse embryonic stem (ES) cells [76] has suggested that CTCF can function as a genome organizer in several ways: (1) by creating a local chromatin hub to facilitate coordinated gene expression, (2) by promoting physical contacts between enhancers and their corresponding promoters, and (3) by establishing boundary structures that can restrict the spreading of the silenced nature of the nuclear lamina into the neighboring regions, or vice versa. Moreover, 10 well-positioned nucleosomes were demonstrated to locate on either side of a CTCF binding site and these nucleosomes are much better positioned than those located near TSSs [77], suggesting that CTCF has a role in influencing nucleosome

occupancy since the binding of CTCF provides an anchor point for positioning nucleosomes.

1.1.4 Locus control regions

A LCR is a cluster of *cis*-regulatory elements that plays critical roles in the regulation of a single gene locus or a gene cluster. The first identified LCR was found at the human β-globin locus [78]. The β-globin LCR is located over 25 kb upstream of the gene cluster, and consists of a group of five DNaseI hypersensitive sites that overlap a cluster of TFBS. This LCR specifies a high-level of erythrocyte expression to the genes within the locus in position-independent manner [78]. A more recent study found that LCR activity at the β-globin locus is orientation-dependent; altering the orientation of the LCR drastically altered the transcription of β-globin genes [79]. Several LCRs have been identified (reviewed in [80]) for example: *CD2* gene [81], *TH2* cytokine locus [82], T cell receptor *α/δ* locus [83], Immunoglobulin heavy chain locus [84], and *CD4* gene [85]. These identified LCRs can also strongly enhance the expression level of their linked genes in a tissue-specific and copy number-dependent manner. The mechanism by which LCRs control their linked genes has been demonstrated using 3C, showing that the β-globin LCR is in contact with its corresponding promoter when the gene is expressed [51]. This chromatin looping mechanism may play an important role in establishing open chromatin domains [51]. However, a recent study demonstrated that deletion of endogenous DNaseI hypersensitive sites of the β-globin LCR impaired the recruitment of factors required for efficient transcriptional elongation of β-globin genes, suggesting a new role of the LCR in β-globin gene transcriptional control [86].

1.2 Genome-wide identification of transcription factor binding sites

As described above, *cis*-regulatory elements are the binding targets of TFs, which often recruit other proteins, and these protein-DNA complexes are involved in gene regulation. In general, TFs recognize and bind DNA sequences at specific recognition sites called transcription factor binding sites (TFBSs). There are several approaches for the genome-wide detection of TFBSs. These approaches can be broadly divided into computational approaches, which focus on using DNA sequence-based analysis, and experimental approaches, which focus on the development of ChIP- and DNaseI HS-based technologies.

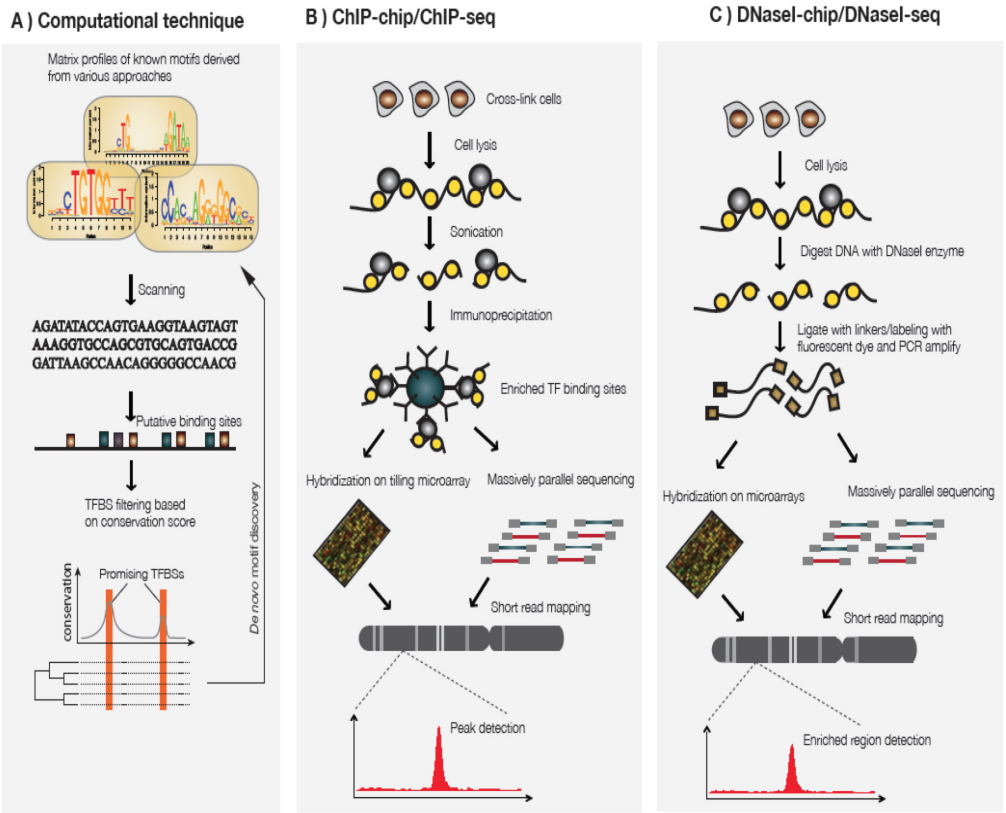

Figure 2. Methods used to identify TFBSs genome-wide. Binding sites can be identified (A) by computational techniques combined with evolutionary sequence conservation, (B) by using direct assays, including ChIP-chip and ChIP-seq, or (C) by using indirect techniques, such as DNaseI-chip and DNaseI-seq. The figure is modified from Figure1 in [87].

1.2.1 Genome-wide identification of *cis*-regulatory elements using computational approaches

The classical approach to identify putative *cis*-regulatory elements is to match genomic DNA sequences with known DNA binding motifs. This method has been used extensively for the prediction of TFBSs (Figure 2A). Known motifs were derived [88,89] from various sources, e.g. systematic mutagenesis and nested deletion experiments, systematic evolution of ligands by exponential enrichment (SELEX) [90], protein-binding microarray (PBM) [91], and ChIP-based technologies [92,93]. In the past several years, ChIP-based genome-wide methods have been used to generate a large number of representative target sequences, from which highly accurate profiles can be derived [94]. Motifs derived from well-characterized TFs, typically represented by position weight matrices (PWMs), have been compiled in databases such as JASPAR [89] and TRANSFAC [88]. The PWM describes the log-likelihood of finding each nucleotide at each position in the DNA motif compared to a chosen background sequence model. It is derived using the nucleotide frequency from the target sequences of known binding sites. To predict TFBSs from input genomic sequences, several programs such as MatInspector [95], MATCH [96], MAST [97], TFBS Perl modules [98], and the matchPWM function from Biostrings (Bioconductor package) [99] can be used to scan a set of PWMs against an input genomic sequence, and return a list of potential TFBSs with the matching score of the matches between the query PWMs and the input sequence. This method, especially for short motifs, often predicts a large number of matches with many false positives [23].

To reduce the number of false positives, one can limit detection only to putative TFBSs with high PWM matching score that are found in regions exhibiting high evolutionary conservation (described below). The latter approach is known as phylogenetic footprinting [97,98], which relies on the evolutionary conservation of genomic sequences over multiple species. Using this approach, functional TFBSs can be detected when orthologous sequences from distantly related species are aligned, thus identified conserved regions will be selected as putative TFBSs. Several methods and tools from different research groups have been developed to facilitate the identification of TFBSs using phylogenetic footprinting, e.g. FootPrinter [100], and

various methods for estimating evolutionary conservation such as phastCons [101], and phyloP [102] can be used for the purpose.

Comparative genomic approaches have been successfully applied for the detection of *cis*-regulatory elements [44,45,103]. However, detection using these methods is limited since the relationship between conservation and function is imperfect. It is clear that a high proportion of experimentally determined binding sites for many factors are not conserved, e.g. (1) 95% of OCT4 (POU5F1) binding sites are not conserved between human and mouse ES cells [104]; and (2) most candidate heart enhancers are not highly conserved in vertebrate evolution [105]. In addition, a comparative ChIP-seq analysis of several factors in five different vertebrates suggested that most binding sites are species-specific, and aligned binding events present in all species are rare [106,107], although it remains to be seen how much of this lineage-specific binding is functional. In contrast, conserved regions do not necessarily retain function [57,87]. Taken together, using computational approaches alone is not sufficient to identify functional *cis*-regulatory elements. For that reason, a number of experimental approaches, described in the following sections have been extensively used for the detection of functional *cis*-regulatory elements.

1.2.2 Genome-wide chromatin immunoprecipitation (ChIP)-based approaches

The most efficient tool to identify *cis*-regulatory elements directly is to use chromatin immunoprecipitation. ChIP allows the isolation and identification of genomic fragments occupied by transcription factors, modified histones, methylated cytosines, or any other molecular feature detectable by antibodies, *in vivo* [108]. ChIP-based approaches were developed more than a decade ago [93,109,110] and have been widely used and extensively reviewed elsewhere [111-114]. Briefly (Figure 2B), cells are cross-linked with a chemical agent, usually formaldehyde. This cross-linked material is randomly sheared by sonication or enzymatic digestion to reduce the fragment size to ~200-600 bp. A specific antibody against a protein of interest (e.g. a transcription factor, a modified histone, and RNA polymerase) is then used to immunoprecipitate nucleoprotein complexes containing the protein of interest. Alternatively, if a specific antibody is not available against the endogenous protein, the direct knock-in approach can be used to introduce small Flag-epitope tags to the

target TFs. Ectopically expressed epitope-tagged proteins can then be generated in cell lines and can be precipitated using commercially available anti-Flag [115-117]. After extensive washing of the immuno-enriched products, cross-links are reversed and DNA fragments associated with the protein of interest are purified and subjected to the size selection typically in the range of ~100-300 bp in preparation for ChIP-chip or ChIP-seq analysis. For ChIP-chip, the ChIP sample and its control sample (input DNA or a non-specific antibody IgG) are labeled with fluorescent dyes and hybridized on microarrays. Protein binding sites can be identified based on the comparison of detected signal between ChIP and its control using ChIP-chip peak-calling software [118,119]. For ChIP-seq, the bound DNA sequencing libraries of the ChIP sample and its control are created and subsequently subjected to the massively parallel sequencing. One of several peak-calling programs [120-123] can then be applied to detect protein-binding sites or to detect regions enriched for histone modifications.

A large number of studies regularly report the use of ChIP-based approaches. As of June 2013, the gene expression omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/), a database repository of high throughput data maintained by the National Center for Biotechnology Information (NCBI), reports ~1,807 ChIP-chip and ~1,837 ChIP-seq experiments (status, taken from http://www.ncbi.nlm.nih.gov/geo/summary/). Recently, the Encyclopedia of DNA elements (ENCODE) consortium project carried out hundreds of ChIP-seq experiments derived from various cell types from human and mouse [124]. ENCODE organizes all of identified binding sites, their corresponding derived motifs, and other associated histone modification patterns for each TF in the FactorBook [125], a public resource for storing the results of the on-going analysis of the ChIP-seq data generated by ENCODE. ENCODE also provides guidelines and practical advice for conducting ChIP-seq experiments based on their experience [126]. The number of ChIP-seq experiments has increased rapidly over ChIP-chip, becoming the dominant method for TFBS detection. There are many reasons why the number of ChIP-seq experiments has increased rapidly (reviewed in [113,114]): (1) since it uses the massively parallel sequencing its resolution is higher than ChIP-chip and does not suffer from the noise generated by the hybridization step in ChIP-chip, (2) its resolution is not limited by the probe sequences fixed on array; and (3) its detected signal has a greater dynamic range compared to ChIP-chip.

ChIP-based approaches are powerful and have been widely used to identify protein-binding sites and enriched regions of modified histones. However, their success depends on the availability of a highly specific antibody. The quality of an antibody is important since e.g. 25% of 200 antibodies suitable for other immunochemical methods (e.g. Western blotting) (tested in *Drosophila*) failed in specificity tests and 20% failed in chromatin immunoprecipitation experiments [127]. The alternative-epitope-tagging approach can be used in cell lines or in model organisms, but might not reflect the endogenous protein binding activity since overexpression of the epitope fusion-protein might cause spurious binding. In addition, ChIP-based approaches require a large number of cells, typically ~10^7 cells yielding 10-100 ng of DNA. Cell types that have small populations such as stem or cancer-initiating cells may limit the application of these approaches due to the limited number of cells. Recently, modified ChIP-seq protocols, Nano-ChIP-seq [128] and LinDA [129], have been developed to address this problem. Nano-ChIP-seq has been successfully carried out using 10,000 cells for histone modifications, whereas LinDA has been successfully applied using 5,000 cells for the estrogen receptor-α transcription factor and 10,000 cells for the H3K4me3 histone modification. Increasing the resolution to precisely detect binding sites is another concern. The resolution of ChIP-based approaches, focusing on ChIP-seq, depends on the library size preparation typically ~100-300 bp long. This much wider than the specific binding position of TFs, which mostly bind only ~6-20 bp of DNA [89]; this limited resolution cannot distinguish binding among closely spaced neighboring sites. ChIP-exo [130], which uses lambda (λ) phage exonuclease to digest the 5' end of protein-bound DNA fragments, has successfully achieved near single-base resolution. In addition, ChIP-exo has been demonstrated to diminish erroneous and missed calls from unbound DNA contaminates [130].

ChIP-based approaches profile the genome-wide binding of a single protein of interest. Therefore, to reveal the dynamics of typical multi-protein regulatory complexes it is necessary to profile binding sites derived from multiple proteins over different conditions. Novel insights can be gained by the integrative analysis of such multiple data sets. For example, the combination of 13 transcription factors revealed that specific combinatorial patterns of binding sites are associated with ES cell-specific transcription circuitry [131]. ChIP-seq analysis of ten key TFs in

hematopoiesis showed uncharacterized combinatorial interactions between transcription factors (SCL/TAL1, LYL1, LMO2, GATA2, RUNX1, ERG and FLI1) in hematopoietic progenitors/stem cell 7 (HPC-7) [132]. These studies propose critical roles for the combination of protein complexes that bind DNA at the same binding sites in the transcriptional control of stem/progenitor-like cells. Since these studies focused on the identification of multiple factors in a specific stage, they cannot be used to explain how protein binding intensity and differences in the combinations of protein complexes can affect transcriptional control during cell development and differentiation process. To this end, a genome-wide study of the dynamics of LDB1 complex [133] (LDB1, GATA1, SCL/TAL1, MTGR1, and ETO2) in mouse erythroleukaemic (MEL) cells before and after erythroid differentiation has shown that the binding intensity of different factors changes during differentiation. Importantly, the repressive factors, ETO2 and MTGR1, showed significant decreases in both binding intensity and the number of binding sites toward MEL cell differentiation [133], indicating that LDB1 complexes acquire stronger gene activation potential toward erythroid maturation. Additionally, integrative analyses of multiple data sets generated by the ENCODE consortium have provided new insights into the mechanisms of gene regulation in the human genome, which can be used to infer the link between mutations in regulatory elements and human diseases [124,134,135].

Finally, protein-binding information alone cannot provide a complete understanding of gene regulation. Integrative analyses, combining protein binding information with various data types, such as gene expression, chromatin conformation, nucleosome positioning, and histone modifications in a various cell types and in specific tissue types are among the key data types to accomplish the understanding of the gene regulation program. Currently, large data sets generated by large-scale consortia like ENCODE, the functional annotation of the mammalian genome (FANTOM), the Blueprint epigenome projects, and the NIH roadmap epigenomics mapping project have become increasingly challenging for data analysis. The challenge is the large requirements for data storage and computational power. Importantly, converting raw data to biologically relevant hypotheses, interpretations and conclusions requires human expertise.

1.3 Genome-wide DNaseI hypersensitivity approaches

DNaseI hypersensitivity assay can detect open chromatin regions that often correspond to nucleosome-depleted regions, which are often associated with binding of regulatory proteins. The DNaseI hypersensitivity assay was developed over 30 years ago and has been widely used to identify genomic regions that are sensitive to cleavage by DNaseI enzyme [136,137]. Conventionally, DNaseI hypersensitive sites (DHSs) have been detected by subjecting isolated nuclei to DNaseI treatment. DNA fragments from DNaseI-digested chromatin are subjected to detect DHSs with the standard Southern blot assay. After this method was developed, hundreds of DHSs were identified, revealing the locations of all types of *cis*-regulatory elements including promoters, enhancers, silencers, insulators and locus control regions [137-139].

With the advent of high-throughout microarray technology, DNase-chip and DNase-array were developed to map DHSs genome-wide [140,141]. DNase-chip (Figure 2C) detects DHSs based on tagging biotinylated linkers to DNaseI-digested fragments, breaking them into smaller size fragments (~200-500 bp), isolating them with the streptavidin, labeling them with fluorescent dye, and hybridizing them on high-density oligonucleotide microarrays [140]. DNase-array relies on two or more cuts per DHS made by DNaseI treatment, releasing smaller fragments of DHSs. The DNA is then isolated by phenol-chloroform-isoamyl and a sucrose gradient is used to isolate appropriately sized fragments. Selected DHSs fragments are then labeled with fluorescent dye, and hybridized on microarrays.

These methods were initially used to detect DHSs across 1% of the human genome for the ENCODE pilot project. More than 2,500 DHSs were identified and were found to fall most often within evolutionarily conserved, and gene-rich regions; some of them were cell-type specific [140]. The majority of them occurred more than 10 kb away from annotated genes. Surprisingly, more than 80% of them formed large DHS super-clusters separated by 100-500 kb implying the existence of higher-order organization of chromatin structure [140,141]. Another study of DNase-chip enabled the detection of ~4000 DHSs in 6 human cell lines [142]. 22% of the DHSs identified

in this study were ubiquitously present across cell lines and a large number of them were bound by CTCF. In addition, a large number of cell-type specific DHSs were enriched for enhancers and were correlated with both cell type-specific gene expression and cell-type specific histone modifications [142].

With the advent of massively parallel sequencing technology, DNaseI microarray-based assays have being replaced by the high-throughput sequencing-based technology. DNase-seq, adapted from DNase-chip [143], was first used to identify ~95,000 DHSs in human primary $CD4^+$ T cells [143]. DNase-seq showed higher sensitivity detection of DHSs over DNase-chip assays. Thus, more studies have used DNase-seq to detect DHSs from various cell types in diverse organisms. In *Drosophila*, DNase-seq was used to profile DHSs and explore the changes of chromatin landscape during five stages of embryonic development [144]. In mammals, ~70,000 DHSs were identified in male and female mouse livers and 1,284 DHSs of them showed robust sex-related differences [145]. Recently, ENCODE identified ~2.9 million DHSs that were derived from 125 diverse cell and tissue types in human, with an integrative analysis revealing novel relationships between chromatin accessibility, transcription, DNA methylation, and regulatory factor occupancy patterns [146]. In addition, correlated locations of distal DHS with their target promoters enabled systematic pairing of different classes of distal DHSs and specific promoter types [146].

The sensitivity of the DNaseI hypersensitivity assay may be limited because the specific regulatory complexes, bound at each open chromatin site, could affect the ability of DNaseI to cut or formaldehyde to crosslink [147]. Therefore, using only one particular technique to detect nucleosome-depleted regions is not sufficient. Integrating DNaseI hypersensitivity with other open-chromatin region detection techniques such as formaldehyde-assisted isolation of regulatory elements (FAIRE) assay (FAIRE-chip/FAIRE-seq) [148], and Sono-seq (a variant method of FAIRE-seq) [149] is essential to enable high-confidence comprehensive detection of open chromatin regions [147]. Although open chromatin regions identified using these approaches cannot determine the particular types of *cis*-regulatory elements, integrating the results with data from other approaches, such as ChIP-seq, can be used to further classify open chromatin regions to each type of *cis*-regulatory element.

1.4 Studying long-range chromatin interactions using chromosome conformation capture (3C)-based technologies

During the past decades, experimental techniques have been developed to study genome architecture and function. One of these techniques, DNA fluorescence *in situ* hybridization (DNA-FISH) [150] has been used as an important tool to allow the identification of some features of spatial genome organization (reviewed in [151-154]). DNA-FISH offers the measurement at the single cell level, which is the greatest advantage of this technique. However, it is limited by its resolution and the severe treatment of chromosomes during sample preparation. Thus, the need for new, minimally invasive experimental techniques to analyze the spatial organization of the genome at high resolution is essential. A considerable portion of this thesis is focused on the application of 3C coupled with massively parallel sequencing (3C-seq) to investigate long-range gene regulation during erythropoiesis. In this section, the various techniques for investigating chromosome capture are described in more detail.

1.4.1 Chromatin Conformation Capture (3C)

The 3C assay was developed by Job Dekker and colleagues in 2002 [155] to analyze the spatial organization of chromatin interactions at high resolution, originally in yeast. The principle of 3C is outlined in Figure 3A. Isolated cells are treated with a cross-linking agent to preserve *in vivo* nuclear proximity between DNA sequences. DNA isolated from these cells is then digested using a primary restriction enzyme, typically a 6-base pair cutting enzyme such as HindIII, BglII, EcoRI or BamHI. A more frequent cutter such as DnpII can also be used for studying smaller loci. The digested DNA products are then ligated under diluted conditions to favor intra-molecular over inter-molecular ligation events. This digested and ligated chromatin yields composite sequences representing (distal) genomic regions that are in close physical proximity in the cell nucleus. The cross-linking in digested and ligated chromatin are then reversed and individual ligation products are detected and quantified by the polymerase chain reaction (PCR) using locus-specific primers.

The initial study by Dekker *et al.* [155] generated a matrix of interaction frequencies to determine the average three-dimensional conformation of yeast chromosome III, demonstrating the shape of a contorted ring of chromosomes in *vivo* [155]. Afterwards, the 3C technique has been applied to detect physical interactions between regulatory elements and their targets in mammalian cells. For instance, 3C was used to demonstrate the looped conformation between the β-globin gene and the LCR that is specific to erythroid cells where the gene is expressed [51].

The 3C technique has also been used to demonstrate physical chromatin interactions at other genomic loci, both within and between chromosomes. For instance, in *cis*, Murrell and colleague used 3C to demonstrate that differentially methylated regions (DMRs) of the imprinted *IGF2* gene and noncoding *H19* gene interact in mice [156]. These interactions described the maternal and paternal allele epigenetic switch for *IGF2* gene when it moved between an active and a silent chromatin domain. In addition, the T helper type2 (T_H2) cytokine gene loci has also been demonstrated to physically interact with its activator LCR, located ~120 kb from the cluster of cytokine coding genes [157]. Furthermore, erythroid-specific DHSs located within the introns of *C16ORF35* physically contact with α -globin target genes [158].

Inter-chromosomal or *trans*-interactions have also been demonstrated by 3C. The initial study investigating *trans*-interactions reported that in murine naïve $CD4^+$ T cells, the *IFNG* gene located on chromosome 10 strongly interacts with the DHSs of the T_H2 LCR located on chromosome 11. The interactions are greatly reduced after the differentiation of naïve T cells to T_H1 or T_H2 cells, whereas non-T cell types did not exhibit these interactions [159]. Another example is the *H* enhancer element located ~75 kb upstream of *MOR28*, one of the odorant cluster receptor genes on chromosome 14, associates with multiple odorant gene promoters on multiple chromosomes. The interaction between the *H* element and an individual odorant gene promoter regulates the expression of the gene [160]. These studies suggested the importance of inter-chromosomal interactions and proposed how the deletion of genetic elements on one chromosome can affect the expression of target genes located on other chromosomes.

3C technology has also demonstrated the role of insulators [59]. As described in the previous section, insulators are DNA elements that have ability to protect a gene or cluster of genes from the influence of neighboring *cis*-regulatory elements. The cellular functions of insulators are dependent on the spatial organization within the nucleus. Using 3C, CTCF was demonstrated to mediate the formation of an ACH to promote the coordinated transcription of β-globin genes during erythroid differentiation [73]. The most obvious example of well characterized CTCF contacts is the mechanism of gene regulation at the imprinting *IGF2/H19* locus. The ICR located ~5 kb upstream of the noncoding *H19* locus has been shown to physically interact with DMRs (DMR1 and DMR2), located around the *IGF2* gene [74]. These contacts are different on the two alleles. On the maternal allele, ICR is bound by CTCF and forms multiple chromatin loops to prevent physical interactions between the *IGF2* gene and its enhancers. On the other hand, on the paternal allele, the ICR is methylated therefore inhibiting the binding of CTCF. Consequently, the *IGF2* gene is able to interact with its enhancers to ensure that the *IGF2* gene is expressed only from the paternal allele.

3C technology has revolutionized the study of spatial chromosome organization and chromatin interactions at specific loci. However, the original technique is still low throughput and hypothesis-driven, since it relies on locus-specific PCR primers and can only be used to interrogate chromatin interactions between preselected sequences. 3C is suitable to detect interactions when target regions are close in the linear chromosomal distance. However, 3C PCR becomes unreliable when the target interacting regions are separated over distances more than a few hundred kb away from each other. For that reason, genome-wide scale methods such as microarray and high-throughput sequencing have been leveraged to develop methods to adapt 3C for generation of high-throughput, high resolution chromatin interaction profiles. A number of high throughput 3C-based methods were developed for the purpose, including: (1) 3C-on-chip and circular 3C (4C)[161,162], a genome-wide scale method to identify interactions involving a specific fragment of choice (a 'viewpoint'), (2) 3C-carbon-copy (5C) [163,164], the identification of interactions with many viewpoints within a confined genomic region, and (3) Hi-C [165], the

identification of interactions between all genomic sites. The summarized description for each 3C-based method and its application is reviewed in [166].

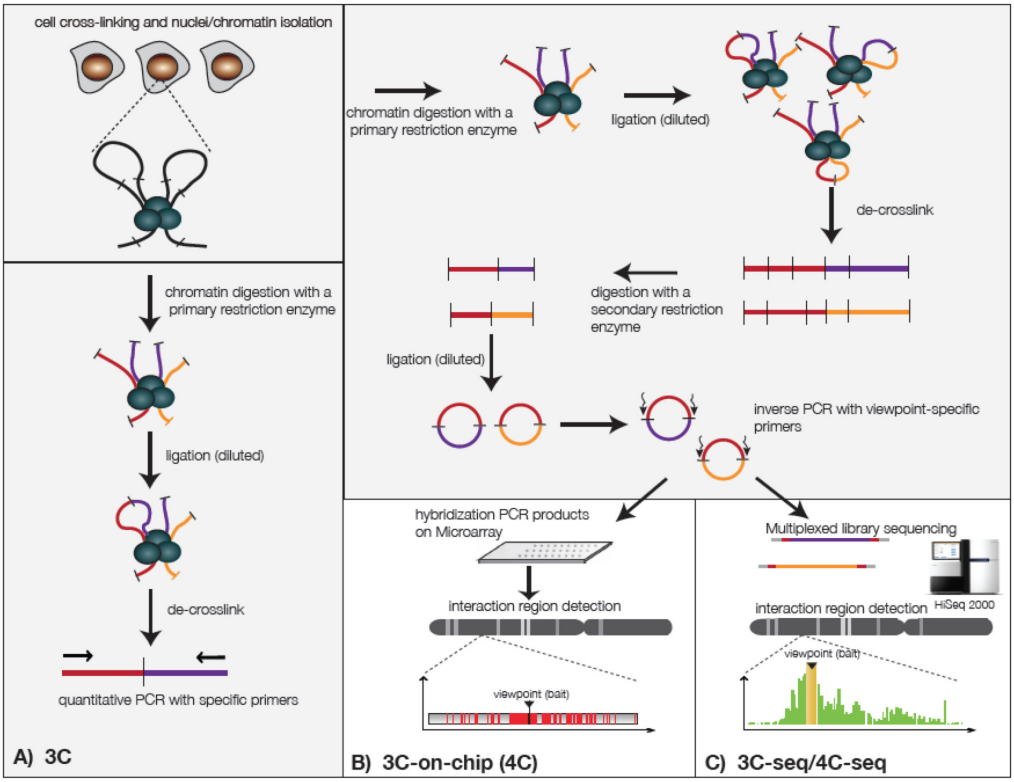

Figure 3. The experimental techniques use to identify chromatin interactions consisting of : (A) 3C, (B) 3C-on-chip (4C), and (C) 3C-seq/4C-seq. 3C is a classical method that is suitable to detect interactions when target regions are close in the linear chromosomal template. 4C/3C-seq/4C-seq are genome-wide scale methods use to identify interactions involving a specific fragment of choice (a 'viewpoint'). The figure is redrawn based on [161,166,167].

1.4.2 4C and 3C-seq/4C-seq

4C is an acronym for chromosome conformation capture on chip (3C-on-chip) or circular chromosome conformation capture (circular 3C). 4C has been developed to identify interactions between all of genomic fragments present on the array and a

selected genomic region ("viewpoint") [161,162]. These two variants of 4C have slightly different protocols. The circular 3C requires circular intra-molecule ligation, where ligation is carried out to form a circle such that both ends of the viewpoint are ligated to both ends of any interacting fragment. In contrast, in 3C-on-chip the ligation is carried out such that only one end of the viewpoint fragment is required to ligate with one end of the interacting fragment.

The protocol for 4C is very similar to 3C (Figure 3B, 3C). Isolated cells are fixed and cross-linked. DNA is isolated from these cells and is digested using a primary restriction enzyme, typically a 6-base pair cutting enzyme such as HindIII, EcoRI or BamHI. The digested products are ligated under diluted conditions and are then de-crosslinked. The de-crosslinked products are subjected to a second restriction digest using secondary restriction enzyme to decrease the fragment sizes (e.g. Nla III or Dpn II). The digested DNA is then ligated again under diluted conditions, creating small circular fragments. These fragments are inverse PCR-amplified using primers specific for any genomic region of interest (e.g. promoter, enhancer, or any other element potentially involved in long-range interactions). The amplified fragments are then purified and either hybridized to a microarray (3C-on-chip), or subjected to massively parallel sequencing (3C-seq/4C-seq). 3C-seq uses the same protocol as 4C-seq, but was developed by different laboratories [133,168]. Although they are the same protocol, 3C-seq normally uses 6-base pair cutter enzymes for both primary and secondary restriction, whereas 4C-seq uses 4-base pair cutters for both restrictions.

4C was initially used to demonstrate that active and inactive genes are engaged in many long-range intra- and inter-chromosomal interactions. 4C analysis of the β-globin gene, using the LCR hypersensitive site 2 (HS2) as the viewpoint, demonstrated significant differences in chromatin contacts between the LCR and the β-globin gene in fetal liver, where the gene is actively transcribed, compared to fetal brain, where the gene is not transcribed [161]. Interaction profiles demonstrated that the β-globin gene in fetal liver interacts with other active regions in both *cis* and *trans* when it is actively transcribed, whereas the silent β-globin gene in fetal brain prefers to contact other inactive regions. This observation suggests that the chromatin conformation of the β-globin locus is tissue-specific.

Significant differences in chromatin conformation between tissue types have also been demonstrated at the *Myb* locus. *Myb,* encoding the c-Myb transcription factor, is a key hematopoietic regulator required for maintaining a proper balance between erythroid cell proliferation and differentiation [169-171]. Many reports have shown that intergenic regions located between *Myb* and *Hbs1l* spanning ~110 kb are the target binding sites of the key hematopoietic TFs [133,172,173]. These binding sites may represent distal regulatory elements involved in the regulation of *Myb* expression. 3C-seq analysis using *Myb* promoter as the viewpoint revealed high interaction signals between *Myb* promoter and the binding sites of TFs complex in fetal liver, which expresses high levels of *Myb*, while interactions in fetal brain, which expresses low levels of *Myb*, were relatively low or absent [174]. This result suggests that the regulatory regions interact with the promoter of the *Myb* gene in tissue-specific manner. This study also demonstrated dynamic changes in the spatial organization of the locus during erythroid differentiation. 3C-seq was used to demonstrate that the ACH formed by the contacts between *Myb* promoter and TFs complex binding sites is destabilized toward differentiation. The enhancers are no longer in contact with the *Myb* gene, in agreement with the loss of TF occupancy at the binding sites, and a loss of transcriptional activity of the locus [174].

HOX genes are sequentially transcribed during the body patterning development in vertebrates [175]. The changes of chromatin modifications and structures regulate the collinear gene expression pattern of these developmental genes [176]. Recent 4C-seq analyses have demonstrated dynamic changes in the chromatin conformation of the *HOX* gene cluster (*HOXA* to *HOXD*) [177]. Interaction profiles generated using multiple genes as viewpoints have revealed comparable domains of contacts in forebrain cells, where all *HOX* genes were inactive. The viewpoints of these inactive genes mostly interacted with random genomic regions with no specific contacts and overlapped with domains of the repressive histone mark H3K27me3. On the contrary, 4C-seq analyses of anterior trunk and posterior trunk cells showed bimodal interaction profiles between the viewpoints and their associated genomic contacts. These interactions separated gene clusters into two distinct chromatin compartments [177]. These results suggested that *HOX* genes associate into a single three-dimensional (3D) chromatin structure when genes in the cluster are inactive. Once genes are

sequentially active during body patterning development, new activated genes progressively migrate from the inactive chromatin structure and cluster into a transcriptionally active 3D chromatin organization that can be demonstrated by the bimodal shape of interaction profile for each progressively active *HOX* gene.

The preference of physical contact between inactive regions seems to be common, since there are some indications that Polycomb group (PcG) proteins can be involved in mediating long-range chromatin interactions in nuclear space. Using 4C, some PREs were shown to interact over long distances to mediate gene silencing in *Drosophila* [53]. In mammals, a repressive chromatin hub formed by multiple chromatin loops, has been demonstrated to be responsible for the silencing of *GATA4* [54]. This chromatin structure was mediated and maintained by PcG. These observations suggest that long-range chromatin interactions among PcG binding domains are a general phenomenon. 4C interaction profiles revealed that PcG target genes prefer to interact with other PcG targets. However, these interactions are constrained by the chromosome architecture since the long-range interactions between PcG targets occur almost exclusively on the same chromosome arm [178]. Therefore, distinct territories of the chromosome structure may limit contact between two chromosome arms [178]. This phenomenon can be observed from other 4C and Hi-C studies in mammals, where intra-chromosomal interactions occur more frequently than inter-chromosomal contacts [161,165,174,177].

4C has also been used to address whether chromatin interactions are a cause or a consequence of gene activity. Many studies confirmed that active genes dynamically co-localize into the nuclear space, termed “transcription factories” [179-181]. These active genes move into or out of these factories resulted in the changes of chromatin conformation structure [180-182]. A modification of 4C protocol, e4C (enhanced ChIP-4C), which includes an additional RNAPII at Ser-5 phosphorylation ChIP enrichment step, was developed to investigate the co-localization in the nuclear space of actively transcribed erythroid responsive genes *Hba* and *Hbb* within transcription factories [180]. This study demonstrated that mouse globin genes preferentially interact with hundreds of other transcribed genomic loci in transcription factories and that the majority of potential globin genes contacts occur with genomic regions from *trans*-chromosomes. These observations suggest that globin genes and other active

genomic regions are moved to the transcription factories with a varied subset of their preferred transcriptional partners (TFs) to facilitate their transcriptional activation, resulting in changes in chromosome conformation. However, various transcription inhibition studies often fail to show a significant influence of chromosome conformational changes [183,184] on gene expression.

4C-seq has been applied to investigate the differences in chromatin conformation between the active and inactive X chromosome in female cells [185]. In mammals, female has two X chromosomes. One X chromosome per cell undergoes X-chromosome inactivation (XCI), in order not to have twice as many X chromosome gene products as in male. Random XCI is initiated by the upregulation of the noncoding *Xist* gene during early embryonic development [186]. During the accumulation of *Xist* RNA, various silencing-related factors are recruited to target silence regions in the X chromosome. These factors consist of PcG protein complexes PRC1 and PRC2; the recruitment of the latter leads to the deposition of H3K27me3 [187]. Thus, most genes on the inactive form of X chromosome are silenced. 4C-seq analyses using multiple gene loci as viewpoints were used to demonstrate in both active and inactive X chromosomal forms. In the active X chromosome, active genes *MeCP2* and *Jarid1C* interact with each other and share a distinct set of interactions with other active genes in both *cis* and *trans*, whereas the inactive genes *Slitrk4* and *Pcdh11x* interact with each other and share a set of interactions with other inactive loci. In the inactive X chromosome, the majority of genes showed a near complete loss of interactions except a small subset that can escape XCI, such as the *Jarid1C* gene. This gene showed many specific interactions with regions on other chromosomes and interacts with other identified genes that escape XCI [185]. These results suggested that the chromatin conformation of active and inactive X chromosomes is completely different. Deletion of *Xist* in the inactive X chromosome showed a dramatic reduction of H3K27me3, whereas the interactions regained from the depletion of *Xist* located on the inactive X chromosome seem to have similar interactions as those on the active X chromosome [185].

4C-based approaches are suitable for studying genomic interacting partners of any given genomic region of interest such as promoter, enhancer, or any other *cis*-regulatory or structural element. The analyses of 4C and 3C-seq/4C-seq for studying

chromatin interactions at specific loci can vary depending on the type of question to be answered (as described in the previous examples). However, these approaches are limited to the detection of only "one versus all" strategy, which is not sufficient to explain the chromatin structure at complex loci. Other technologies such as 5C and Hi-C can be used to address this limitation of 4C-based methods. The resolution of 4C-based approaches is dependent on the restriction enzyme used. A 6-cutter restriction enzyme is typically used in the 4C protocol and cuts once every few kb. A more frequent cutter such as a 4-cutter can potentially increase the resolution but may also increase the noise and decrease statistical power per fragment, requiring windowing approaches to establish significant interactions [168]. 4C-based approaches normally capture high signals of interactions near the viewpoint because DNA sequences near the viewpoint have an increased chance of being non-specifically captured during chromatin cross-linking and digestion. Therefore, interaction signals near the viewpoint are biased and may not be sufficiently reliable for the detection of interactions involving regions that are located very close to the viewpoint. During the data analysis, one has to take the nature of this bias into account to quantify detectable interaction signals. Other potential biases come from a number of steps that can locally differ in efficiency. These biases could be due to differences in the efficiency of crosslinking, restriction digestion, ligation, PCR amplification, and microarray hybridization or sequencing. Some of the biases are relatively common to other high-throughput protocols. Thus, to reduce the biases, biological and/or technical replicates and robust data analysis methods are required to ensure reproducible biological interpretation.

1.5 Transcription factor binding complexes and their role in gene regulation during erythroid differentiation

The hematopoietic system has served as a classical model for studying transcriptional control during multi-lineage differentiation [188-190]. Pluripotent stem cells in this system generate progenitor cells that can further differentiate into more than ten distinct types of mature blood cells [191]. Differentiation from hematopoietic stem cells (HSCs) to a specific lineage requires the trigger of a relatively small number of

critical transcription factors that are sequentially expressed and are largely restricted to that specific lineage. These factors include well-studied TFs such as GATA1, GATA2, FOG1, SCL/TAL1, LDB1, ETO2, EKLF, PU1, LYL1, LMO2, GFI1B and RUNX1 (reviewed in [192-194]). Gene-targeting knockout in mouse models for these crucial factors leads to the development of specific hematological malignancies [190,195]. Distinct multi-protein complexes of these factors have been demonstrated to have critical roles in specification and biological function of HSCs, and also in the development of specific mature blood lineages.

One of the branches of hematopoiesis that has been well studied with respect to TF involvement is the formation of red blood cells, called erythropoiesis. Formation of distinct multi-protein complexes including GATA1, SCL/TAL1, LMO2, LDB1, and KLF1 is important for the control of transcription during erythropoiesis [195]. The repertoire of multi-protein complexes and the cellular mechanisms involved in the activation and repression of specific genes is not completely known. Therefore, significant effort has been expended, using new techniques such as ChIP-seq, to identify the binding sites of individual regulators, as well as to characterize the relationships between them. To understand the roles of regulators in transcriptional activation or repression, ChIP-seq can be integrated with gene expression profiles to provide new information and to build regulatory networks that describe the regulatory mechanisms underlying erythropoiesis. Recently, several studies have reported a number of TFBS identified by ChIP-seq, for both individual factor and multiple factors of key erythroid factors. For instance, GATA1 binding sites have been reported in several mouse and human erythroid cell lines [133,196-199]. Genome-wide occupancy of GATA1 from all studies revealed that ~10-15% of GATA1 binding sites are located at the proximal promoter regions, whereas the remainder occurs at distal regulatory elements distributed in both intra- and inter-genic regions [133,196-199]; these identified elements are also enriched for active histone marks such as H3K4me1, H3K9ac, and H3K27ac [199]. Integrating patterns of GATA1 occupancy with gene expression profiles demonstrated that GATA1 activates a large number of genes in concert with other factors and co-factors known as the LDB1 complex. This complex consists of several crucial factors such as GATA1, LMO2, E2A, and SCL/TAL1 as well as co-factors like ETO2, MTGR1, and CDK9 [133]. This activation complex assembles at the E-box-GATA1 DNA motifs spaced 9-11

nucleotides apart [195]; derived motifs from ChIP-seq data revealed a preference for TG dinucleotide upstream of WGATAR, with a preferred consensus (C)TGN$_{7-8}$WGATAR [133]. This protein complex activates target genes via the formation of a chromatin loop, as demonstrated by 3C-seq analyses at both β-globin and *Myb* locus [133,174]. GATA1 has a role not in only gene activation but also in the repression of a large number of target genes during erythropoiesis [198]. However, the components of the GATA1 repressive complex are not completely known or understood. A recent study suggested that GATA1 potentially recruits components of the repressor GFI1B complex via LSD1. This recruitment may remove H3K4 methylation at genes expressed during the earlier stages of erythroid development such as at the *GATA2* gene. The PRC2 complex may then act to mediate and stabilize gene silencing at a subset of genes [198].

Similar to GATA1, SCL/TAL1 is present at all GATA binding elements that activate target genes, and depleted at sites where GATA1 acts as a repressor [200]. ChIP-seq analysis of SCL/TAL1 revealed approximately 3,000-5,000 binding sites, mostly found at distal regulatory regions similar to those observed for GATA1 binding [133,200,201]. SCL/TAL1 physically interacts with E2A with the help of LMO2 to form the LDB1 complex, which is involved in the activation of a large number of target genes. To repress target genes, SCL/TAL1 may recruit the co-repressors ETO2 and MTGR1 to mediate gene silencing, since some SCL/TAL1 target genes are de-repressed when the binding level of ETO2 and MTGR1 is depleted during differentiation [133,200].

As described previously, GATA1 and SCL/TAL1 are components of the LDB1 complex. The LDB1 complex is a key regulatory complex during erythroid maturation [202]. ChIP-seq analysis of the LDB1 complex in MEL cells during the induction of differentiation revealed that the LDB1 complex changes during differentiation [133]. Importantly, the co-repressors ETO2 and MTGR1 showed significant decreases in both binding intensity and the number of binding sites towards the terminal stage of differentiation, suggesting that the full LDB1 complex binds to genes that are poised to be expressed in the earlier stage. After induction of

differentiation, the LDB1 complex dynamically changes to activate target genes while the levels of the co-repressors ETO2 and MTGR1 are decreased [133].

KLF1, a zinc finger transcription factor that binds to DNA at the CACC box motifs, is one of the key erythroid factors, and is remarkably erythroid-restricted [203]. ChIP-seq analysis of KLF1 revealed between 945 and 1,380 binding sites in primary erythroid cells; ~10% of these sites located within 1 kb of the TSS, whereas the remainder were distributed > 10 kb away from TSSs [172]. About a half of these identified KLF1 binding sites are within 1 kb of the GATA1 binding sites that overlap with a small number of SCL/TAL1 binding sites. This suggests a KLF1/GATA1 complex, which is distinct from the complexes that contain SCL/TAL1. Surprisingly, many binding sites of KLF1 are not co-occupied by P300, suggesting KLF1 functions at these sites in a P300-independent manner [204]. As demonstrated by the integration of ChIP-seq and expression profiles of wild-type and the $Klf1^{-/-}$ cells, these two distinct KLF1/GATA1 complexes activate a large number of target genes involving in terminal erythroid differentiation and maintaining of homeostasis within the erythroid compartment [172]. Recent mRNA-seq analysis identified additional KLF1 target genes and demonstrated that KLF1 targets are not only the genes responsible for the production of hemoglobin, but also include regulators of the cell cycle, membrane and cytoskeletal components, and apoptosis [204].

2. Present investigation

This thesis was a part of, and supported by, the European Transcriptome, Regulome and Cellular Commitment Consortium (EuTRACC), therefore the aims for this thesis have been proposed to support the scientific objectives of EuTRACC, specifically focusing on following topics:

- To study protein complexes of basic, general and tissue-specific transcription factors, and their interacting partners in hematopoietic cell types; this thesis has focused on the identification of TFBSs of several key factors and the characterization of the binding patterns of protein complexes involved in gene regulation during hematopoietic development, primarily during erythropoiesis.

- To develop computational procedures, bioinformatics methods, and tools for the analysis and visualization of next-generation sequencing data; specifically focusing on the development of high-throughput sequencing data analysis pipelines, including data visualization software (implemented in R statistical programming environment) to allow the public dissemination of findings.

- To perform an integrative computational analysis of large-scale experimental data, specifically using data generated from ChIP-seq, microarray and/or RNA-seq, and 3C-seq technologies, with the aim of generating new hypotheses and acquiring new biological knowledge on mammalian gene regulation.

2.1 The genome-wide dynamics of the binding of LDB1 complexes during erythroid differentiation (Paper I)

LDB1 is known to form a protein complex with GATA1 and TAL1. This complex is critical for the differentiation of the erythroid cell lineage [200,205,206]. To study the importance of the LDB1 complex during erythropoiesis, we took the advantage of the MEL cells, which correspond to the proerythroblast stage of erythroid differentiation. MEL cells were induced to differentiate with 2% dimethylsulfoxide (DMSO) for 4 days to erythrocyte-like cells. To identify binding sites of the LDB1 complex, we

performed ChIP-seq analysis for individual factors that are known to be involved, consisting of LDB1, GATA1, SCL/TAL1, ETO2 and MTGR1, both before and after differentiation. Using a comprehensive bioinformatic analysis to identify binding sites, we were able to detect 4,271 and 4,982 LDB1 binding sites, 5,368 and 5,205 GATA1 binding sites, 671 and 4,173 SCL/TAL1 binding sites, and 2,159 and 480 ETO2/MTGR1 binding sites before and after differentiation, respectively. The binding site analysis showed that a large number of LDB1 occupied regions are bound by GATA1 and TAL1 in both stages. Motif analysis of the top scoring LDB1 binding sites confirmed a preference for binding of the LDB1 complex to the E-box-GATA1 motif with a preferred consensus DNA motif of (C)TGN$_{7–8}$WGATAR, which is known to be the binding motif of SCL/TAL1 and GATA1. We noted that the observed E-box motif in our study is different from the previously published E-box sequence (CANNTG) [207]. Our observations have established that the LDB1 complex is the most prevalent complex involved in the regulation of terminal erythropoiesis.

We observed that the LDB1 complex changes the number of binding sites and the binding intensity (peak height) of its protein complexes during MEL cell differentiation. There was a significant net decrease in the relative binding intensity of co-repressors ETO2 and MTGR1, whereas the relative binding intensity of LDB1 and TAL1 was increased towards an activating state. This change was observed at *Epb4.2*, *Alas2*, and *Slc22a4* genes, which are target genes of the LDB1 complex and are induced late during erythroid differentiation. To allow genome-wide identification of target genes of the LDB1 complex, gene expression profiles from MEL cells both before and after differentiation were analyzed and integrated with regions that were identified as being occupied by the LDB1 complex. Integrative analysis showed that the LDB1 complex acts almost exclusively as an activator, since most of upregulated genes are direct targets of the complex (with binding sites located within 50 kb of the TSS), whereas the most strongly downregulated genes showed no evidence of binding of the LDB1 complex. Many of the identified direct target genes were involved in different pathways such as heme biosynthesis, cell cycle, apoptosis, and gas transport. This result suggested that the complete LDB1 complex, which contains ETO2 and MTGR1, binds to target genes that are poised to be expressed during the earlier stage (before differentiation). After the induction of differentiation, the LDB1 complex

dynamically changes to activate target genes while the levels of ETO2/MTGR1 are decreased and LDB1/TAL1 are increased toward the terminal stage of differentiation.

The binding distribution of the LDB1 complex around target genes also showed that the complex frequently binds downstream from the TSS, and is typically found in the first intron of a gene. We observed a clear difference in the binding pattern of the LDB1 complex between target genes with high-CpG and low-CpG promoters. In high-CpG promoters, the binding sites were commonly found between 1 and 3 kb downstream from the TSS, whereas in low-CpG promoters, most binding sites were found in the promoter. This result suggested that the LDB1 complex binds differently to different types of core promoters, with low-CpG promoters being associated with specific expression in terminally differentiated tissues [9] and high-CpG promoters being associated with either housekeeping genes or developmentally regulated genes [208].

Although the binding sites of the LDB1 complex are distributed around promoter regions, we observed that clusters of binding sites are often located at a long distance from genes, which are (possibly) upregulated and are developmental genes. These genes may be regulated by long-range interactions. To determine whether the LDB1 complex is involved in mediating long-range interactions involving target genes, we performed 3C-seq using the β-globin gene (β-major promoter) as the viewpoint. 3C-seq analyses revealed that β-major promoter interacts with binding sites of the LDB1 complex that are located within the LCR. This result suggested that binding sites of the LDB1 complex mark positions of sites involved in long-range interactions.

2.2 The updated JASPAR database with new matrix profiles derived from high-throughput sequencing data (Paper II)

While we were analyzing LDB1 complex-binding sites from ChIP-seq data, we derived PWMs for individual factors using a thousand of actual binding DNA sequences. Our derived profiles from ChIP-seq data, using a large number of representative target sequences, obviously improved existing matrices found in JASPAR database [89] because ChIP-seq provides higher information content than

the original ones, which were derived using the SELEX assay. We therefore updated existing matrix profiles in JASPAR database with the first batch of matrix profiles derived from ChIP-seq data. We have expanded the number of matrix profiles and updated existing matrices for the JASPAR core database and also derived and curated PWMs from other genome-wide methods such as ChIP-chip, a comprehensive protein binding matrix (PBM) experiment [91], and new literature-based profiles from PAZAR [209]. Currently, JASPAR contains 457 non-redundant matrix profiles.

2.3 Dynamic long-range chromatin interactions control *Myb* proto-oncogene transcription during erythroid development (Paper III)

During the analysis of the binding of components of the LDB1 complex [Paper I], we observed that, although the LDB1 complex was found to bind at promoter regions, its bindings sites were most frequent in gene introns and often located more than 100 kb away from annotated genes and showed a preference towards intergenic regions. Using an integrative analysis of ChIP-seq and 3C-seq, we demonstrated that the cluster of multiple distal LDB1 binding sites interacts with β-globin locus, and these interactions appear to increase upon differentiation, suggesting that the LDB1 complex marks the formation of chromatin loops, which are important for β-globin gene activation [Paper I].

Several clusters of LDB1 complex binding sites were observed in intergenic regions. One interesting example was identified within the *Myb-Hbs1l* intergenic region, a 135 kb region known to harbor a set of common intergenic single nucleotide polymorphisms (SNPs) and other variants that were highly associated with clinically important human erythroid traits [210-212]. Since *Myb*, encoding the c-Myb transcription factor, is a key hematopoietic regulator and plays a pivotal role in maintaining a proper balance between erythroid cell proliferation and differentiation [169-171], we performed an integrative analysis of ChIP-seq and 3C-seq to identify the presence of erythroid-specific long-range interactions between intergenic LDB1 complex-binding regions and *Myb* gene during MEL cell differentiation. ChIP-seq analysis showed five LDB1 binding sites in MEL cells and in primary mouse erythroid progenitor cells from fetal liver, located -36, -61, -68, -81, and -109 kb

upstream of the *Myb* transcription start site. Using a luciferase reporter gene assay, we found that these regions appeared to function as enhancers, and were associated with strong binding of P300, RNAPII, H3K4me1, and H3K27ac. Interestingly, the -81 kb binding site was co-occupied by KLF1, one of the master erythroid-restricted regulators [171] .

3C-seq experiments were performed using the *Myb* promoter as a viewpoint on fetal liver erythrocytes (expressing high levels of *Myb*) and fetal brain cells (expressing undetectable levels of *Myb*) to investigate whether these binding sites interact with the *Myb* gene via chromatin looping. The fetal brain cells were used as a negative control. Using our *r*3Cseq analysis pipeline [Paper IV], we reported that 3C-seq signals were detected to coincide with the binding of LDB1 complex, KLF1, and CTCF. Interaction frequencies of these regions in fetal liver erythrocytes were statistically significant and substantially higher than in fetal brain, where they were either absent or low. Interestingly, the repeat experiment using the -36, and -81 kb binding sites as the viewpoints clearly demonstrated that these sites interact with the *Myb* promoter and the adjacent CTCF-bound intron 1 element. Thus, 3C-seq data confirmed the presence of erythroid-specific long-range interactions between intergenic binding sites, *Myb* promoter, and the first intron bound by CTCF, suggesting the formation of an ACH coincident with upregulation of *Myb* expression.

We further investigated the association between the CTCF-bound intron 1 element and markers of productive transcription elongation, Ser2-phosphorylated RNAPII and H3K36me3. We observed that the transition to productive transcription elongation occurs around the CTCF-bound intron 1 element.

Since *Myb* expression is decreased during differentiation of MEL cells, while nearby genes (*Hbs1l* or *Ahi1*) are unaffected, we decided to investigate the dynamics of long-range chromatin interactions that lead to the loss of *Myb* expression during cellular differentiation. To this end, we analyzed 3C-seq data obtained from MEL cells before and after differentiation, using the *Myb* promoter as a viewpoint. 3C-seq data analysis during differentiation demonstrated that in non-induced MEL cells (expressing high levels of *Myb*), interacting regions are similar to those found in fetal liver erythrocytes and often overlapping with intergenic enhancers. Strikingly, upon induction of

differentiation these interacting regions showed much lower interaction signals. These results revealed the loss of interactions between the promoter and intergenic regulatory regions upon cellular differentiation, suggesting that *Myb* downregulation upon erythroid differentiation is accompanied by loss of communication between *Myb* promoter, *Myb* intron 1 element (bound by CTCF), and intergenic enhancers. We therefore proposed a *Myb* ACH model, which consists of the *Myb* promoter, intron 1 element, and distal enhancers. In erythroid progenitors, this ACH primarily maintains high-level of the *Myb* expression with a local high concentration of RNAPII, transcription and elongation factors. During terminal differentiation, the ACH is destabilized due to a loss of contacts between the *Myb* promoter and intergenic regulatory regions, resulting in decreased *Myb* transcription.

Our analysis provides a framework for further study in human erythroid cells, where the *MYB-HBS1L* intergenic region has long been suspected to possess a regulatory function. The region harbors several SNPs associated with clinically relevant human erythrocyte traits, e.g. the persistence of fetal hemoglobin in adults and pain crises in sickle cell disease [210-212]. The analysis of association between TF-bound elements that contain SNPs and the long-range chromatin contacts in *Myb* gene regulation may provide crucial information to investigate the functional impact of the erythroid phenotype-associated variants in human diseases.

2.4 *r3Cseq*: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data (Paper IV)

We performed 3C-seq studies to demonstrate the phenomena of long-range gene regulation via chromatin looping [Paper I and Paper III], while 3C-seq data analysis and tools were not available at that time. For this reason, we have developed a R/Bioconductor package called *r3Cseq* to perform the analysis of data generated by 3C-seq technology. The package is built on and extends the functionality of existing Bioconductor packages. It is composed of several basic functions consisting of

aligned read manipulation, data processing, data normalization, identification of interacting regions and visualization. The package supports different experimental designs, e.g. either with or without a control experiment and either with or without replicates.

We adapted existing methods used in high-throughput sequencing data analysis and applied them for data normalization and statistical testing to allow the identification of *cis*- and *trans*-interactions. To normalize 3C-seq data, we fitted the reverse-cumulative distribution of reads per region to a power-law distribution. This normalization technique increases the statistical power of interaction signal detection. To detect significant interactions in both *cis* and *trans,* we adapted methods used in previous 4C [161] and 5C [213] studies. Our detection method corrects for any bias resulting from background signal and assigns an interaction score *(q*-value*)* to a specific restriction fragment or a defined window.

Our *r3Cseq* package supports the identification of interacting regions in both restriction fragment-based and window-based methods, which is a great help to allow scientists to compare different ways of analyzing their dataset and to select the most suitable analysis for the interpretation of their data. To show the usefulness of the package, we have successfully demonstrated the use of *r3Cseq* to characterize long-range interactions at the mouse β-globin locus [Paper I] and to study chromatin interactions in a structurally unexplored *Myb* locus [Paper III] in erythroid cells.

Finally, *r3Cseq* is an R user-friendly tool that produces a range of plots specifically designed for the visualization of genomic regions at both the genome-wide and user-defined level with additional genomic features, such as gene models. Its generated output consists of plain text and *bedGraph* files compatible with other visualization tools, such as the UCSC Genome Browser [214].

2.5 Genome-wide dynamics of P300 transcription factor complexes during erythroid differentiation (Paper V)

In paper I, we reported the genome-wide dynamics of the LDB1 complex (LDB1, GATA1, SCL/TAL1, MTGR1, and ETO2) during MEL cell differentiation. To gain a detailed insight in the dynamics of regulatory complexes in this system, we extended this work by conducting ChIP-seq experiments for additional factors consisting of RUNX1, GFI1B, FOG1, LSD1, LMO2, LMO4, P300, TIF1γ, CTCF and RNAPII, both before and after MEL cell differentiation. In this study, we focused on the analysis of the P300 complex, because it is known to mark regulatory sequences both in *vitro* and in *vivo* and is required for hematopoiesis [215-221].

Multiple peak calling software were used to detect binding sites from ChIP-seq data, and we selected only consensus-binding sites generated using the combined detected regions from different software as representative binding sites for each factor. We reported the number of identified binding sites for all factors in both stages. These numbers varied between ~2,000 (FOG1 after differentiation) and ~110,000 (RNAPII before differentiation) binding sites. We observed that for most of the factors the number of identified binding sites is different before and after differentiation. For example, the number of binding sites for TAL1, P300, and LMO2 were significantly increased, whereas the number of binding sites for RNAPII, RUNX1, ETO2, GFI1B, FOG1, LSD1, and TIF1γ were significantly decreased during differentiation. Interestingly, the key regulators, such as GATA1, LDB1, LMO4 and the insulator CTCF, showed negligible differences in the number of identified binding sites during differentiation.

We have investigated how the binding intensity of each factor changes at individual binding sites during differentiation. One of the main problems for this investigation was that ChIP-seq library sizes from before and after differentiation are highly different in many data sets. These differences may bias the analysis of the binding intensity comparison between two stages. To this end, we applied quantile normalization to normalize peak heights of binding sites before and after differentiation. ChIP-qPCR validation of a number of selected binding sites was found to validate the normalized binding signals from ChIP-seq data. We therefore applied this normalization technique genome-wide. We used P300 binding sites as reference points to identify its binding complexes. We assumed that P300 forms protein complexes with other factors when binding sites from other factors are located

within ±250bp of P300 binding sites. Using this technique, we could retrieve normalized binding signals (Z-scores) for all factors in individual P300 binding sites to generate a P300 complex-binding matrix. This matrix was then subjected to unsupervised *K*-means clustering. Importantly, this analysis identified twelve distinct patterns of P300 binding complexes. These patterns exhibited different compositions of the P300 complex during MEL cell differentiation. Importantly, we did not find a clear pattern where P300 was alone in the complex (although P300 may form a complex with other factors which were not examined in this study), suggesting that all P300-containing complexes during erythropoiesis also contain key erythroid factors, and that no other types of P300-containing complexes exist in significant number.

We have classified twelve binding patterns of the P300 complex into three different classes consisting of (1) binding patterns with no/low/moderate RNAPII binding signal (clusters A, C, D, E, F, H, I, K and L), (2) binding patterns with high signal of RNAPII (clusters B and J), and (3) the binding pattern with high signal of CTCF (cluster G). The distribution of these P300 binding patterns around the TSS showed that these classes are associated with enhancers, promoters, and insulators respectively. Each class showed different levels of DNaseI hypersensitive sites (DHSs), H3K27ac, H3K4me1 and H3K4me3. Indeed, cluster A, B, C, D, and J showed high signals of DHSs and H3K27ac. Clusters A, C and D, which were classified into the enhancer class, were marked by DHSs with high level of H3K4me1 and low level of H3K4me3. Clusters B and J, classified in the promoter class, were associated with DHSs marked by high level of H3K4me3 and low level of H3K4me1. These results revealed the characteristics of P300 binding patterns and confirmed the existence of biological relevance between computationally identified binding patterns and chromatin signatures.

We next demonstrated that the P300 complex dynamically changes during differentiation. We observed that five out of twelve binding patterns (cluster A, C, E, F and L) contain all of the factors, while other clusters lose one or more factors in the complexes. Cluster A, C, D, and I showed strong binding intensity of P300 and key factors GATA1, LDB1, TAL1, GFI1B, and LSD1 were significantly increased, whereas FOG1, LMO2, LMO4, and TIF1γ were significantly decreased within the

complex during differentiation. Individual gene loci that are known to be induced late during erythroid differentiation were associated with multiple bindings of cluster A and C, such as *Alas2* and *Gata1*. These clusters likely include the previously identified activating LDB1 complex [Paper I]. Interestingly, we showed that P300 binding-intensity of cluster F, K and L significantly decreased during differentiation, resulted in the decreasing of binding-intensity in other factors in these complexes.

Integrative analysis using P300 binding patterns and RNA-seq data generated both before and after differentiation showed a clear bias for cluster A, C, and E to contain significantly upregulated genes. Although less visible, clusters D, H and I also occurred preferentially around upregulated genes. This result suggested that these complexes activate their target genes, which may due to the increased binding intensity of P300. We thus defined these P300 binding patterns as the P300 activation complex. In contrast, we observed that clusters F and K, corresponding with the decrease of P300 binding intensity during differentiation, are strongly associated with downregulated genes. We thus defined these binding patterns as the P300-containing erythroid repression complex. Motif analysis revealed preferentially binding DNA motifs of both P300 activation and repression complexes. Consistent with the previous study [Paper I] the activation complex contained a high fraction of TAL1::GATA1, suggesting that TAL1::GATA1 motif is important for erythroid gene activation. In contrast, the repressive complex contained a high fraction of the GATA1 motif only. Our findings were consistent with recent studies, which reported that the activation complex in erythroid cells tends to be assembled at the TAL1::GATA1 motif, whereas the repressive complex has either lost or has lowered levels of TAL1 [200,222-224]. We also observed that conservation (phastCons 30-way) scores of TAL1::GATA1 motif found in the P300 activation complex are significantly higher than GATA1 motif found in the repression complex. Our results showed that the constraint on TAL::GATA1 motif occurs more frequently for transcriptional enhancement than for repression.

We further demonstrated that the majority of differentially expressed genes are associated with multiple P300 binding patterns within 50kb upstream and 10kb downstream of genes. Multiple clusters occurred to be biased towards up- and down-regulated genes when compared to genomic background. We next asked which pairs

of clusters were associated to the same target genes more often than expected by chance. Using cluster pair association analysis, we observed that the frequency of co-occurrence of clusters A, C and/or E is significant in the proximity of upregulated genes, whereas clusters F and K are significantly co-occur in the proximity of downregulated genes. We also demonstrated how the co-occurrence of multiple P300 binding patterns correlate with different gene expression levels. The cluster pair association analysis at different gene expression levels showed that clusters E and I are frequently co-occurred at lowly expressed genes, whereas clusters A, C, E, and B increase their co-occurrences from mid to highly expressed genes. Gene ontology analysis of multiple P300 complex-binding targets showed that target genes with multiple clusters were widely associated with some of the most important erythroid-specific processes, including heme biosynthesis, hemoglobin metabolic process and erythroid differentiation.

We finally associated our patterns of P300 complex-binding sites with eRNA derived from erythroid cells generated by Kowalczyk et al [225]. We identified 943 extragenic and 1,266 intragenic P300 binding sites, which are associated with eRNAs. Interestingly, cluster A, C, and D, classified as the P300 activation complex, showed high fraction of intragenic/extragenic binding sites that produced eRNAs, whereas other clusters seems to have very low signal of eRNAs. Importantly, clusters A and C contain a high level of eRNA signal at binding sites that were associated with significantly upregulated genes. In contrast, there was very low/no signal of eRNAs at binding sites that were associated with significantly downregulated genes, and highly or lowly expressed genes which were not differentially expressed. These results suggest that our identified P300 activation complexes show evidence of transcription at active enhancers, which may imply that P300 activation complexes bind RNAPII and produce eRNAs to facilitate active transcription.

3. Discussion

Technological advances, especially in high-throughput sequencing technologies, have significantly improved the identification and characterization of regulatory regions. This thesis describes my contributions to the analysis of massively parallel sequencing data for studying regulatory elements, their associated TF complexes and chromatin interactions that are essential for gene regulation during erythropoiesis.

The identification and characterization of regulatory elements has been significantly improved in the past several years from using computational techniques to using ChIP-based genome-wide methods, which currently are the most efficient tool to directly identify TFBSs. Here, we used ChIP-seq technology to study regulatory elements generated from multiple TFs involving in erythropoiesis. ChIP-seq was conducted to profile binding sites bound by multiple proteins that are involved in the LDB1 complex (LDB1, GATA1, SCL/TAL1, ETO2, and MTGR1), both before and after MEL cell differentiation [Paper I]. To gain a detailed insight of TF involvement in this system, additional ChIP-seq experiments and data analyses involving other key factors (RUNX1, GFI1B, FOG1, LSD1, LMO2, LMO4, P300, TIF1γ, CTCF, and RNA Polymerase II (RNAPII) were performed [Paper V]. Short reads were generated from each factor yielding between ~6 million and ~53 million mapping reads [Paper I and V]. In total, there were about a half billion informative reads from ChIP experiments used in the downstream analysis. This massive amount of informative reads required the development and use of a set of bioinformatics tools for in-depth analysis. The results from the analysis of ChIP-seq data provided an invaluable resource for studying regulatory elements during erythropoiesis.

Comprehensive ChIP-seq analysis pipelines have been implemented using the existing R/Bioconductor packages [99] such as ShortRead [226], Rsamtools, GenomicRanges [227], rtracklayer [228], and BSgenome packages, to facilitate data processing, manipulation, mining and visualization. These analysis tools have been used to identify the binding sites of multiple factors and to both discover the binding patterns of protein complexes and the correlation of these patterns with the expression level of associated genes [Paper I and V]. A customized set of analysis tool has been

used to demonstrate the association between identified protein-binding sites and their relationship with their target genes, as well as their target promoter types. When promoters were classified according to their CpG content; we discovered a clear difference in LDB1-binding distribution between high-CpG and low-CpG target promoters, demonstrating that different promoter types respond differently to proximal promoter versus mid/long-range regulatory inputs [Paper I]. The analysis tools also support the *de novo* motif discovery using parallelized MEME [229] on a supercomputing platform, which can handle inputs of many thousands of sequences generated from identified binding sites. Using these tools, motif analyses such as motif logo generation, motif scanning, and motif identification can be efficiently performed, thus greatly facilitating the interpretation of experimental results. One of the great advantages of using these tools is to derive new matrix profiles from ChIP-seq data for improving existing matrices found in the JASPAR database [Paper II].

The analysis of ChIP-seq data showed that several clusters of LDB1 complex-binding sites are often found at long distances from erythroid genes. For instance, five binding sites were located at the LCR of β-globin gene [Paper I]. At the *Hbs1l-Myb* intergenic region, the cluster of LDB1 complex-binding sites was also observed upstream of the *Myb* gene. Other large intergenic regions of genes, such as *Klf3*, *Ets2*, *Max*, *Mef2c*, or *Pim1,* were occupied by LDB1. These genes are known to be involved in hematopoiesis and are regulated by long-range enhancers [44,45,230]. We therefore suggested that LDB1 complex-binding sites are involved in long-range gene regulation. We tested this possibility at the β-globin and *Myb* loci using 3C-seq [Paper I, III, and IV]. To analyze 3C-seq data, we developed a R/Bioconductor package called *r3Cseq* to provide basic functions and to facilitate 3C-seq data analyses [Paper IV]. Using *r3Cseq,* we demonstrated the importance of long-range chromatin contacts involving in β-globin and *Myb* gene regulation. Our integrative method that combines the analysis of multiple ChIP-seq and 3C-seq data would be a good example to show the advancement in studying long-range gene regulation. Our method would pave the way to go beyond the integration of individually analyzed TF binding profiles to instead study how TFs or *cis*-regulatory elements act cooperatively to affect the transcriptional level of their target genes [Paper III].

Most studies of TF binding sites in hematopoietic system have analyzed TFs in a

particular cell type or lineage [132,196,231-233] to demonstrate the importance of TFs in that cell populations. Therefore, these studies cannot be used to explain the dynamics of protein complexes during cell development and differentiation process. Here, in our ChIP-seq studies, we analyzed ChIP-seq data over a time course to understand how diverse TF complexes are established during the differentiation of lineage-specific cell types. We studied the dynamics of regulatory complexes, in particular the LDB1 and P300 complexes during differentiation of proerythroblast-like cells to fully differentiated erythrocyte-like cells. Our integrative analysis produced significant findings relating to how protein complexes change during differentiation and how these changes relates to gene regulation. We showed that dynamic changes in the composition of the LDB1 complex have major impacts on the expression of its targets, in particular, the binding of the co-repressor ETO2 and MTGR1 significantly decreases during differentiation. The LDB1 complex that contains ETO2 and MTGR1 directly binds to genes that are involved in important pathways (e.g. heme biosynthesis, cell cycle, apoptosis, and gas transport) and these genes are not expressed during the earlier stage. This suggests that this version of the LDB1 complex can act as the repressor. During cell differentiation, the composition of the LDB1 complex changes to activate these genes, while the binding level of its composition ETO2 and MTGR1 is decreased toward the terminal stage of differentiation. This suggests that the LDB1 complex acts as an activator by binding a very specific subset of genes that are induced late during erythroid differentiation.

Another example is the genome-wide dynamics of P300 complexes. We extended the study of the LDB1 complex to involve more TFs. In this study, computational techniques have been intensively used to discover the pattern of co-occupancy by multiple TFs at each binding site. We computationally discovered twelve distinct patterns of P300 binding complexes, according to dynamic changes of binding intensity during differentiation. These patterns exhibited specific characteristics in each group of binding patterns, such as the motif composition, the distribution in relation to genes, and the type and level of histone modifications. Importantly, integrative analysis of these identified binding patterns with gene expression profiles showed specific P300 binding patterns that are associated with up- and downregulation of gene during erythropoiesis. This study presents how a comprehensive bioinformatics analysis can be used to mine multifactor ChIP-seq data

and to reveal new insights into the combination of TF complexes that are essential for transcriptional control.

We studies TF complexes by using the protein of interest as the reference. For instance, P300 was used as a reference for the identification of P300 complex-binding patterns. Using this strategy, we have successfully described how P300 complexes change their composition during differentiation. The advantage of this technique is to reduce the complexity of the combinatorial analysis, since finding all of the combinations in all binding sites from all factors may generate a large number of combinations, which can be difficult to discover significant and biologically relevant binding patterns. However, using this technique may limit the identification of important binding patterns if there is low co-occupancy between the references factor and other factors in binding regions. Thus, the integrative results from individual protein complex analysis may be required in order to provide all relevant protein-binding information using this method. The techniques presented in this thesis could provide the ideas for the analysis of other key erythroid factors such as GATA1, TAL1, and RUNX1 complexes.

In summary, the thesis work demonstrated the analysis of large-scale experimental data in particular for ChIP-seq, and 3C-seq technologies. Computational methods and tools were developed to identify and to characterize TF complexes and their binding patterns as well as to study the dynamics of protein complexes and their associated chromatin interactions that are involved in gene regulation during erythropoiesis. Our comprehensive data analysis presented in this thesis work may facilitate the analysis of high-throughput sequencing data that can be used for a broader research community in epigenomics and mammalian gene regulation.

4. References

1. Claverie, J.M. (2001) **Gene number. What if there are only 30,000 human genes?** *Science*, **291**, 1255-1257.
2. Levine, M. and Tjian, R. (2003) **Transcription regulation and animal diversity**. *Nature*, **424**, 147-151.
3. Reik, W. (2007) **Stability and flexibility of epigenetic gene regulation in mammalian development**. *Nature*, **447**, 425-432.
4. Thomas, M.C. and Chiang, C.M. (2006) **The general transcription machinery and general cofactors**. *Crit Rev Biochem Mol Biol*, **41**, 105-178.
5. Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H. and Kim, V.N. (2004) **MicroRNA genes are transcribed by RNA polymerase II**. *Embo J*, **23**, 4051-4060.
6. Juven-Gershon, T., Hsu, J.Y., Theisen, J.W. and Kadonaga, J.T. (2008) **The RNA polymerase II core promoter - the gateway to transcription**. *Curr Opin Cell Biol*, **20**, 253-259.
7. Butler, J.E. and Kadonaga, J.T. (2002) **The RNA polymerase II core promoter: a key component in the regulation of gene expression**. *Genes Dev*, **16**, 2583-2592.
8. Smale, S.T. and Kadonaga, J.T. (2003) **The RNA polymerase II core promoter**. *Annu Rev Biochem*, **72**, 449-479.
9. Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C. et al. (2006) **Genome-wide analysis of mammalian promoter architecture and evolution**. *Nat Genet*, **38**, 626-635.
10. Lenhard, B., Sandelin, A. and Carninci, P. (2012) **REGULATORY ELEMENTS Metazoan promoters: emerging characteristics and insights into transcriptional regulation**. *Nature Reviews Genetics*, **13**, 233-245.
11. Baumann, M., Pontiller, J. and Ernst, W. (2010) **Structure and basal transcription complex of RNA polymerase II core promoters in the mammalian genome: an overview**. *Mol Biotechnol*, **45**, 241-247.
12. Juven-Gershon, T. and Kadonaga, J.T. (2010) **Regulation of gene expression via the core promoter and the basal transcriptional machinery**. *Dev Biol*, **339**, 225-229.
13. Ptashne, M. and Gann, A. (1997) **Transcriptional activation by recruitment**. *Nature*, **386**, 569-577.
14. Orphanides, G., Lagrange, T. and Reinberg, D. (1996) **The general transcription factors of RNA polymerase II**. *Genes Dev*, **10**, 2657-2683.
15. Naar, A.M., Lemon, B.D. and Tjian, R. (2001) **Transcriptional coactivator complexes**. *Annu Rev Biochem*, **70**, 475-501.
16. Narlikar, G.J., Fan, H.Y. and Kingston, R.E. (2002) **Cooperation between complexes that regulate chromatin structure and transcription**. *Cell*, **108**, 475-487.
17. Gaston, K. and Jayaraman, P.S. (2003) **Transcriptional repression in eukaryotes: repressors and repression mechanisms**. *Cell Mol Life Sci*, **60**, 721-741.
18. Thiel, G., Lietz, M. and Hohl, M. (2004) **How mammalian transcriptional repressors work**. *European Journal of Biochemistry*, **271**, 2855-2862.

19. Perissi, V., Jepsen, K., Glass, C.K. and Rosenfeld, M.G. (2010) **Deconstructing repression: evolving models of co-repressor action**. *Nature Reviews Genetics*, **11**, 109-123.
20. Zhao, J., Sun, B.K., Erwin, J.A., Song, J.J. and Lee, J.T. (2008) **Polycomb Proteins Targeted by a Short Repeat RNA to the Mouse X Chromosome**. *Science*, **322**, 750-756.
21. Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E. et al. (2007) **Functional demarcation of active and silent chromatin domains in human HOX loci by Noncoding RNAs**. *Cell*, **129**, 1311-1323.
22. Kaplan, C.D., Laprade, L. and Winston, F. (2003) **Transcription elongation factors repress transcription initiation from cryptic sites**. *Science*, **301**, 1096-1099.
23. Wasserman, W.W. and Sandelin, A. (2004) **Applied bioinformatics for the identification of regulatory elements**. *Nat Rev Genet*, **5**, 276-287.
24. Smith, A.D., Sumazin, P., Xuan, Z. and Zhang, M.Q. (2006) **DNA motifs in human and mouse proximal promoters predict tissue-specific expression**. *Proc Natl Acad Sci U S A*, **103**, 6275-6280.
25. Roider, H.G., Lenhard, B., Kanhere, A., Haas, S.A. and Vingron, M. (2009) **CpG-depleted promoters harbor tissue-specific transcription factor binding signals--implications for motif overrepresentation analyses**. *Nucleic Acids Res*, **37**, 6305-6315.
26. Gardinergarden, M. and Frommer, M. (1987) **Cpg Islands in Vertebrate Genomes**. *J Mol Biol*, **196**, 261-282.
27. Antequera, F. and Bird, A. (1993) **Number of Cpg Islands and Genes in Human and Mouse**. *P Natl Acad Sci USA*, **90**, 11995-11999.
28. Saxonov, S., Berg, P. and Brutlag, D.L. (2006) **A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters**. *P Natl Acad Sci USA*, **103**, 1412-1417.
29. Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) **Cpg Islands as Gene Markers in the Human Genome**. *Genomics*, **13**, 1095-1107.
30. Yamashita, R., Suzuki, Y., Sugano, S. and Nakai, K. (2005) **Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity**. *Gene*, **350**, 129-136.
31. Deaton, A.M. and Bird, A. (2011) **CpG islands and the regulation of transcription**. *Genes Dev*, **25**, 1010-1022.
32. Rach, E.A., Winter, D.R., Benjamin, A.M., Corcoran, D.L., Ni, T., Zhu, J. and Ohler, U. (2011) **Transcription Initiation Patterns Indicate Divergent Strategies for Gene Regulation at the Chromatin Level**. *Plos Genet*, **7**.
33. Istrail, S. and Davidson, E.H. (2005) **Logic functions of the genomic cis-regulatory code**. *Proc Natl Acad Sci U S A*, **102**, 4954-4959.
34. Davidson, E.H. and Erwin, D.H. (2006) **Gene regulatory networks and the evolution of animal body plans**. *Science*, **311**, 796-800.
35. Beer, M.A. and Tavazoie, S. (2004) **Predicting gene expression from sequence**. *Cell*, **117**, 185-198.
36. Banerji, J., Rusconi, S. and Schaffner, W. (1981) **Expression of a Beta-Globin Gene Is Enhanced by Remote Sv40 DNA-Sequences**. *Cell*, **27**, 299-308.

37. Blackwood, E.M. and Kadonaga, J.T. (1998) **Going the distance: A current view of enhancer action**. *Science*, **281**, 60-63.
38. Gillies, S.D., Morrison, S.L., Oi, V.T. and Tonegawa, S. (1983) **A Tissue-Specific Transcription Enhancer Element Is Located in the Major Intron of a Rearranged Immunoglobulin Heavy-Chain Gene**. *Cell*, **33**, 717-728.
39. Pirrotta, V. and Gross, D.S. (2005) **Epigenetic silencing mechanisms in budding yeast and fruit fly: Different paths, same destinations**. *Mol Cell*, **18**, 395-398.
40. Ogbourne, S. and Antalis, T.M. (1998) **Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes**. *Biochem J*, **331**, 1-14.
41. Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E. and de Graaff, E. (2003) **A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly**. *Hum Mol Genet*, **12**, 1725-1735.
42. Dong, X., Navratilova, P., Fredman, D., Drivenes, O., Becker, T.S. and Lenhard, B. (2010) **Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons**. *Nucleic Acids Res*, **38**, 1071-1085.
43. Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K. et al. (2005) **Highly conserved non-coding sequences are associated with vertebrate development**. *PLoS Biol*, **3**, e7.
44. Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D. et al. (2006) **In vivo enhancer analysis of human conserved non-coding sequences**. *Nature*, **444**, 499-502.
45. Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A.Z., Engstrom, P.G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K. et al. (2007) **Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates**. *Genome Res*, **17**, 545-555.
46. Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L. and McCallion, A.S. (2006) **Conservation of RET regulatory function from human to zebrafish without sequence similarity**. *Science*, **312**, 276-279.
47. Ellingsen, S., Laplante, M.A., Konig, M., Kikuta, H., Furmanek, T., Hoivik, E.A. and Becker, T.S. (2005) **Large-scale enhancer detection in the zebrafish genome**. *Development*, **132**, 3799-3811.
48. Whitfield, T.W., Wang, J., Collins, P.J., Partridge, E.C., Aldred, S.F., Trinklein, N.D., Myers, R.M. and Weng, Z. (2012) **Functional analysis of transcription factor binding sites in human promoters**. *Genome Biol*, **13**, R50.
49. Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M. and Sidow, A. (2008) **Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data**. *Nat Methods*, **5**, 829-834.
50. Zhang, C., Xuan, Z., Otto, S., Hover, J.R., McCorkle, S.R., Mandel, G. and Zhang, M.Q. (2006) **A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome**. *Nucleic Acids Res*, **34**, 2238-2246.

51. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. and de Laat, W. (2002) **Looping and interaction between hypersensitive sites in the active beta-globin locus**. *Mol Cell*, **10**, 1453-1465.
52. de Laat, W. and Grosveld, F. (2003) **Spatial organization of gene expression: the active chromatin hub**. *Chromosome Res*, **11**, 447-459.
53. Lanzuolo, C., Roure, V., Dekker, J., Bantignies, F. and Orlando, V. (2007) **Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex**. *Nat Cell Biol*, **9**, 1167-1174.
54. Tiwari, V.K., McGarvey, K.M., Licchesi, J.D., Ohm, J.E., Herman, J.G., Schubeler, D. and Baylin, S.B. (2008) **PcG proteins, DNA methylation, and gene repression by chromatin looping**. *PLoS Biol*, **6**, 2911-2927.
55. West, A.G., Gaszner, M. and Felsenfeld, G. (2002) **Insulators: many functions, many mechanisms**. *Genes Dev*, **16**, 271-288.
56. Gaszner, M. and Felsenfeld, G. (2006) **Insulators: exploiting transcriptional and epigenetic mechanisms**. *Nat Rev Genet*, **7**, 703-713.
57. Maston, G.A., Evans, S.K. and Green, M.R. (2006) **Transcriptional regulatory elements in the human genome**. *Annu Rev Genomics Hum Genet*, **7**, 29-59.
58. Geyer, P.K., Spana, C. and Corces, V.G. (1986) **On the Molecular Mechanism of Gypsy-Induced Mutations at the Yellow Locus of Drosophila-Melanogaster**. *Embo J*, **5**, 2657-2662.
59. Chung, J.H., Whiteley, M. and Felsenfeld, G. (1993) **A 5' Element of the Chicken Beta-Globin Domain Serves as an Insulator in Human Erythroid-Cells and Protects against Position Effect in Drosophila**. *Cell*, **74**, 505-514.
60. Chung, J.H., Bell, A.C. and Felsenfeld, G. (1997) **Characterization of the chicken beta-globin insulator**. *P Natl Acad Sci USA*, **94**, 575-580.
61. Bell, A.C., West, A.G. and Felsenfeld, G. (1999) **The protein CTCF is required for the enhancer blocking activity of vertebrate insulators**. *Cell*, **98**, 387-396.
62. Pikaart, M.I., Recillas-Targa, F. and Felsenfeld, G. (1998) **Loss of transcriptional activity of a transgene is accompanied by DNA methylation and histone deacetylation and is prevented by insulators**. *Gene Dev*, **12**, 2852-2862.
63. Recillas-Targa, F., Pikaart, M.J., Burgess-Beusse, B., Bell, A.C., Litt, M.D., West, A.G., Gaszner, M. and Felsenfeld, G. (2002) **Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities**. *P Natl Acad Sci USA*, **99**, 6883-6888.
64. Li, Q.L., Zhang, M.H., Duan, Z.J. and Stamatoyannopoulos, G. (1999) **Structural analysis and mapping of DNase I hypersensitivity of HS5 of the beta-globin locus control region**. *Genomics*, **61**, 183-193.
65. Bell, A.C. and Felsenfeld, G. (2000) **Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene**. *Nature*, **405**, 482-485.
66. Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M. and Tilghman, S.M. (2000) **CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus**. *Nature*, **405**, 486-489.
67. Klenova, E.M., Nicolas, R.H., Paterson, H.F., Carne, A.F., Heath, C.M., Goodwin, G.H., Neiman, P.E. and Lobanenkov, V.V. (1993) **Ctcf, a**

Conserved Nuclear Factor Required for Optimal Transcriptional Activity of the Chicken C-Myc Gene, Is an 11-Zn-Finger Protein Differentially Expressed in Multiple Forms. *Molecular and Cellular Biology*, **13**, 7612-7624.

68. Moon, H., Filippova, G., Loukinov, D., Pugacheva, E., Chen, Q., Smith, S.T., Munhall, A., Grewe, B., Bartkuhn, M., Arnold, R. et al. (2005) **CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator**. *Embo Rep*, **6**, 165-170.
69. Burgess-Beusse, B., Farrell, C., Gaszner, M., Litt, M., Mutskov, V., Recillas-Targa, F., Simpson, M., West, A. and Felsenfeld, G. (2002) **The insulation of genes from external enhancers and silencing chromatin**. *Proc Natl Acad Sci U S A*, **99 Suppl 4**, 16433-16437.
70. Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V. and Ren, B. (2007) **Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome**. *Cell*, **128**, 1231-1245.
71. Xie, X., Mikkelsen, T.S., Gnirke, A., Lindblad-Toh, K., Kellis, M. and Lander, E.S. (2007) **Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites**. *Proc Natl Acad Sci U S A*, **104**, 7145-7150.
72. Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.Y., Cui, K. and Zhao, K. (2009) **Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains**. *Genome Res*, **19**, 24-32.
73. Splinter, E., Heath, H., Kooren, J., Palstra, R.J., Klous, P., Grosveld, F., Galjart, N. and de Laat, W. (2006) **CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus**. *Gene Dev*, **20**, 2349-2354.
74. Kurukuti, S., Tiwari, V.K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z.H., Lobanenkov, V., Reik, W. and Ohlsson, R. (2006) **CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2**. *P Natl Acad Sci USA*, **103**, 10684-10689.
75. Ribeiro de Almeida, C., Stadhouders, R., de Bruijn, M.J., Bergen, I.M., Thongjuea, S., Lenhard, B., van Ijcken, W., Grosveld, F., Galjart, N., Soler, E. et al. (2011) **The DNA-binding protein CTCF limits proximal Vkappa recombination and restricts kappa enhancer interactions to the immunoglobulin kappa light chain locus**. *Immunity*, **35**, 501-513.
76. Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W., Ye, C., Ping, J.L., Mulawadi, F. et al. (2011) **CTCF-mediated functional chromatin interactome in pluripotent cells**. *Nat Genet*, **43**, 630-638.
77. Fu, Y., Sinha, M., Peterson, C.L. and Weng, Z. (2008) **The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome**. *Plos Genet*, **4**, e1000138.
78. Grosveld, F., van Assendelft, G.B., Greaves, D.R. and Kollias, G. (1987) **Position-independent, high-level expression of the human beta-globin gene in transgenic mice**. *Cell*, **51**, 975-985.
79. Tanimoto, K., Liu, Q., Bungert, J. and Engel, J.D. (1999) **Effects of altered gene order or orientation of the locus control region on human beta-globin gene expression in mice**. *Nature*, **398**, 344-348.

80. Li, Q., Peterson, K.R., Fang, X. and Stamatoyannopoulos, G. (2002) **Locus control regions**. *Blood*, **100**, 3077-3086.
81. Lang, G., Mamalaki, C., Greenberg, D., Yannoutsos, N. and Kioussis, D. (1991) **Deletion analysis of the human CD2 gene locus control region in transgenic mice**. *Nucleic Acids Res*, **19**, 5851-5856.
82. Lee, G.R., Fields, P.E., Griffin, T.J. and Flavell, R.A. (2003) **Regulation of the Th2 cytokine locus by a locus control region**. *Immunity*, **19**, 145-153.
83. Diaz, P., Cado, D. and Winoto, A. (1994) **A locus control region in the T cell receptor alpha/delta locus**. *Immunity*, **1**, 207-217.
84. Madisen, L. and Groudine, M. (1994) **Identification of a locus control region in the immunoglobulin heavy-chain locus that deregulates c-myc expression in plasmacytoma and Burkitt's lymphoma cells**. *Genes Dev*, **8**, 2212-2226.
85. Wurster, A.L., Siu, G., Leiden, J.M. and Hedrick, S.M. (1994) **Elf-1 binds to a critical element in a second CD4 enhancer**. *Mol Cell Biol*, **14**, 6452-6463.
86. Bender, M.A., Ragoczy, T., Lee, J., Byron, R., Telling, A., Dean, A. and Groudine, M. (2012) **The hypersensitive sites of the murine beta-globin locus control region act independently to affect nuclear localization and transcriptional elongation**. *Blood*, **119**, 3820-3827.
87. Maston, G.A., Landt, S.G., Snyder, M. and Green, M.R. (2012) **Characterization of Enhancer Function from Genome-Wide Analyses**. *Annu Rev Genom Hum G*, **13**, 29-57.
88. Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V. et al. (2003) **TRANSFAC: transcriptional regulation, from patterns to profiles**. *Nucleic Acids Res*, **31**, 374-378.
89. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W. and Lenhard, B. (2004) **JASPAR: an open-access database for eukaryotic transcription factor binding profiles**. *Nucleic Acids Res*, **32**, D91-94.
90. Tuerk, C. and Gold, L. (1990) **Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase**. *Science*, **249**, 505-510.
91. Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D., Snyder, M., Young, R.A. and Bulyk, M.L. (2004) **Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays**. *Nat Genet*, **36**, 1331-1339.
92. Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) **Genome-wide mapping of in vivo protein-DNA interactions**. *Science*, **316**, 1497-1502.
93. Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E. et al. (2000) **Genome-wide location and function of DNA binding proteins**. *Science*, **290**, 2306-2309.
94. Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W. and Sandelin, A. (2010) **JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles**. *Nucleic Acids Res*, **38**, D105-110.
95. Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) **MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data**. *Nucleic Acids Res*, **23**, 4878-4884.

96. Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V. and Wingender, E. (2003) **MATCH: A tool for searching transcription factor binding sites in DNA sequences**. *Nucleic Acids Res*, **31**, 3576-3579.
97. Bailey, T.L. and Gribskov, M. (1998) **Combining evidence using p-values: application to sequence homology searches**. *Bioinformatics*, **14**, 48-54.
98. Lenhard, B. and Wasserman, W.W. (2002) **TFBS: Computational framework for transcription factor binding site analysis**. *Bioinformatics*, **18**, 1135-1136.
99. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. et al. (2004) **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol*, **5**, R80.
100. Blanchette, M. and Tompa, M. (2003) **FootPrinter: A program designed for phylogenetic footprinting**. *Nucleic Acids Res*, **31**, 3840-3842.
101. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S. et al. (2005) **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res*, **15**, 1034-1050.
102. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S. and Sidow, A. (2005) **Distribution and intensity of constraint in mammalian genomic sequence**. *Genome Res*, **15**, 901-913.
103. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S. and Haussler, D. (2004) **Ultraconserved elements in the human genome**. *Science*, **304**, 1321-1325.
104. Kunarso, G., Chia, N.Y., Jeyakani, J., Hwang, C., Lu, X.Y., Chan, Y.S., Ng, H.H. and Bourque, G. (2010) **Transposable elements have rewired the core regulatory network of human embryonic stem cells**. *Nat Genet*, **42**, 631-U111.
105. Blow, M.J., McCulley, D.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F. et al. (2010) **ChIP-Seq identification of weakly conserved heart enhancers**. *Nat Genet*, **42**, 806-810.
106. Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S. et al. (2010) **Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding**. *Science*, **328**, 1036-1040.
107. Odom, D.T., Dowell, R.D., Jacobsen, E.S., Gordon, W., Danford, T.W., MacIsaac, K.D., Rolfe, P.A., Conboy, C.M., Gifford, D.K. and Fraenkel, E. (2007) **Tissue-specific transcriptional regulation has diverged significantly between human and mouse**. *Nat Genet*, **39**, 730-732.
108. Carey, M.F., Peterson, C.L. and Smale, S.T. (2009) **Chromatin immunoprecipitation (ChIP)**. *Cold Spring Harb Protoc*, **2009**, pdb prot5279.
109. Lieb, J.D., Liu, X., Botstein, D. and Brown, P.O. (2001) **Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association**. *Nat Genet*, **28**, 327-334.
110. Weinmann, A.S., Yan, P.S., Oberley, M.J., Huang, T.H. and Farnham, P.J. (2002) **Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis**. *Genes Dev*, **16**, 235-244.

111. Furey, T.S. (2012) **ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions**. *Nat Rev Genet*, **13**, 840-852.
112. Hao, H. (2012) **Genome-wide occupancy analysis by ChIP-chip and ChIP-Seq**. *Adv Exp Med Biol*, **723**, 753-759.
113. Park, P.J. (2009) **ChIP-seq: advantages and challenges of a maturing technology**. *Nat Rev Genet*, **10**, 669-680.
114. Farnham, P.J. (2009) **Insights from genomic profiling of transcription factors**. *Nat Rev Genet*, **10**, 605-616.
115. Zhang, X., Guo, C., Chen, Y., Shulha, H.P., Schnetz, M.P., LaFramboise, T., Bartels, C.F., Markowitz, S., Weng, Z., Scacheri, P.C. et al. (2008) **Epitope tagging of endogenous proteins for genome-wide ChIP-chip studies**. *Nat Methods*, **5**, 163-165.
116. He, A. and Pu, W.T. (2010) **Genome-wide location analysis by pull down of in vivo biotinylated transcription factors**. *Curr Protoc Mol Biol*, **Chapter 21**, Unit 21 20.
117. Kim, J.S., Bonifant, C., Bunz, F., Lane, W.S. and Waldman, T. (2008) **Epitope tagging of endogenous genes in diverse human cell lines**. *Nucleic Acids Res*, **36**, e127.
118. Benoukraf, T., Cauchy, P., Fenouil, R., Jeanniard, A., Koch, F., Jaeger, S., Thieffry, D., Imbert, J., Andrau, J.C., Spicuglia, S. et al. (2009) **CoCAS: a ChIP-on-chip analysis suite**. *Bioinformatics*, **25**, 954-955.
119. Johnson, W.E., Li, W., Meyer, C.A., Gottardo, R., Carroll, J.S., Brown, M. and Liu, X.S. (2006) **Model-based analysis of tiling-arrays for ChIP-chip**. *Proc Natl Acad Sci U S A*, **103**, 12457-12462.
120. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. et al. (2008) **Model-based analysis of ChIP-Seq (MACS)**. *Genome Biol*, **9**, R137.
121. Rozowsky, J., Euskirchen, G., Auerbach, R.K., Zhang, Z.D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M. and Gerstein, M.B. (2009) **PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls**. *Nat Biotechnol*, **27**, 66-75.
122. Jothi, R., Cuddapah, S., Barski, A., Cui, K. and Zhao, K. (2008) **Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data**. *Nucleic Acids Res*, **36**, 5221-5231.
123. Xu, H., Handoko, L., Wei, X., Ye, C., Sheng, J., Wei, C.L., Lin, F. and Sung, W.K. (2010) **A signal-noise model for significance analysis of ChIP-seq with negative control**. *Bioinformatics*, **26**, 1199-1204.
124. Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R. et al. (2012) **An integrated encyclopedia of DNA elements in the human genome**. *Nature*, **489**, 57-74.
125. Wang, J., Zhuang, J., Iyer, S., Lin, X.Y., Greven, M.C., Kim, B.H., Moore, J., Pierce, B.G., Dong, X., Virgil, D. et al. (2013) **Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium**. *Nucleic Acids Res*, **41**, D171-D176.
126. Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P. et al. (2012) **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia**. *Genome Res*, **22**, 1813-1831.

127. Egelhofer, T.A., Minoda, A., Klugman, S., Lee, K., Kolasinska-Zwierz, P., Alekseyenko, A.A., Cheung, M.S., Day, D.S., Gadel, S., Gorchakov, A.A. et al. (2011) **An assessment of histone-modification antibody quality**. *Nat Struct Mol Biol*, **18**, 91-93.
128. Adli, M. and Bernstein, B.E. (2011) **Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq**. *Nat Protoc*, **6**, 1656-1668.
129. Shankaranarayanan, P., Mendoza-Parra, M.A., Walia, M., Wang, L., Li, N., Trindade, L.M. and Gronemeyer, H. (2011) **Single-tube linear DNA amplification (LinDA) for robust ChIP-seq**. *Nat Methods*, **8**, 565-567.
130. Rhee, H.S. and Pugh, B.F. (2011) **Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution**. *Cell*, **147**, 1408-1419.
131. Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J. et al. (2008) **Integration of external signaling pathways with the core transcriptional network in embryonic stem cells**. *Cell*, **133**, 1106-1117.
132. Wilson, N.K., Foster, S.D., Wang, X., Knezevic, K., Schutte, J., Kaimakis, P., Chilarska, P.M., Kinston, S., Ouwehand, W.H., Dzierzak, E. et al. (2010) **Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators**. *Cell Stem Cell*, **7**, 532-544.
133. Soler, E., Andrieu-Soler, C., de Boer, E., Bryne, J.C., Thongjuea, S., Stadhouders, R., Palstra, R.J., Stevens, M., Kockx, C., van IJcken, W. et al. (2010) **The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation**. *Gene Dev*, **24**, 277-289.
134. Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y. et al. (2012) **Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors**. *Genome Res*, **22**, 1798-1812.
135. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. and Snyder, M. (2012) **Linking disease associations with regulatory information in the human genome**. *Genome Res*, **22**, 1748-1759.
136. Wu, C. (1980) **The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I**. *Nature*, **286**, 854-860.
137. Wu, C., Wong, Y.C. and Elgin, S.C. (1979) **The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity**. *Cell*, **16**, 807-814.
138. Fritton, H.P., Sippel, A.E. and Igo-Kemenes, T. (1983) **Nuclease-hypersensitive sites in the chromatin domain of the chicken lysozyme gene**. *Nucleic Acids Res*, **11**, 3467-3485.
139. Gross, D.S. and Garrard, W.T. (1988) **Nuclease hypersensitive sites in chromatin**. *Annu Rev Biochem*, **57**, 159-197.
140. Crawford, G.E., Davis, S., Scacheri, P.C., Renaud, G., Halawi, M.J., Erdos, M.R., Green, R., Meltzer, P.S., Wolfsberg, T.G. and Collins, F.S. (2006) **DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays**. *Nature Methods*, **3**, 503-509.
141. Sabo, P.J., Kuehn, M.S., Thurman, R., Johnson, B.E., Johnson, E.M., Hua, C., Man, Y., Rosenzweig, E., Goldy, J., Haydock, A. et al. (2006) **Genome-scale**

mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nature Methods*, **3**, 511-518.

142. Xi, H., Shulha, H.P., Lin, J.M., Vales, T.R., Fu, Y., Bodine, D.M., McKay, R.D., Chenoweth, J.G., Tesar, P.J., Furey, T.S. et al. (2007) **Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome**. *Plos Genet*, **3**, e136.
143. Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. and Crawford, G.E. (2008) **High-resolution mapping and characterization of open chromatin across the genome**. *Cell*, **132**, 311-322.
144. Thomas, S., Li, X.Y., Sabo, P.J., Sandstrom, R., Thurman, R.E., Canfield, T.K., Giste, E., Fisher, W., Hammonds, A., Celniker, S.E. et al. (2011) **Dynamic reprogramming of chromatin accessibility during Drosophila embryo development**. *Genome Biol*, **12**, R43.
145. Ling, G., Sugathan, A., Mazor, T., Fraenkel, E. and Waxman, D.J. (2010) **Unbiased, genome-wide in vivo mapping of transcriptional regulatory elements reveals sex differences in chromatin structure associated with sex-specific liver gene expression**. *Mol Cell Biol*, **30**, 5531-5544.
146. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B. et al. (2012) **The accessible chromatin landscape of the human genome**. *Nature*, **489**, 75-82.
147. Song, L.Y., Zhang, Z.C., Grasfeder, L.L., Boyle, A.P., Giresi, P.G., Lee, B.K., Sheffield, N.C., Graf, S., Huss, M., Keefe, D. et al. (2011) **Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity**. *Genome Res*, **21**, 1757-1767.
148. Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. and Lieb, J.D. (2007) **FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin**. *Genome Res*, **17**, 877-885.
149. Auerbach, R.K., Euskirchen, G., Rozowsky, J., Lamarre-Vincent, N., Moqtaderi, Z., Lefrancois, P., Struhl, K., Gerstein, M. and Snyder, M. (2009) **Mapping accessible chromatin regions using Sono-Seq**. *P Natl Acad Sci USA*, **106**, 14926-14931.
150. Rudkin, G.T. and Stollar, B.D. (1977) **High resolution detection of DNA-RNA hybrids in situ by indirect immunofluorescence**. *Nature*, **265**, 472-473.
151. Cremer, T. and Cremer, C. (2001) **Chromosome territories, nuclear architecture and gene regulation in mammalian cells**. *Nat Rev Genet*, **2**, 292-301.
152. Cremer, T., Cremer, M., Dietzel, S., Muller, S., Solovei, I. and Fakan, S. (2006) **Chromosome territories--a functional nuclear landscape**. *Curr Opin Cell Biol*, **18**, 307-316.
153. Cremer, T. and Cremer, M. (2010) **Chromosome territories**. *Cold Spring Harb Perspect Biol*, **2**, a003889.
154. Pombo, A. and Branco, M.R. (2007) **Functional organisation of the genome during interphase**. *Curr Opin Genet Dev*, **17**, 451-455.
155. Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002) **Capturing chromosome conformation**. *Science*, **295**, 1306-1311.
156. Murrell, A., Heeson, S. and Reik, W. (2004) **Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops**. *Nat Genet*, **36**, 889-893.

157. Spilianakis, C.G. and Flavell, R.A. (2004) **Long-range intrachromosomal interactions in the T helper type 2 cytokine locus**. *Nat Immunol*, **5**, 1017-1027.
158. Vernimmen, D., De Gobbi, M., Sloane-Stanley, J.A., Wood, W.G. and Higgs, D.R. (2007) **Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression**. *Embo J*, **26**, 2041-2051.
159. Spilianakis, C.G., Lalioti, M.D., Town, T., Lee, G.R. and Flavell, R.A. (2005) **Interchromosomal associations between alternatively expressed loci**. *Nature*, **435**, 637-645.
160. Lomvardas, S., Barnea, G., Pisapia, D.J., Mendelsohn, M., Kirkland, J. and Axel, R. (2006) **Interchromosomal interactions and olfactory receptor choice**. *Cell*, **126**, 403-413.
161. Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. and de Laat, W. (2006) **Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)**. *Nat Genet*, **38**, 1348-1354.
162. Zhao, Z., Tavoosidana, G., Sjolinder, M., Gondor, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U. et al. (2006) **Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions**. *Nat Genet*, **38**, 1341-1347.
163. Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C. et al. (2006) **Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements**. *Genome Res*, **16**, 1299-1309.
164. Tiwari, V.K., Cope, L., McGarvey, K.M., Ohm, J.E. and Baylin, S.B. (2008) **A novel 6C assay uncovers Polycomb-mediated higher order chromatin conformations**. *Genome Res*, **18**, 1171-1179.
165. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O. et al. (2009) **Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome**. *Science*, **326**, 289-293.
166. de Wit, E. and de Laat, W. (2012) **A decade of 3C technologies: insights into nuclear organization**. *Gene Dev*, **26**, 11-24.
167. Stadhouders, R., Kolovos, P., Brouwer, R., Zuin, J., van den Heuvel, A., Kockx, C., Palstra, R.J., Wendt, K.S., Grosveld, F., van Ijcken, W. et al. (2013) **Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions**. *Nat Protoc*, **8**, 509-524.
168. van de Werken, H.J., Landan, G., Holwerda, S.J., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B.A. et al. (2012) **Robust 4C-seq data analysis to screen for regulatory DNA interactions**. *Nat Methods*, **9**, 969-972.
169. Lieu, Y.K. and Reddy, E.P. (2009) **Conditional c-myb knockout in adult hematopoietic stem cells leads to loss of self-renewal due to impaired proliferation and accelerated differentiation**. *P Natl Acad Sci USA*, **106**, 21689-21694.

170. Ramsay, R.G. and Gonda, T.J. (2008) **MYB function in normal and cancer cells**. *Nat Rev Cancer*, **8**, 523-534.
171. Vegiopoulos, A., Garcia, P., Emambokus, N. and Frampton, J. (2006) **Coordination of erythropoiesis by the transcription factor c-Myb**. *Blood*, **107**, 4703-4710.
172. Tallack, M.R., Whitington, T., Yuen, W.S., Wainwright, E.N., Keys, J.R., Gardiner, B.B., Nourbakhsh, E., Cloonan, N., Grimmond, S.M., Bailey, T.L. et al. (2010) **A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells**. *Genome Res*, **20**, 1052-1063.
173. Kassouf, M.T., Hughes, J.R., Taylor, S., McGowan, S.J., Soneji, S., Green, A.L., Vyas, P. and Porcher, C. (2010) **Genome-wide identification of TAL1's functional targets: Insights into its mechanisms of action in primary erythroid cells**. *Genome Res*, **20**, 1064-1083.
174. Stadhouders, R., Thongjuea, S., Andrieu-Soler, C., Palstra, R.J., Bryne, J.C., van den Heuvel, A., Stevens, M., de Boer, E., Kockx, C., van der Sloot, A. et al. (2012) **Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development**. *Embo J*, **31**, 986-999.
175. Pearson, J.C., Lemons, D. and McGinnis, W. (2005) **Modulating Hox gene functions during animal body patterning**. *Nat Rev Genet*, **6**, 893-904.
176. Chambeyron, S. and Bickmore, W.A. (2004) **Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription**. *Genes Dev*, **18**, 1119-1130.
177. Noordermeer, D., Leleu, M., Splinter, E., Rougemont, J., De Laat, W. and Duboule, D. (2011) **The Dynamic Architecture of Hox Gene Clusters**. *Science*, **334**, 222-225.
178. Tolhuis, B., Blom, M., Kerkhoven, R.M., Pagie, L., Teunissen, H., Nieuwland, M., Simonis, M., de Laat, W., van Lohuizen, M. and van Steensel, B. (2011) **Interactions among Polycomb Domains Are Guided by Chromosome Architecture**. *Plos Genet*, **7**.
179. Branco, M.R. and Pombo, A. (2006) **Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations**. *PLoS Biol*, **4**, e138.
180. Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J.A., Umlauf, D., Dimitrova, D.S. et al. (2010) **Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells**. *Nat Genet*, **42**, 53-U71.
181. Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W. et al. (2004) **Active genes dynamically colocalize to shared sites of ongoing transcription**. *Nat Genet*, **36**, 1065-1071.
182. Papantonis, A., Larkin, J.D., Wada, Y., Ohta, Y., Ihara, S., Kodama, T. and Cook, P.R. (2010) **Active RNA polymerases: mobile or immobile molecular machines?** *PLoS Biol*, **8**, e1000419.
183. Palstra, R.J., Simonis, M., Klous, P., Brasset, E., Eijkelkamp, B. and de Laat, W. (2008) **Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription**. *PLoS One*, **3**, e1661.
184. Tumbar, T., Sudlow, G. and Belmont, A.S. (1999) **Large-scale chromatin unfolding and remodeling induced by VP16 acidic activation domain**. *J Cell Biol*, **145**, 1341-1354.

185. Splinter, E., de Wit, E., Nora, E.P., Klous, P., van de Werken, H.J.G., Zhu, Y., Kaaij, L.J.T., van IJcken, W., Gribnau, J., Heard, E. et al. (2011) **The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA**. *Gene Dev*, **25**, 1371-1383.
186. Senner, C.E. and Brockdorff, N. (2009) **Xist gene regulation at the onset of X inactivation**. *Current Opinion in Genetics & Development*, **19**, 122-126.
187. Chow, J. and Heard, E. (2009) **X inactivation and the complexities of silencing a sex chromosome**. *Current Opinion in Cell Biology*, **21**, 359-366.
188. Orkin, S.H. and Zon, L.I. (2008) **Hematopoiesis: an evolving paradigm for stem cell biology**. *Cell*, **132**, 631-644.
189. Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T. et al. (2011) **Densely interconnected transcriptional circuits control cell states in human hematopoiesis**. *Cell*, **144**, 296-309.
190. Cantor, A.B. and Orkin, S.H. (2002) **Transcriptional regulation of erythropoiesis: an affair involving multiple partners**. *Oncogene*, **21**, 3368-3376.
191. Chao, M.P., Seita, J. and Weissman, I.L. (2008) **Establishment of a normal hematopoietic and leukemia stem cell hierarchy**. *Cold Spring Harb Symp Quant Biol*, **73**, 439-449.
192. Miranda-Saavedra, D. and Gottgens, B. (2008) **Transcriptional regulatory networks in haematopoiesis**. *Curr Opin Genet Dev*, **18**, 530-535.
193. Wilson, N.K., Calero-Nieto, F.J., Ferreira, R. and Gottgens, B. (2011) **Transcriptional regulation of haematopoietic transcription factors**. *Stem Cell Res Ther*, **2**, 6.
194. Wilson, N.K., Tijssen, M.R. and Gottgens, B. (2011) **Deciphering transcriptional control mechanisms in hematopoiesis:the impact of high-throughput sequencing technologies**. *Exp Hematol*, **39**, 961-968.
195. Kerenyi, M.A. and Orkin, S.H. (2010) **Networking erythropoiesis**. *J Exp Med*, **207**, 2537-2541.
196. Fujiwara, T., O'Geen, H., Keles, S., Blahnik, K., Linnemann, A.K., Kang, Y.A., Choi, K., Farnham, P.J. and Bresnick, E.H. (2009) **Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy**. *Mol Cell*, **36**, 667-681.
197. Cheng, Y., Wu, W., Kumar, S.A., Yu, D., Deng, W., Tripic, T., King, D.C., Chen, K.B., Zhang, Y., Drautz, D. et al. (2009) **Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression**. *Genome Res*, **19**, 2172-2184.
198. Yu, M., Riva, L., Xie, H., Schindler, Y., Moran, T.B., Cheng, Y., Yu, D., Hardison, R., Weiss, M.J., Orkin, S.H. et al. (2009) **Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis**. *Mol Cell*, **36**, 682-695.
199. Xu, J., Shao, Z., Glass, K., Bauer, D.E., Pinello, L., Van Handel, B., Hou, S., Stamatoyannopoulos, J.A., Mikkola, H.K., Yuan, G.C. et al. (2012) **Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis**. *Dev Cell*, **23**, 796-811.
200. Tripic, T., Deng, W., Cheng, Y., Zhang, Y., Vakoc, C.R., Gregory, G.D., Hardison, R.C. and Blobel, G.A. (2009) **SCL and associated proteins**

distinguish active from repressive GATA transcription factor complexes. *Blood*, **113**, 2191-2201.

201. Kassouf, M.T., Hughes, J.R., Taylor, S., McGowan, S.J., Soneji, S., Green, A.L., Vyas, P. and Porcher, C. (2010) **Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells**. *Genome Res*, **20**, 1064-1083.
202. Meier, N., Krpic, S., Rodriguez, P., Strouboulis, J., Monti, M., Krijgsveld, J., Gering, M., Patient, R., Hostert, A. and Grosveld, F. (2006) **Novel binding partners of Ldb1 are required for haematopoietic development**. *Development*, **133**, 4913-4923.
203. Miller, I.J. and Bieker, J.J. (1993) **A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the Kruppel family of nuclear proteins**. *Mol Cell Biol*, **13**, 2776-2786.
204. Tallack, M.R., Magor, G.W., Dartigues, B., Sun, L., Huang, S., Fittock, J.M., Fry, S.V., Glazov, E.A., Bailey, T.L. and Perkins, A.C. (2012) **Novel roles for KLF1 in erythropoiesis revealed by mRNA-seq**. *Genome Res*, **22**, 2385-2398.
205. Wadman, I.A., Osada, H., Grutz, G.G., Agulnick, A.D., Westphal, H., Forster, A. and Rabbitts, T.H. (1997) **The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins**. *Embo J*, **16**, 3145-3157.
206. Brand, M., Ranish, J.A., Kummer, N.T., Hamilton, J., Igarashi, K., Francastel, C., Chi, T.H., Crabtree, G.R., Aebersold, R. and Groudine, M. (2004) **Dynamic changes in transcription factor complexes during erythroid differentiation revealed by quantitative proteomics**. *Nat Struct Mol Biol*, **11**, 73-80.
207. Murre, C., Voronova, A. and Baltimore, D. (1991) **B-cell- and myocyte-specific E2-box-binding factors contain E12/E47-like subunits**. *Mol Cell Biol*, **11**, 1156-1160.
208. Akalin, A., Fredman, D., Arner, E., Dong, X., Bryne, J.C., Suzuki, H., Daub, C.O., Hayashizaki, Y. and Lenhard, B. (2009) **Transcriptional features of genomic regulatory blocks**. *Genome Biol*, **10**, R38.
209. Portales-Casamar, E., Arenillas, D., Lim, J., Swanson, M.I., Jiang, S., McCallum, A., Kirov, S. and Wasserman, W.W. (2009) **The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences**. *Nucleic Acids Res*, **37**, D54-60.
210. Thein, S.L., Menzel, S., Peng, X., Best, S., Jiang, J., Close, J., Silver, N., Gerovasilli, A., Ping, C., Yamaguchi, M. et al. (2007) **Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults**. *Proc Natl Acad Sci U S A*, **104**, 11346-11351.
211. Lettre, G., Sankaran, V.G., Bezerra, M.A., Araujo, A.S., Uda, M., Sanna, S., Cao, A., Schlessinger, D., Costa, F.F., Hirschhorn, J.N. et al. (2008) **DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease**. *Proc Natl Acad Sci U S A*, **105**, 11869-11874.
212. Farrell, J.J., Sherva, R.M., Chen, Z.Y., Luo, H.Y., Chu, B.F., Ha, S.Y., Li, C.K., Lee, A.C., Li, R.C., Yuen, H.L. et al. (2011) **A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression**. *Blood*, **117**, 4935-4945.

213. Sanyal, A., Lajoie, B.R., Jain, G. and Dekker, J. (2012) **The long-range interaction landscape of gene promoters**. *Nature*, **489**, 109-113.
214. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) **The human genome browser at UCSC**. *Genome Res*, **12**, 996-1006.
215. Merika, M., Williams, A.J., Chen, G., Collins, T. and Thanos, D. (1998) **Recruitment of CBP/p300 by the IFN beta enhanceosome is required for synergistic activation of transcription**. *Mol Cell*, **1**, 277-287.
216. Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A. et al. (2007) **Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome**. *Nat Genet*, **39**, 311-318.
217. Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F. et al. (2009) **ChIP-seq accurately predicts tissue-specific activity of enhancers**. *Nature*, **457**, 854-858.
218. Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A. and Wysocka, J. (2011) **A unique chromatin signature uncovers early developmental enhancers in humans**. *Nature*, **470**, 279-283.
219. Kasper, L.H., Boussouar, F., Ney, P.A., Jackson, C.W., Rehg, J., van Deursen, J.M. and Brindle, P.K. (2002) **A transcription-factor-binding surface of coactivator p300 is required for haematopoiesis**. *Nature*, **419**, 738-743.
220. Sandberg, M.L., Sutton, S.E., Pletcher, M.T., Wiltshire, T., Tarantino, L.M., Hogenesch, J.B. and Cooke, M.P. (2005) **c-Myb and p300 regulate hematopoietic stem cell proliferation and differentiation**. *Dev Cell*, **8**, 153-166.
221. Kimbrel, E.A., Lemieux, M.E., Xia, X., Davis, T.N., Rebel, V.I. and Kung, A.L. (2009) **Systematic in vivo structure-function analysis of p300 in hematopoiesis**. *Blood*, **114**, 4804-4812.
222. Wozniak, R.J., Keles, S., Lugus, J.J., Young, K.H., Boyer, M.E., Tran, T.M., Choi, K. and Bresnick, E.H. (2008) **Molecular hallmarks of endogenous chromatin complexes containing master regulators of hematopoiesis**. *Mol Cell Biol*, **28**, 6681-6694.
223. Cheng, Y., Wu, W.S., Kumar, S.A., Yu, D.N., Deng, W.L., Tripic, T., King, D.C., Chen, K.B., Zhang, Y., Drautz, D. et al. (2009) **Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression**. *Genome Res*, **19**, 2172-2184.
224. Soler, E., Andrieu-Soler, C., de Boer, E., Bryne, J.C., Thongjuea, S., Stadhouders, R., Palstra, R.J., Stevens, M., Kockx, C., van Ijcken, W. et al. (2010) **The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation**. *Genes Dev*, **24**, 277-289.
225. Kowalczyk, M.S., Hughes, J.R., Garrick, D., Lynch, M.D., Sharpe, J.A., Sloane-Stanley, J.A., McGowan, S.J., De Gobbi, M., Hosseini, M., Vernimmen, D. et al. (2012) **Intragenic Enhancers Act as Alternative Promoters**. *Mol Cell*, **45**, 447-458.
226. Morgan, M., Anders, S., Lawrence, M., Aboyoun, P., Pages, H. and Gentleman, R. (2009) **ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data**. *Bioinformatics*, **25**, 2607-2608.

227. Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T. and Carey, V.J. (2013) **Software for computing and annotating genomic ranges**. *PLoS Comput Biol*, **9**, e1003118.
228. Lawrence, M., Gentleman, R. and Carey, V. (2009) **rtracklayer: an R package for interfacing with genome browsers**. *Bioinformatics*, **25**, 1841-1842.
229. Bailey, T.L. and Elkan, C. (1994) **Fitting a mixture model by expectation maximization to discover motifs in biopolymers**. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, **2**, 28-36.
230. Sandelin, A., Bailey, P., Bruce, S., Engstrom, P.G., Klos, J.M., Wasserman, W.W., Ericson, J. and Lenhard, B. (2004) **Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes**. *BMC Genomics*, **5**, 99.
231. Tijssen, M.R., Cvejic, A., Joshi, A., Hannah, R.L., Ferreira, R., Forrai, A., Bellissimo, D.C., Oram, S.H., Smethurst, P.A., Wilson, N.K. et al. (2011) **Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators**. *Dev Cell*, **20**, 597-609.
232. Mylona, A., Andrieu-Soler, C., Thongjuea, S., Martella, A., Soler, E., Jorna, R., Hou, J., Kockx, C., van Ijcken, W., Lenhard, B. et al. (2013) **Genome-wide analysis shows that Ldb1 controls essential hematopoietic genes/pathways in mouse early development and reveals novel players in hematopoiesis**. *Blood*, **121**, 2902-2913.
233. Li, L., Freudenberg, J., Cui, K., Dale, R., Song, S.H., Dean, A., Zhao, K., Jothi, R. and Love, P.E. (2013) **Ldb1-nucleated transcription complexes function as primary mediators of global erythroid gene activation**. *Blood*, **121**, 4575-4585.