From the Science for Life Laboratory,
Department of Biosciences and Nutrition (BioNut),
Karolinska Institutet, Stockholm, Sweden

DECIPHERING THE TRANSCRIPTIONAL REGULATION CODE IN COLORECTAL CANCER GENOME

Jian Yan

Stockholm 2014

All constituent papers were reproduced with permission from the publishers.
Front cover shows the mouse model used to validate the regulatory elements found in this study (Photo from Dr. Inderpreet K. Sur, PhD)

Published and printed by Karolinska University Press
Box 200, SE-171 77 Stockholm, Sweden

ISBN 978-91-7549-486-9

To my dear parents

献给我最爱的父母

ABSTRACT

Colorectal cancer (CRC) is the third most common cancer type to threaten life for both men and women in the developed world. The molecular mechanism of CRC is very complicated and involves changes in different categories of biological processes, among which a large number of mutations affecting either coding sequence of transcription factors (TFs) or their binding sites in genome are included. This indicated the importance of fine-tuned transcriptional regulation in normal colon function and the significance of its disruption in tumorigenesis and metastasis. However, the comprehensive knowledge of transcriptional regulation in CRC is still inadequate, leaving obstacles for clinical prognosis and therapy.

In order to better understand the transcriptional regulation network in CRC, we designed a study including three individual projects to systematically investigate the mammalian transcriptional regulation *in vitro*, *ex vivo* and *in vivo*.

In the *in vitro* study, we performed the High Throughput Systematic Evolution of Ligands by Exponential Enrichment (HT-SELEX) for the vast majority of mammalian TFs in order to profile their DNA sequence binding specificities. Eventually, we obtained binding profiles for 303 human DNA binding domains (DBDs), 84 mouse DBDs and 151 human full length TFs, representing 411 different TFs in total, which exceeds any existing database for mammalian TF DNA binding motifs and provides rich information for research on transcriptional regulation. By analyzing this data, we also determined some factors affecting TF-DNA binding such as adjacent base stacking and DNA shape, and suggested two advanced models to improve the computational prediction of TF binding.

In the *ex vivo* study, we carried out chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) for over 500 different TFs in a single CRC cell line LoVo, and a relatively smaller scale in another CRC cell line GP5d. We observed that most TFs tended to bind to DNA forming highly dense clusters around cohesin and occupying an unexpectedly small fraction of human genome. Our data suggested that cohesin binding could function as a cellular memory to mark the TF binding sites and facilitate the quick re-establishment of TF binding within the limited time during each cycle of cell division.

To test the function of some TF clusters, we generated a conditional knock out mouse strain lacking a 1.3kb TF cluster fragment (Myc-335) 335 kb upstream of Myc gene transcription starting site (TSS). We discovered that Myc-335 was a tumor specific enhancer for Myc gene and it was dispensable for normal intestinal development and function.

The study greatly improved our knowledge of TF-DNA interaction and its biological function in the relevant fields of transcriptional regulation, cell cycle, epigenetics, epigenomics and cancer biology, and also provided the whole scientific community with enormous data sets for further analyses.

LIST OF PUBLICATIONS

I. **Yan J***, Enge M*, Whitington T, Dave K, Liu J, Sur I, Schmierer B, Jolma A, Kivioja T, Taipale M, Taipale J. Transcription Factor Binding in Human Cells Occurs in Dense Clusters Formed around Cohesin Anchor Sites. *Cell.* 2013 Aug 15; 154(4): 801-13.

II. Jolma A*, **Yan J***, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al. DNA-binding specificities of human transcription factors. *Cell.* 2013 Jan 17; 152(1-2): 327-39.

III. Sur IK, Hallikas O, Vähärautio A, **Yan J**, Turunen M, Enge M, Taipale M, Karhu A, Aaltonen LA, Taipale J. Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science*. 2012 Dec 7; 338(6112): 1360-3.

IV. Tuupanen S*, **Yan J***, Turunen M, Gylfe AE, Kaasinen E, Li L, Eng C, Culver DA, Kalady MF, Pennison MJ, et al. Characterization of the colorectal cancer-associated enhancer MYC-335 at 8q24: the role of rs67491583. *Cancer Genet.* 2012 Jan-Feb; 205(1-2): 25-33.

V. Huang Q, Whitington T, Gao P, Lindberg JF, Yang Y, Sun J, Väisänen MR, Szulkin R, Annala M, **Yan J**, et al. A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nat Genet.* 2014 Feb; 46(2): 126-35.

VI. Mäkinen N, Mehine M, Tolvanen J, Kaasinen E, Li Y, Lehtonen HJ, Gentile M, **Yan J**, Enge M, Taipale M, et al. MED12, the mediator complex subunit 12 gene, is mutated at high frequency in uterine leiomyomas. *Science.* 2011 Oct 14; 334(6053): 252-5.

VII. Wei GH, Badis G, Berger MF, Kivioja T, Palin K, Enge M, Bonke M, Jolma A, Varjosalo M, Gehrke AR, **Yan J**, et al. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J.* 2010 Jul 7; 29(13): 2147-60.

VIII. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, Taipale M, Vaquerizas JM, **Yan J**, Sillanpää MJ, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*. 2010 Jun; 20(6): 861-73.

IX. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Björklund M, Wei G, **Yan J**, Niittymäki I, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet.* 2009 Aug; 41(8): 885-90.

* Equal contribution to the work

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

TF	Transcription Factor
DNA	Deoxyribonucleic Acid
DBD	DNA-Binding Domain
GTF	General Transcription Factor
HT-SELEX	High Throughput Systematic Evolution of Ligands by Exponential Enrichment
ChIP-seq	Chromatin Immunoprecipitation followed by Sequencing
CRC	Colorectal Cancer
TSS	Transcription Starting Site
DHS	DNase I Hypersensitibity
SCPL	Sister Chromosome Proximity Ligation
RNA	Ribonucleic Acid
NRF	Nuclear Respiratory Factor
ETS	E26 Transformation-Specific
bHLH	basic-Helix-Loop-Helix
bZIP	beta-Zipper
TAD	Trans-Activating Domain
SSD	Signal Sensing Domain
IUPAC	International Union of Pure and Applied Chemistry
PBM	Protein Binding Microarray
EMSA	Electrophoretic Mobility Shift Assay
MITOMI	Mechanically Induced Trapping of Molecular Interaction
ENCODE	Encyclopedia of DNA Elements
dUTP	Deoxyuridine Triphosphate
UP element	Upstream Promoter element
TBP	TATA box Binding Protein
GWAS	Genome-Wide Association Studies
kb	kilobase
ncRNA-a	non-coding RNA activating
3C	Chromosome Confirmation Capture
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
CTCF	CCCTC-binding Factor

Hi-C	High-throughput 3C
SNP	Single Nucleotide Polymorphism
HAT	Histone Acetyltransferase
HDAC	Histone Deacetylase
PolII	RNA polymerase II
CBP	CREB Binding Protein
H3K27ac	Histone H3 Lysine 27 acetylation
H3K4me1	Histone H3 Lysine 4 monomethylation
H3K4me3	Histone H3 Lysine 4 trimethylation
PP2A/4/6	Protein Phosphotase 2A/4/6
K	Lysine
S	Serine
NCR	Nucleosome Core Particle
SMC1/2/3/4/5/6	Structural Maintenance of Chromosomes 1/2/3/4/5/6
WHO	World Health Organization
DMR	DNA Mismatch Repair
HNPCC	Hereditary Non-Polypsis Colorectal Cancer
MSI	Microsatellite Instability
APC	Adenomatous Polypsis Coli
Wnt	Wingless-Int
FAP	Familial Adenomatous Polyposis
MAPK	Mitogen-Activated Protein Kinase
PI3K	Phosphoinositide-3 Kinase
SOS	Son Of Sevenless
APCmin	APC multiple intestinal neoplasia
USA	the United States of America
ATCC	the American Type Culture Collection
ECACC	European Collection of Cell Cultures
HEK293FT	Human Embryonic Kidney 293 Fast-growing T-antigen transformed
cKO	conditional Knock Out
DMEM	Dulbecco's Modified Eagle Medium
FBS	Fetal Bovine Serum
PEI	Polyethylemine
RT	Room Temperature

PBS	Phosphate Buffered Saline
RIPA buffer	Radio-Immunoprecipitation Assay buffer
SDS	Sodium Dodecyl Sulfate
BEA	Karolinska Bioinformatics and Expression Analysis core facility
NEB	New England Biolabs
FDR	False Discovery Rate
KHTC	Karolinska High Throughput Center
PWM	Position Weight Matrix
PCR	Polymerase Chain Reaction
GEO	Gene Expression Omnibus
ENA	European Nucleotide Archive
CRE	cAMP Response Elements
UPRE	UPR-dependent *cis*-acting Elements
ADM	Adjacent Dinucleotide Model
QC	Quality Control
GO	Gene Ontology
EEL	Enhancer Element Locator

1 INTRODUCTION

1.1 TRANSCRIPTIONAL REGULATION IN HUMAN CELLS

1.1.1 Transcription Factors (TFs) and their binding models

Transcription factors (TFs) are a class of protein that can bind to the DNA sequence and regulate the transcriptional activity. The most important feature of TF is the DNA binding, which is accomplished by its DNA binding domain (DBD)[2,3]. TF also contains other domains such as trans-activating domain (TAD), which can interact with other protein or protein complex[4], or signal sensing domain (SSD) which senses external signal, i.a. vitamins A and D, and in turn transmits it to the transcription machinery and changes the gene expression activity[4].

There are mainly two types of TFs, the general transcription factor (GTF) and the sequence specific transcription factor. GTFs include the proteins involved in the transcription pre-initiation and elongation complex, such as transcription factor II A (TFIIA), transcription factor II D (TFIID), transcription factor II B (TFIIB), transcription factor II E (TFIIE), transcription factor II F (TFIIF) and transcription factor II H (TFIIH)[5]. They sequentially bind to the core promoter of a gene and recruit RNA polymerase II to initiate the transcription.

The main class of TF is the sequence specific transcription factors. There are about 1200-2000 TFs in the human genome (reviewed by [6]). They can be classified into 29 different main families based on the protein sequence and tertiary structure similarity of their DBDs, such as zinc finger, homeodomain, E26 transformation-specific (ETS), fork-head, nuclear hormone receptor, beta-zipper (bZIP), and basic helix-loop-helix (bHLH)[7-9]. The number of different TF family member could vary from over 600 of C2H2 zinc finger TFs[10] to only a single member of nuclear respiratory factor (NRF) family.

TF can sequence-specifically bind to DNA either via a direct contact of its side chain to the accessible edge of the base pair or an indirect contact to the DNA backbone[11]. The direct contact between the base and the amino acid allows the discrimination of sequence depending on the different patterns of hydrogen bond donors and acceptors or van der Waals interaction between amino acid and the methyl group of thymidine[12]. Although in most cases an amino acid could form hydrogen bonds with more than one base, the preference does exist. Luscombe et al.[12] found from the compilation of over 120 crystal structures of DNA-Protein complexes that the two most common cases are the hydrogen bonds formed between arginine and guanosine nucleotide and between asparagine or glutamine and adenosine nucleotide.

Although without direct dependence on the base sequence, the indirect contact of TF to the DNA backbone generally replies on the deformation of the standard B-form DNA, either with a broader or narrower major groove or being bent, which is also sequence relevant[13].

The large variety of interactions between DNA and TF amino acids avoid the extreme sequence specificity. The relatively loose specificity allows TF to bind to a number of different sites in the genome with various affinities but still much more specific than random. Such a continuum of the binding affinity allows cooperative binding of TF pairs with adjacent sites[14-16].

In order to describe the TF-DNA binding, different types of binding models are used. The simplest model is to directly describe the consensus sequence to which TF prefers to bind[17]. Since TF can bear variance of DNA sequence at each position, degenerate code (IUPAC code) could be applied in which M represents A or C, W for A or T, etc (more details see **Table 1.1**). This model even allows gaps between two half sites, such as the consensus sequence of AGGTGTGANNTCACACCT for T-box family member MGA[18].

Table 1.1 IUPAC code for nucleotide

Code letter	Bases covered
A	A
T	T
C	C
G	G
R	A, G
Y	C, T
W	A, T
S	C, G
M	A, C
K	G, T
B	C, G, T
D	A, G, T
H	A, C, T
V	A, C, G
N	A, C, G, T

However, taking into account TF could bind to different sites with distinct affinity, this model could only tell which sites TF is able to bind instead of quantitatively discriminating the affinity to different sites. To avoid this problem, the position weight matrix (PWM) model has been introduced and widely implemented to model the TF-DNA binding[19]. In this model, a numeric score is assigned to all possible bases at each position to describe the TF binding affinity or probability to the DNA sequence. Given any DNA sequence, one can calculate the score by summing up the scores of all positions for such sequence, and compare it with the score of a random sequence or of any other sequence for its competence to bind to the TF. The consensus sequence could also be easily described with PWM by concatenating the base with the highest score at each position. For example, the PWM model for TF BARHL2 (**Table 1.2**)[9] is shown below. The consensus sequence of BARHL2 is AACCAATTAA. The score for an individual nucleotide at each position is the ratio of the counts of such nucleotide to the total counts of all nucleotides. For example, the score for A nucleotide at the first position is 78/(78+4+13+5)=0.78. Score for consensus sequence is the sum of the highest score at each position (0.78+0.9+0.44+0.94+0.95+0.96+1.00+1.00+0.99+0.73=8.69), while the average score for a random sequence is 0.25*10=2.5. Accordingly, any given sequence with the score higher than 2.5 has higher affinity to BARHL2 than random.

Table 1.2 PWM model for BARHL2

A	78	91	19	0	95	96	0	0	99	73
C	4	2	44	94	0	1	0	0	0	2
G	13	7	17	0	4	0	0	0	0	14
T	5	1	20	6	1	3	99	99	1	11

Besides the basic PWM model which is based on the assumption of independent contribution of each position to TF binding, other advanced models were also developed, such as binding energy model[20], transcription factor flexible model[21], and connecting matrix model[18], that takes into account the factors affecting TF-DNA interaction e.g. dinucleotide interdependency, dimer formation, etc.

The TF-DNA binding model could be generated with different methods (**Figure 1.1**). In early stage, electrophoretic mobility shift assay (EMSA) was carried out to determine the affinity of TF to different DNA sequences[22,23]. The mobility of large molecules in gel is determined by its size and charge. Based on this, if the tested DNA oligo is bound by the TF, the DNA-TF complex will move slower in gel than the DNA molecule alone. To confirm the TF identity, an antibody against the TF can also be added and the larger complex will move even slower. By comparing the mobility of TF-incubated DNA with the non-bound control DNA, one can tell whether such DNA sequence is preferred by the TF or not. However, the number of tested oligos is relatively low so that it takes lots of efforts to generate a PWM model though still with low quantitative accuracy.

Protein binding microarray (PBM) has later on been widely applied to model a large number of TF DNA binding specificities[24,25]. PBM first generates the double-strand DNA using the microarray probes as templates and labels the DNA with Cy3-dUTP before TF proteins are applied to the DNA array for sequence selection. To detect the TF binding, antibody with conjugated fluorescence is added. By reading the fluorescence of both double-strand DNA and protein antibody, one can tell which sequences are preferred by the tested TF. PBM uses all combinations of 8-mer sequences for selection and yields a much higher number of counts than EMSA to generate the PWM model. However, the probe length of microarray (8 bp) inherently limits the analysis of the longer binding sites, which is very common for proteins that form dimer when binding to DNA[18].

With the development of high throughput sequencing technology, Systematic Evolution of Ligands by Exponential Enrichment (SELEX) incorporated with the massively in parallel sequencing, emerged to be very efficient in both throughput and accuracy to model the TF binding specificities[26,27]. Random oligos are synthesized and double-stranded. They are incubated and selected with immobilized TF protein. The selected DNA is amplified and sequenced to obtain the preferred sequence of the tested TF. In general, the selected DNA will be amplified by PCR and used for the next cycle of selection to reduce the noise. SELEX can be applied to analyze the TF binding sites without any length limit; albeit it also has some disadvantage as the competition and PCR based method may lose the very low affinity sites or become saturated for the very high affinity sites.

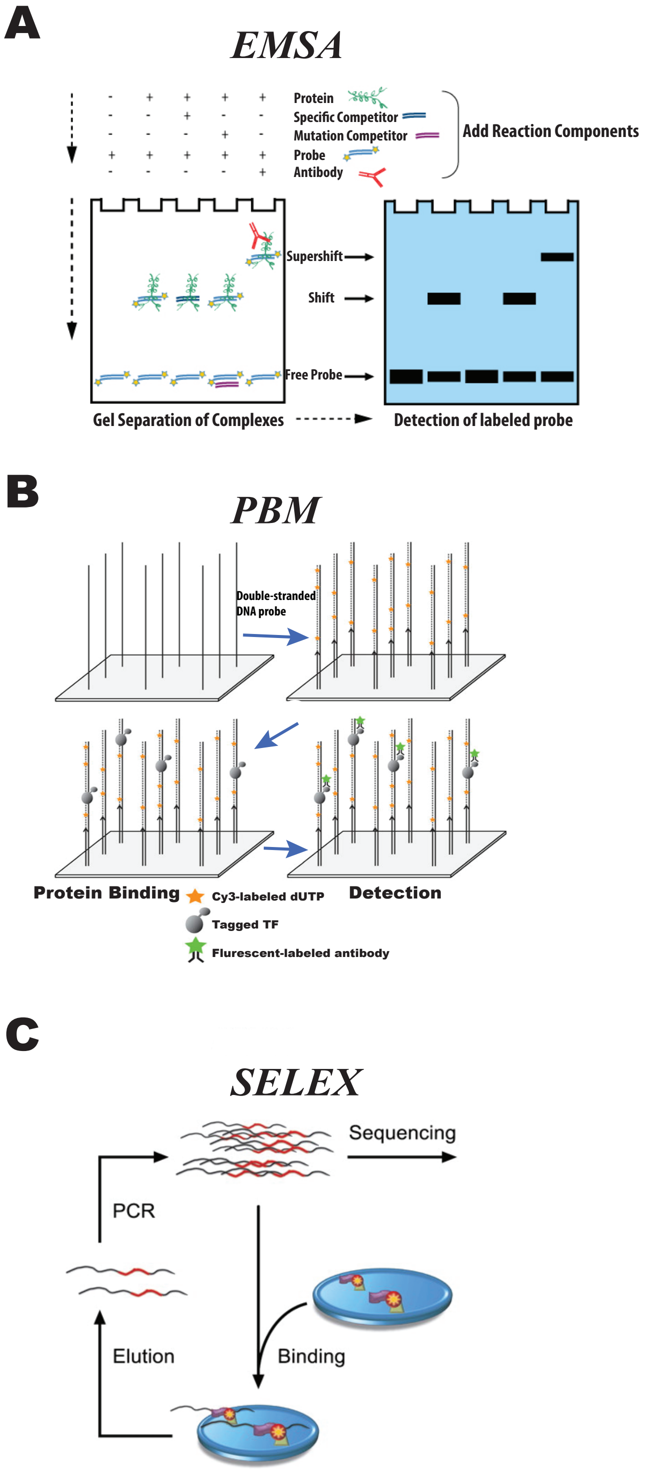

Figure 1.1 Different methods to study DNA binding specificities of TFs

(A) Electrophoretic mobility shift assay (EMSA). Different oligos could be tested for the binding affinity with proteins. After bound by TF protein, the mobility of DNA oligo in the gel electrophoresis will be affected. In addition, competitor oligos are used to test the affinity of the binding and antibody against the TF could be applied to double confirm the binding of TF. By comparing the shift of the probes in the gel, affinity could be concluded.

(B) Protein binding microarray (PBM). Microarray probes are used as templates to synthesize the double-strand DNA and labeled with Cy3-dUTP. TF protein is then applied to select the preferred sequences before being detected with specific antibody that is conjugated with fluorescent dye. After washing, fluorescence for both the double strand DNA and antibody will be read. The detected DNA sequences will be used to build the binding motifs.

(C) SELEX. TF protein is immobilized and incubated with different DNA oligos. After washing, elution and amplification, the DNA will be sequenced in order to determine the binding affinity of the TF. Normally several cycles of enrichment are necessary to decrease the noise.

Figures are adapted from http://www.piercenet.com/method/gel-shift-assays-emsa and [26,30]

There are also other ways to investigate the TF binding models, such as the mechanically induced trapping of molecular interaction (MITOMI)[28], bacteria-one-hybrid[29], ChIP-seq[27], etc. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) uses antibody to specifically pull down the TF which is cross-linked to chromatin DNA with formaldehyde. After non-specific binding being washed away with different buffers, the bound DNA could be de-crosslinked and sequenced. The sequence could be used to map the binding sites of the TF in the genome and build the binding motifs from the significant binding sites. This technique has been widely applied to query the *in vivo* binding sites of TFs in different cell types. However based on the complex context of the genome, the binding specificities of only a small number of TFs could be modeled with ChIP-seq.

1.1.2 Transcription *cis*-regulatory element

In the mammalian genome, the coding region only takes up little more than 1% of the total sequence. The function of the vast majority of the genome is yet to be elucidated. ENCODE project, started in 2007, has performed different sequence-based methods to annotate the human genome and claimed strikingly that 80% of the genome participates in RNA and/or chromatin associated biochemical events in at least one cell type they studied[31,32]. The broadest elements they covered are different types of transcribable genome, taking up to over 60% to 70% of the analyzed genome[33]. However, the knowledge to understand the function of all these RNAs is still lacking. A majority of these transcribed sequences are not conserved; hence it is still questionable whether they are really functional or biological meaningful[34,35].

A fraction of the non-coding sequence that includes information to control when, where and how the gene is expressed, is called regulatory elements. Such information is mainly embedded in different types of function elements: promoters, enhancers, silencers and insulators/boundary elements[36].

Promoter is a position and orientation dependent *cis*-regulatory element, located immediate upstream of its target gene with the length of 100 to 1000 base pairs (bp) and plays an essential role in the initiation of the transcription. Promoter contains the DNA sequence that is recognized by sequence specific TFs, general TFs and concomitantly RNA polymerase. In prokaryotic genome, promoter often contains binding sites for RNA polymerase and associated protein called σ factor to form the holoenzyme and initiate the transcription[37]. There are two conserved consensus sequences in the bacteria promoter region, which are located at approximately 10 bp and 35 bp upstream of the transcription start site (TSS), called -10 element (consensus sequence as "TATTA") and -35 element (consensus sequence as "TTGACA") respectively. Besides, some bacterial promoters even contain other conserved upstream promoter elements (UP elements) for recognition of RNA holoenzymes with different σ factors[38].

The promoter in eukaryotic genome is much more complex, mostly composed of three classes: the core promoter, proximal promoter and distal promoter. The core promoter is the minimal essential composition for transcription initiation. It should contain the TSS, RNA polymerase binding site, general TF binding sites, and spaces between them[39]. Since there are three different types of RNA polymerases in human genome to transcribe different types of RNAs that require different general TFs, the composition of core promoter differs for different RNA polymerases[40].

The TATA box (also called Goldberg-Hogness box) is an important consensus element (consensus sequence “TATAAA”) in human core promoter, located at 25 bp upstream of TSS and found in ca. 20% of human genes[41]. TATA box is a well-studied element in the core promoter to recruit the TATA box binding protein (TBP), the general TFs and RNA polymerase II. However, the prevalence of TATA-less gene is over 80% and various alternative core promoters are identified, i.a. X core promoter element 2 for the second TSS of hepatitis B virus X gene mRNA transcription[42]. The proximal and distal promoters are similar in structure and function, both containing a cluster of sequence specific TF binding sites, differing from each other with respect to their distance from the TSS. The proximal promoter resides around 250 bp upstream of TSS and the distal promoter is accommodated further upstream in the same strand as the TSS, and their regulatory activity is location- and orientation-dependent.

Enhancer is a major type of the regulatory element which contains a cluster of sequence specific TF binding sites. The regulatory activity of enhancer is position and orientation independent[43]. The main function of enhancer is to enhance the transcriptional activity, hence its name. A single gene could be regulated by several enhancers. And enhancer could locate either upstream or downstream of the target gene, and some are even over hundreds of kilobases (kb) away from the gene. It is estimated that in each cell type there are over thousands of active enhancers, and in total approximately 1 million enhancers are embedded in the human genome[31,44,45].

The enhancers cooperate to control the tissue specific expression of the genes. The regulatory activity of enhancer is more dynamic and varies in different development stages and amongst different cell types[46]. Evidence has shown that in a given cell type some TFs could serve as the master regulators whose occupancy at the enhancer could determine the fate of other TFs and is predictive of the enhancer activity[15,47,48]. Because of the importance of the enhancer function during embryonic development, the sequence of enhancer is positively selected by evolution and highly conserved across mammals[49]. The mutations and variants in the enhancer sequence have a higher potential to affect the gene expression than those located outside enhancers[50]. Indeed, genome wide association studies (GWAS) confirm that the disease-associated variants are more frequently found within the enhancers across the genome[51,52]. A more recent study verified that genetic variation affecting the binding site of lineage-determining TF would affect the enhancer activity and function[53].

Active enhancer is commonly hypersensitive to DNase I and associated with specific histone modifications, such as monomethylation of histone H3 lysine 4 or acetylation of histone H3 lysine 27[54,55], and histone variants such as H2A.Z and H3.3[56], which will be discussed in the next section. It has also been reported that active enhancer could be transcribed to RNA, called eRNA, which might contribute to the enhancer activity[57-59]. However, the mechanism remains unclear although rising evidence suggests its function in recruitment of regulatory protein[59].

Since enhancer could be over hundreds of kilobases (kb) off the target gene, the physical interaction between the enhancer and its target promoter or gene body is necessary for its regulatory function. Some structural proteins like mediator or cohesin could mediate the interaction[60-62]. A subclass of long noncoding RNA, termed noncoding RNA activating (ncRNA-a) has also been found to play a role in mediating the chromatin interaction[57,63]. To detect the physical interaction between different regions of chromatin, the chromosome conformation capture (3C) based methods[64-69] have been developed. The 3C method is based on the model that the physically closer chromatin regions would have higher probability of interaction. It uses restriction endonuclease to cut open the chromatin and relegate the ends. Two physically adjacent chromatin ends are supposed to be ligated more frequently than the distant ends. So one can test the ligation frequency by quantitative PCR with two primers next to two given ends respectively. Such ligation frequency reflects the relative physical proximity, thus the interaction probability. Based on 3C, more advanced technologies 4C, 5C, 6C, Hi-C and ChIA-PET have also been developed[64-69] to detect the interactions between multiple regions, or even genome-wide. More recently, a method based on RNA fluorescent in situ hybridization has been developed to detect the chromosome interaction[70].

Insulator (also called boundary element) is a class of regulatory element which protects the active gene from the heterochromatin or blocks the cross talk between enhancer(s) and promoter. The insulating function is thought to be associated with the control of histone modification or DNA methylation and/or be accomplished by affecting the chromatin looping by a C2H2 zinc finger protein CCCTC-binding factor (CTCF)[71-73].

Silencer is a less well-known type of regulatory element that could block the promoter activity of a gene. There are two types of silencers: type I is the classical, position or orientation independent ‘silencer’; type II is the non-classical, position or orientation dependent ‘negative regulatory element’ (NRE)[74]. The classical silencer uses an active mechanism and does not need to bind to any protein but directly interferes the general TF assembly[75,76]. The non-classical NRE is a type of passive silencer, which recruits a transcription repressor to disrupt the transcription machinery[77-79].

Super enhancer has recently been categorized as a new type of regulatory elements that regulates the expression of the key genes to control the cell state. Super enhancer is a cluster of enhancers, spanning over tens of kb at the genes, and densely occupied by master TFs and mediator[48]. In addition, the disease-associated variants are highly enriched in the super enhancers of the disease relevant cells. Interestingly, in cancer cells, super enhancer can be acquired for the key oncogene and provides a biomarker for tumor specific pathologies, which may be valuable for cancer therapies[80].

1.1.3 Epigenetic factors and cohesin in transcriptional regulation

1.1.3.1 Over view of epigenetic factors in transcriptional regulation

In most cases, knowing the DNA sequence specificities that a TF prefers is not sufficient to explain all its binding sites in the genome, which suggests that TF binding is context dependent, e.g. protein-protein interaction, chromatin open accessibility or DNA methylation may affect the TF binding. It is very difficult to predict the transcription level of a gene merely based on DNA sequence itself. There are more and more epigenetic factors being discovered to regulate the transcription in our cells.

There are over 200 different cell types in our human body, all of which contain almost the identical copy of the genome. However, the transcriptome in each individual cell type is unique[81]. The information of cell identity must not be directly reflected from the genomic sequence and arising evidence points out that such information could be coded to the epigenome of cell. With the rapid progress of high throughput methods and sequencing technology, we are able to access the epigenome data and associate them with the gene expression and cell identity.

There are three different types of epigenomic information: 1) DNA methylation, 2) histone covalent modifications and 3) histone variants. Covalent histone modification could be either active or repressive marks for the neighbouring regulatory element. Histone acetylation is a general active mark for the open chromatin, which takes place on the lysine residue within the N-termini of histone tail on the nucleosome core surface. Earlier, people thought that adding the acetyl group to the histone would bring negative charge and neutralize the positive charge of the protein surface, which will decrease the histone electrostatic interaction with DNA backbone and loosen the DNA from nucleosome. Nevertheless, this hypothesis has been challenged by recent studies, according to which the turnover rate of histone acetylation is more rapid at the more actively transcribed gene[82,83]. If acetylation only functions on neutralizing the positive charge of histone surface, rapid turnover rate will however increase the DNA interaction with these histones than the constantly acetylated histones.

1.1.3.2 DNA methylation

In all cell types in human, except germline cells and pre-implantation embryos, the genomic DNA methylation pattern is relatively stable and unique. DNA methylation is a principal epigenetic mark for gene expression, imprinting and X chromosome inactivation, and essential for stem cell pluripotency and development. About 70-80% of CpGs in mammalian genome are methylated. Most of the methylated CpGs are constant among cell types whereas less than 10% of genome region is variable in methylation. There are some regions in the genome, ranging about 200 to 500 bp in length, with over 50% of the CG dinucleotide content and a ratio of observed to expected CpG >0.6, called CpG islands (CGI)[84]. Methylation of CGIs in the promoter is believed to correlate with the silencing of the gene, although recent studies suggested that only a subset of the genes with methylated CGIs were defective of transcription[85].

The tissue-specific differentially methylated regions (tDMRs) occur in distal *cis*-regulatory elements and are enriched in disease-related single nucleotide polymorphisms (SNPs). Also the DMRs mark some enhancers that are active during embryo development but become silent or ‘vestigial’ in adult cells[86,87]. With some recent study in addition to CG sequence, the CHG and CHH (H for A, C and T) methylations, which were thought common in plant genome, have also been discovered in some mammalian tissue[88] although their role remains unclear.

1.1.3.3 Histone modifications

Two enzymes are involved in the dynamic acetylation and deacetylation of histone tail: Histone acetyltransferase (HAT) and Histone Deacetylase (HDAC). The genome-wide mapping of HAT and HDAC binding sites showed that both enzymes bind to the transcribed sites, recruited by phosphorylated RNA Polymerase II[89]. HADC could reset the chromatin state by removal of the acetyl group from histone tail and mutations of HDACs are linked to tumor development as they may cause aberrant expression of key genes for cell cycle or apoptosis[90]. Two homologous HAT family members, p300 and CREB binding protein (CBP), have been characterized as transcription co-activators. Their binding has been lately shown to highly correlate to the tissue specific enhancer activity[91]. One of their products, histone H3 lysine 27 acetylation (H3K27ac), is now widely recognized as an active enhancer marker[92-94].

Histone acetylation and deacetylation are also related to the long-lasting transcription-dependent memory formation. Individuals with increments of histone acetylation are better at learning and memory, whereas lack of it could cause cognitive impairment[95].

Besides histone acetylation, histone methylation has also consistently revealed strong correlation with transcription activity. Histone lysine methylation is more stable than acetylation and could act as either an active or repressive mark. The best-known active histone methylation marks are histone H3 lysine 4 with mono-, or trimethylation though their distribution in the genome is different. The monomethyllysine 4 of histone H3 (H3K4me1) had been known as an enhancer hallmark even before H3K27ac. In spite of their similar function in marking active enhancers, H3K4me1 is less dynamic than H3K27ac during transient stimulation and embryonic development, and hereby H3K4me1 marks both the active enhancers and vestigial enhancers that were active at earlier development stages[96-98]. The enzyme that deposits this mark has not been disclosed yet, leaving the question whether it comes from the *de novo* methylation of a native lysine or demethylation from di- or trimethylation of H3K4.

Histone H3 lysine 4 trimethylation (H3K4me3) is the most well studied chromatin hallmark, which records the active promoter of protein coding gene[54]. In yeast, the writer of this modification is only a single methyltransferase called Set1[99], while in human there are several proteins or protein complexes to accomplish the task, including SET1A, SET1B and MLL1-4[100-102].

The most well-known repressive histone modifications are histone H3 lysine 9 di- and tri-methylation (H3K9me2 and H3K9me3, respectively), and histone H3 lysine 27 trimethylation (H3K27me3). They are all enriched in the promoters of low and silent genes. Notably, H3K9me2 and H3K9me3, deposited by histone methyltransferase SETDB1, are well-characterized heterochromatin markers, which recruit heterochromatin protein HP1 and in turn other repressive chromatin modification enzymes[103-105].

All nucleosomal histone tails could be phosphorylated or dephosphrylated by kinases or phosphatases. This type of modification includes the target residues of serine, threonine and tyrosine. It is best known to take place during chromatin remodeling in response to DNA damage repair. More recent evidence points out that it can be also involved in other nuclear processes, such as transcription and DNA compaction in cell cycle and apoptosis[106-108]. In response to the double strand breakage, serines in histone H2A (e.g. H2AS129) will be phosphorylated by protein kinases Tel1 and Mec1[109,110]. Subsequently, the phosphorylated histone could be either replaced or dephosphorylated by phosphotases PP2A, PP4, PP6[111-113]. In addition to the DNA damage response, phosphorylation of serine 10, threonine 11 and serine 28 of histone H3 are found related to the activation of transcription[114-116].

1.1.3.4 Histone variants related to transcriptional regulation

The histone variants H3.3 and H2A.Z contained nucleosome core particles (NCRs) are less stable than the canonical NCR. It has been suggested that they may be involved in histone exchange or nucleosome eviction during chromatin remodeling[117]. It is even more regularly observed that these histone variants marked the nucleosome free regions or protein accessible regions of the genome, which are potentially the regulatory element regions, e.g. promoters, enhancers or insulators as we discussed above[56,118,119].

1.1.3.5 Cross talk between different epigenetic factors

The large number of chromatin modification in the cell provides the scope of precise regulation (Listed in **Table 1.3**). Due to the cross talk between all these modifications, a higher level of complexity has been added to the system to fine-tune the overall control. As far as is known, at least but not limited to four different types of cross talks exist between two different marks: i) two modifications compete for the same substrate as antagonists: the best example for this is H3K27, of which the acetylation is an active mark and trimethylation is a repressive mark; ii) one modification depends on another: histone H3K9me2 could help to recruit the chromomethylase to the nucleosome and direct local DNA methylation[120]; iii) the binding of a protein to one modification is disrupted by another adjacent modification: HP1 and H3K9me2/3 binding is inhibited if histone H3 serine 10 is phosphorylated[121]; iv) two or multiple modifications cooperate to recruit proteins to the chromatin: PHF8 binds to H3K4me3 stronger when H3K9 and H3K14 are phosphorylated[122].

Table 1.3 Examples of Histone Code (adapted from[54,123-127]; K=lysine, S=serine)

Histone	*Amino Acid*	*Modification*	*Associated Transcription Activity*
H2B	K5	Monomethylation	Transcription activation
		Trimethylation	Transcription repression
H3	K4	Mono-, Di- Tri-methylation	Transcription activation
	K9	Di-, Tri-methylation	Transcription repression
	S10	Phosphorylation	Transcription activation
	K27	Acetylation	Transcription activation
		Trimethylation	Transcription repression
	K36	Di-, Tri-methylation	Transcription elongation
		Acetylation	Transcription activation
	K79	Mono-, Di-, Tri-methylation	Transcription activation
H4	K5	Acetylation	Transcription activation
	K8	Acetylation	Transcription activation
	K12	Acetylation	Transcription activation
	K20	Monomethylation	Transcription activation

1.1.3.6 Cohesin and its function in transcriptional regulation

Unlike the epigenomic factors discussed above, there are some heritable epigenetic features that are not the covalent modification of chromatin but their function is related to gene expression or cellular phenotype, e.g. the chromatin accessibility for protein binding, the gene expression pattern and the chromosome topological domain territories.

For epigenetic factors to function on the regulation of transcription, a physical interaction between the regulatory element and its target gene must be formed. Among others, cohesin and mediator are demonstrated to connect the gene expression and chromatin architecture[60].

Cohesin is a ring-shape protein complex, composed of four subunits, SMC1, SMC3, SCC1/RAD21 and SCC3/SA (**Figure 1.2**). SMC1 and SMC3 belong to the structural maintenance of chromosome (SMC) family, which is a chromosomal ATPase and conserved from yeast to human. Upon its discovery, the main function of cohesin was identified as the establishment of cohesion between two sister chromatids during mitosis[128]. Other family members include SMC2 and SMC4 of condensin complex, and SMC5 and SMC6 that function in DNA damage repair.

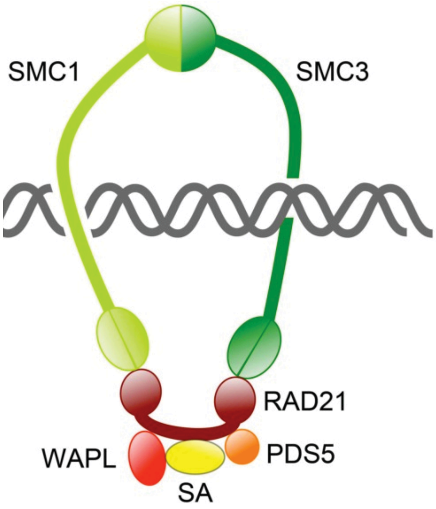

Figure 1.2 Structure of cohesin complex structure

Cohesin is a ring shape protein complex composed of four subunits, SMC1 and SMC3, SCC1/RAD21, SCC3/SA. Cohesin could not directly bind to the chomatin but be loaded by its loading factor NIPBL. During mitosis, cohesin ring will be released from the chromatin by WAPL and PDS5. Figure is adapted from [129].

RAD21/SCC1 belongs to another protein family kleisin and interacts with both SMC proteins to close the ring and encircle the DNA strands. Kleisin is a superfamily consisting of multiple members such as RAD21, REC8, CAP-H, and its function is exclusively to form the protein ring complexes with SMC proteins (reviewed in [130]). Kleisin is also playing a role of guarding the protein ring. During cell cycle, cohesin ring complex is formed in G1 phase and loaded by cohesin loading factors NIPBL and MAU2 in an ATP-dependent manner. The binding of cohesin on chromatin is dynamic, and the release is promoted by two known proteins associated with RAD21 called PDS5 and WAPL[131,132]. When cell enters S phase, the cohesin subunit SMC3 is acetylated by cohesin acetyltransferases, triggering a confirmational change and neutralizing the PDS5-WAPL unloading action. Hence it could stabilize cohesin to entrap both sister chromatids until their segregation later in mitosis[133,134]. At the onset of prophase during mitosis, most cohesin gets phosphorylated and starts to dissociate off chromatin in a process driven by PDS5 and WAPL. However, the pericentromeric cohesin could stay bound to chromatin and oppose the pulling force from spindle and allow the chromosome to align in metaphase before two daughter cells separate. Stepping into anaphase, the cell will unveil the degradation of securin, so that separase could cleave RAD21 and release the sister chromatids to two different poles. After segregation, cohesin will be deacetylated by HDAC8 and re-associated with PDS5-WAPL which would stabilize and protect it from being cleaved by separase and enables the exit from mitosis[131,135,136].

In human somatic cells, there are two alternative SCC3 proteins SA1 and SA2, with mutually exclusively binding to RAD21 depending on different cell types. The function of SCC3 is less known than the other subunits and has been suggested to bridge two cohesin rings together in a handcuff shape with one chromosome in each ring[137].

Cohesin has lately been proven to be involved in organizing the chromosome topological domains and 3D interaction[62,138-140]. The chromatin looping provides a preferable and more efficient environment for DNA replication and RNA transcription. For replication, looping would allow multiple loci to be replicated at the same time. For transcription, looping allows the interaction of enhancers to regulate the transcription machinery. Cohesin is enriched in replication origins and could interact with pre-replication complex (e.g. MCM) proteins; nonetheless MCM and cohesin loading is independent, and not affected by each other. Upon the depletion of cohesin, S-phase progression is expected to slow down with reduced number of active origins[141]. However, how cohesin behaves during the progress of DNA replication has not been elucidated: does cohesin move together with the replication fork or remain bound to the orgins during the whole S phase? If latter, what is the topology when cohesin and replication fork clash?

The chromatin loops organized by cohesin could also serve as the physical interaction between promoters and distal regulatory elements. Cohesin can be loaded to chromatin via the insulator protein CTCF and may be involved in organizing long-range topological domain territories[142]. Cohesin can also be loaded by NIPBL and mediator complex, and cooperate to connect the transcription regulation between enhancer and its target promoter[60,138].

Mouse embryo with complete loss of one subunit (SCC3) of cohesin experiences the severe developmental defect and die from E12.5 before birth[143]. There are two human syndromes related to dysfunction of cohesin: Cornelia de Lange (CdLS) and Roberts Syndromes (RBS). CdLS affects 1 in 30,000 children and shows both physical and mental developmental anomalis. Over half of the CdLS patients carried heterozygous mutations in cohesin loading factor NIPBL but very rare in cohesin subunits themselves[136]. RBS is very rare syndrome, characterized as prenatal growth retardation. The main homozygous mutation found in the patients is ESCO2 which encodes a cohesin acetyltransferase[144]. Mutations in genes coding different subunits of cohesin are found in a number of human cancers with aneuploidy, as well as acute myeloid leukemia, which is not characterized as aneuploidy[145,146].

1.2 COLORECTAL CANCER (CRC)

1.2.1 Overview

Cancer has become the major risk to threaten human health and caused 7.6 million deaths worldwide in 2008, being the leading death factor of all diseases. Over 11 million new cases of cancer were diagnosed in 2007 and the number has been continuously increasing. The World Health Organization (WHO) estimated that the global cancer death toll would increase 45% (from 7.9 million to 11.5 million deaths) from 2007 to 2030 (Statistics from WHO official website: http://www.who.int/cancer/en/). Each year February 4th is set as the World Cancer Day to promote ways to ease the global burden of cancer. What is cancer?

There are over 200 different types of human cells, rising from the zygote through cell proliferation, cell death and differentiation. All these processes are under tight control to make sure the body maintains the particular number of individual type of cells at different developmental stages. When the growth control is lost, the cell will develop to neoplasm or tumor. Most of the tumors are benign, and therefore are harmless for human health, although sometimes the benign tumor cells secret the aberrant level of hormone or grow so big as to push the adjacent tissues with heavy pressure. The benign tumor is self-limited and has a clear margin with the surrounding normal tissue. In contrast, the malignant tumor, which is also called cancer, could invade and spread to other organs, termed metastasis. Metastasis is the main cause of death for cancer patients.

Cancer has been characterized with seven different features: invasion and metastasis, self-sufficiency of growth signal, insensitive to antigrowth signal, escape of apoptosis, unlimited proliferative potential, sustained angiogenesis[147] and inflammatory microenvironment[148]. Cancer can rise from different cell type origins: the most frequently diagnosed cancer type is carcinoma which is derived from epithelial cells; the mesenchymal tissue could develop to sarcoma and leukemia originates from blood cells.

The prominent factors contributing to cancer development are environment and lifestyle, which include smoking, frequent exposure to radiation, high alcohol consumption, lack of exercise, poor diet, obesity, frequent contact with carcinogens (e.g. benzene and asbestos), bacteria or virus infection[149-151]. Cancer could also result from genetic factors: mutations in oncogene or tumor repressor genes are frequently detected in patients' genome, and hereditary variants in the high-penetrant cancer associated genes or the combination of several low-penetrant variants confer high risks of cancer onset[152].

Colorectal cancer is the third most common cancer type in developed countries for both genders. It has very high mortality rate, in average about 5 to10 per 100,000 of population (data extracted from GLOBOCAN, http://globocan.iarc.fr/). However, ninety percent of CRC cases showed little or low genetic factor, and if diagnosed early and treated adequately, most CRC can be cured. So it is so important to determine the risk factors of CRC at different levels, from environmental factors to molecular mechanisms, so as to ease and facilitate the early diagnosis and therapies.

About 15% of CRC patients show the microsatellite instability (MSI) caused by dysfunction of DNA mismatch repair (DMR) pathway. There are two main subtypes of the MSI colorectal cancers: 20% of the MSI cases belong to the well known hereditary non-polyposis colorectal cancer (HNPCC, also called Lynch syndrome) and the other 80% MSI arise from sporadic mutation[153]. The Lynch syndrome is developed from conventional adenomas, in which familial germline mutations or insertions or deletions (indels) of Adenomatous polyposis coli (APC), β-catenin, K-Ras and some genes involved in DMR such as MLH1, MSH2, MSH6 and PMS2 are frequently detected. In contrast, sporadic MSI colon cancer cases are commonly found either with hypermethylated promoter of MLH1 gene, coding for a protein involved in DMR or with mutations in B-Raf, a serine/threonine-protein kinase that could control the cell growth[154,155]. Compared with microsatellite stable CRC, the MSI CRC has better prognosis with significant clinical features, lymphocytic infiltration, mucin secretion and poor differentiation[155]. However, the response to chemotherapy is unfortunately worse than the microsatellite stable CRC[154]. To characterize both CRCs in molecular level could provide important information for clinical diagnosis as well as for discovering targets for CRC therapy.

1.2.2 Transcriptional regulation in development and CRC

TFs are also found important in most biological processes. The Nobel Prize of medicine and physiology in 2012 was awarded to Dr. Shinya Yamanaka for his finding that introducing four TFs, Oct3/4, Sox2, c-Myc and Klf4, into mouse fibroblast cells could reprogram the cells to pluripotent cells[156]. In addition, mutations in some TFs are frequently identified in severe diseases such as diabetes and cancers. For example, mutation of the genes encoding for TFs p53 and c-Myc are commonly found in many cancer types[157,158]; Peroxisome proliferator-activated receptor gamma (PPARγ) is involved in the pathophysiology of metabolic syndromes, like diabetes and obesity[159].

CRC arises from colonic epithelium, forming invaginations into the lamina propria mucosae, known as crypt which is the functional unit of colon[160]. Intestinal self-renewal is one of the fastest in all human tissues, and the complete replacement of all the cells only takes about five days[161]. All cell types present in the intestinal epithelium are derived from the intestinal stem cell which sits at the base of the crypt. The stem cell maintains and proliferates in the stem cell niche supported by pericryptal myofibroblast from a mesenchymal lineage, migrates upwards through the crypt and differentiates to other cell types. There are four principal types of epithelial cells in colon: the absorptive enterocytes, the mucus-secreting goblet cells, the endocrine cells and paneth cells (**Figure 1.3**).

After a study showed that the crypt was a clonal population originating from an intestinal stem cell[162,163], it brought great interest to study the dynamics of intestinal stem cell differentiation. Several signaling pathways are involved in the stem cell maintenance and differentiation, among which Wnt-β-catenin signaling pathway is thought to be of central importance. Wnt signaling pathway is highly conserved from *Drosophila* to vertebrates and plays an important role in embryonic development and body patterning[164]. It can regulate cell growth, differentiation and apoptosis, and especially control the balance between proliferation and cell lineage commitment (**Figure 1.4**).

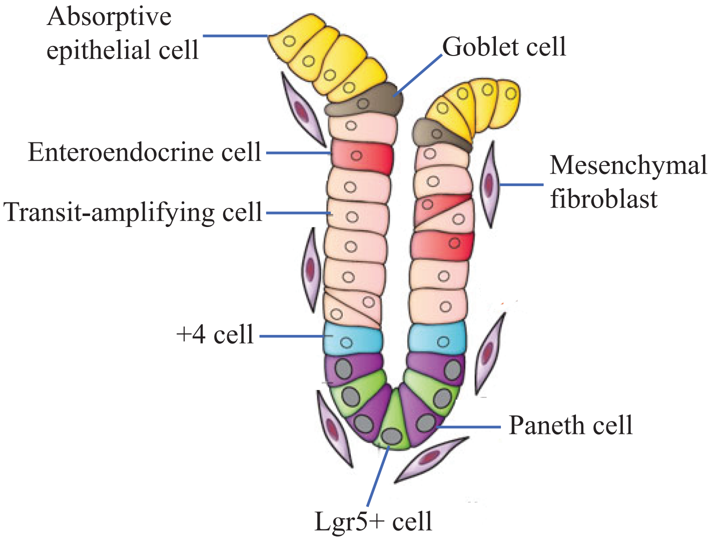

Figure 1.3 The different cell types in intestinal crypt

The stem cell is located at the bottom of the crypt, interspersed between paneth cells. The proliferated cells migrate up and differentiated to different cell types under signaling control in their niches. Figure is modified from [1].

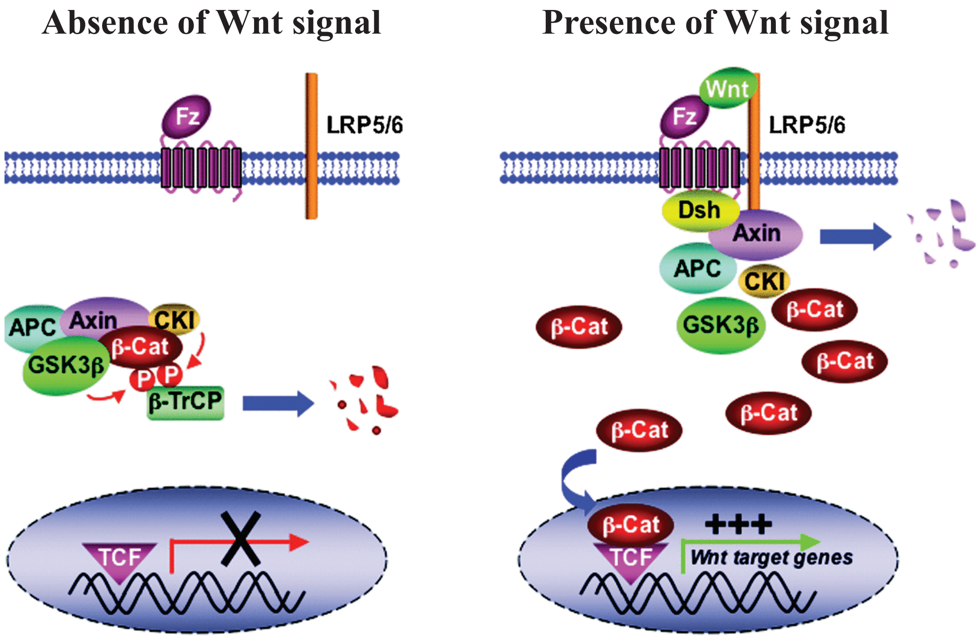

Figure 1.4 Silencing and activation of Wnt signaling pathway

Left: when Wnt signal is absent, β-catenin is bound and phosphorylated by other proteins, and then subjected to degradation in cytosol. Right: when Wnt signal is recognized by Frizzled receptor, it recruits Axin and APC to the cell membrane and releases β-catenin. The free protein will be translocated into nucleus and co-binds to chromatin with TCF transcription factor to drive their target gene expression. Figure is modified from [165].

Frizzled receptor on the surface of epithelial stem cell receives the Wnt-signaling ligand secreted from pericryptal myofibroblast cells. Upon activation by Wnt ligand, the cell would allow the translocation of β-catenin from cytosol to nuclei and interact with transcription factor TCF or LEF, to drive the expression of downstream target genes. Targeted conditional deletion of one TCF transcription factor Tcf4 or inhibition of Wnt pathway in mouse disrupts the stem cell population and resulted in lack of cell proliferation and stem cell compartment in the mouse[166-168]. The most important target of Wnt pathway in CRC is Myc proto-oncogene, which will be discussed later in **Section 1.2.3**.

In addition, *APC* gene has been identified as being responsible for familial adenomatous polyposis (FAP)[169,170]. *APC* is located in chromosome 5 in human, and somatic mutations in APC are also detected in sporadic CRC patients. *APC* encodes a protein with 2543 amino acids (aa) in length, and can interact with β-catenin to regulate its concentration in cytoplasm[171]. Disruption of APC-β-catenin interaction will cause the elevated concentration of β-catenin in the cytosol and foster its translocation into nucleus. The nuclear β-catenin could bind to TCF-4/LEF-1 transcription factor and co-activates the target genes, including MYC oncogene.

The mutation of the small G protein RAS has also been widely described in CRC patients. RAS is involved in two main cellular pathways, mitogen-activated protein kinases (MAPK) pathway and phosphoinositide-3 kinase (PI3K) pathway. After ligand binding, RAS will be recruited to the cell surface by the receptor and activated by guanine nucleotide exchange factor son of sevenless (SOS). SOS replaces the RAS bound GDP to GTP. RAS-GTP in turn recruits RAF to the cell surface and activates BRAF. Activated BRAF will then phosphorylated ERK which will be translocated to nuclei to phosphorylated the transcription factor Jun and Fos, and hence increases the transcription of targeted genes. Mutations in KRAS genes are very common in CRC, taking up to 40% of the cases. In general, the mutant KRAS is maintained constitutively active in cells amplifying the MAPK pathway signals[172], leading to the growth of adenomas.

The transforming growth factor beta TGF-β pathway is also commonly mutated in CRC patient cells. In normal cells, when ligand binds to the receptor in the cell surface, R-SMAD will be phosphorylated and interact with Co-SMAD. The protein complex will together be translocated to nuclei and drive the downstream targets. That the mutations and deletions of the members in the TGF-β pathway would cause unlimited proliferation indicates the tumor repressor function of TGF-β pathway. Besides, another TGF-β pathway factor called BMP, also contributes to the maintenance of stem cell niche through its gradient expression level along the crypt, with the highest level at the top of the colonic crypt[173]. It is demonstrated that BMP could antagonize Wnt signaling pathway via Pten tumor repressor protein and restrict cell proliferation[174].

Vogelstein et al. (1990) suggested that CRC is a result of multiple somatic mutations happening in different stages of the tumorigenesis[176] (**Figure 1.5**). In the first stage, APC mutation is commonly discovered to lead to the initial formation of polyps in the intestinal epithelium. The tumor growing from small polyps to large adenomas is commonly associated with mutations of KRAS and BRAF, amplifying the MAPK pathway gene expression. The final carcinoma transformation would require mutations in the tumor repressor gene p53 or transforming growth factor TGF-β.

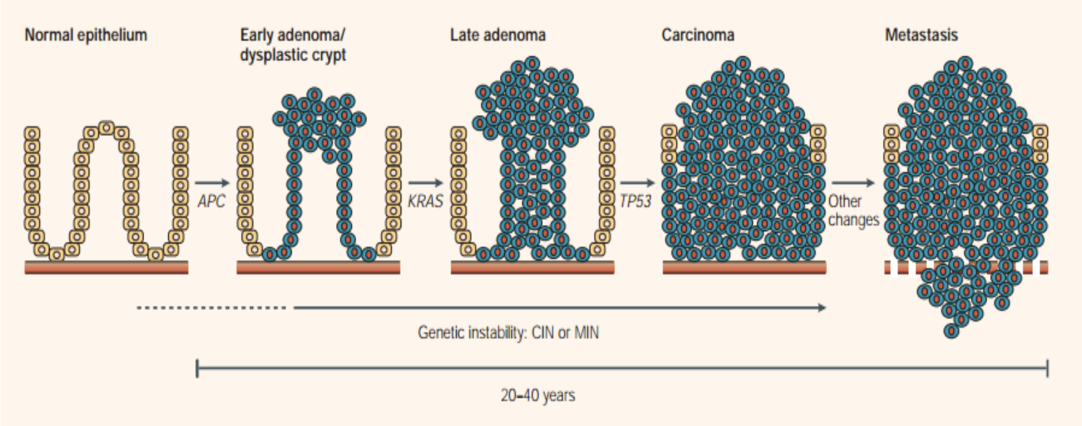

Figure 1.5 Sequential Mutation Model of CRC (adapted from [175])
When genome becomes instable, the mutations start to happen which increases the risk of CRC. In early stage, when APC mutations are acquired, small adenomas initiate to form in intestines. If additional mutations in KRAS happen, the small adenomas grow to large adenomas. Under such circumstances, if the tumor repressor gene TP53 obtains mutations the adenomas will transform to carcinomas and metastasis. The total tumorigenesis generally takes 20 to 40 years.

Genome-wide Association Studies (GWAS) from different populations have identified several SNPs affecting a few genes and linked to human CRCs (**Table 1.4**)[177-182], among which one of the most well studied SNPs rs6983267 at 8q24 was located 335 kb upstream of MYC gene in a gene desert region. Following study revealed that the region physically interacted with MYC promoter, and the risk 'G' allele of the SNP conferred stronger binding of Wnt-signaling pathway transcription factor TCF7L2 and β-catenin and could therefore potentially affect the MYC transcription[51,183].

Table 1.4 Selected high confidential SNPs linked to human CRC[177-182]

SNP ID	*Genomic Location*	*Affected Gene*	*Minor Allele*	*Odds Ratio (95% Confident Interval)*	*p value*
rs6983267	8q24	*MYC*	G	1.43 (1.26-1.63)	1.72x10^{-7}
rs4939827	18q21	*SMAD7*	C	0.71 (0.63-0.81)	3.07x10^{-7}
rs16892766	8q23	*EIF3H*	C	NA	7.4x10^{-8}
rs10795668	10p14		A	NA	9.8x10^{-5}
rs4779584	15q13	*GREM1*	T	1.35 (1.14-1.60)	4.34x10^{-4}
rs3802842	11q23		C	1.20 (1.05-1.37)	6.7x10^{-3}
rs4444235	14q22	*BMP4*	C	1.12 (1.07-1.18)	1.8x10^{-6}
rs9929218	16q22	*CDH1*	T	0.88 (0.83-0.92)	1.4x10^{-6}
rs10411210	19q13	*RHPN2*	T	0.79 (0.72-0.86)	4.9x10^{-8}
rs961253	20p12		A	1.13 (1.08-1.19)	8.9x10^{-7}

1.2.3 Epigenetic mutations in CRC

Assessment of CRC epigenome has revealed that virtually all CRCs showed aberrant epigenetic changes, particularly the DNA methylome, affecting on average hundreds to thousands of genes. A subgroup of CRC (ca. 20% CRCs) has exhibited CpG island methylator phenotype (CIMP), presenting a distinct epigenome and high frequency of methylated genes, particularly in the standardized CIMP marker genes, i.a. NEUROG1, SOCS1, RUNX3, IGF2 and CACNA1G[184,185]. Hinoue et al.[85] carried out a high throughput genome-wide analysis of the DNA methylome profile of 125 CRCs and classified them into three different subtypes: CIMP-high CRC, CIMP-low CRC and non-CIMP CRC. CIMP-high CRC showed a strong association with mutation in BRAF (T1799A), methylated MLH1 and cancer specific DNA hypermethylation. CIMP-low CRC has instead the mutation in KRAS and is hypermethylated in only a subset of CIMP-high-associated methylated genes[186]. Non-CIMP CRC exhibits a high frequency with mutant p53 (71%) and intermediate frequency with KRAS mutation (33%)[187].

It had been initially proposed that the dysregulation of DNA methylation in CRC was just a bystander phenomenon, which was later disproven by innovative studies (reviewed in [188]). Fitting Knudson's 'two-hit' hypothesis for cancer development, epigenetic alteration in CRC could be the 'second hit' following the genetic mutation in oncogenes or tumor repressor genes as the 'first hit'. There seems to be two mechanisms for the DNA methylation change events in CRC: selective targeting and selective growth advantage. With regards to the selective targeting, DNA methyltransferases DNMTs are detected with elevate expression level, activity or aberrant enzyme-substrate fidelity in some tumors, compared with the normal tissue[189-191]. In addition to the overexpression, the misdirection of DNMTs to their target sequences was also observed in many different loci in tumors, explaining why DNA methylation pattern was altered in some CRCs. In contrast, regarding the selective growth advantage, the selective hypermethylation of some specific genes gives the host cells the clonal growth advantage. Although the ectopic DNA methylation has been observed to foster the metastasis of CRC, the genes affected are more predominantly involved at the early stage of tumorigenesis than progression event[192].

1.2.4 MYC oncogene in CRC

c-MYC proto-oncogen encodes a transcription factor belonging to a basic-helix-loop-helix (bHLH) family, ubiquitously expressed during embryogenesis and adult tissue compartment. It forms heterodimer with another TF MAX to bind to the DNA sequence, E-box 'CACGTG'[193]. The dimer binding could recruit chromatin remodelers resulting in acetylation of nucleosomal histones and consequently activate the target genes[194,195]. However, c-MYC is not the only partner of MAX; another transcription factor MAD/MNT could also heterodimerize with MAX in competition with MYC. The MAD-MAX heterodimer, which could also bind to E-box, would in contrast recruit transcription co-repressor (e.g. SIN3A) or HDAC to shut down the expression of MYC-target genes and therefore antagonize the MYC function (**Figure 1.6**)[196,197]. Generally, c-MYC is expressed in immature and dividing cells while MAD can only be detected in terminally differentiated cells. Coincidence of upregulation of MAD family members with diminished level of MYC is commonly noticed, and such cells start to exit the cell cycle and acquire the terminal differentiated phenotype[198].

In general, by checking the targets of c-Myc transcription factor, several groups identified its function in protein synthesis, cell proliferation and mitochondria biogenesis. The early notion that c-Myc could control cell growth emerged from the correlation between c-Myc level and expression of translation initiation factors eIF4E and eIF2α, which are both demonstrated as c-Myc target genes[200-202]. Accumulating evidences revealed that c-Myc also directly bound to the general transcription factor TFIIIB, a specific general TF for RNA polymerase III (polIII)[203]. polIII is the main RNA polymerase for generation of tRNA and 5S ribosomal RNA, both of which are important RNA molecules in protein synthesis. Not only does c-MYC regulate the protein translation pathway, but it also promotes the cell proliferation via activation or repression of genes involved in cell cycle progression. c-Myc on the one hand could induce the activity of cyclin E-CDK2 complex and boost cell to go through G1-S progression[204,205]. On the other hand, c-Myc could inhibit the expression of CDK inhibitors p15 and p21, by binding to their regulatory TF MIZ1 to compete its co-activator p300 from binding, leading to activation of CDK and progression of cell cycle[206].

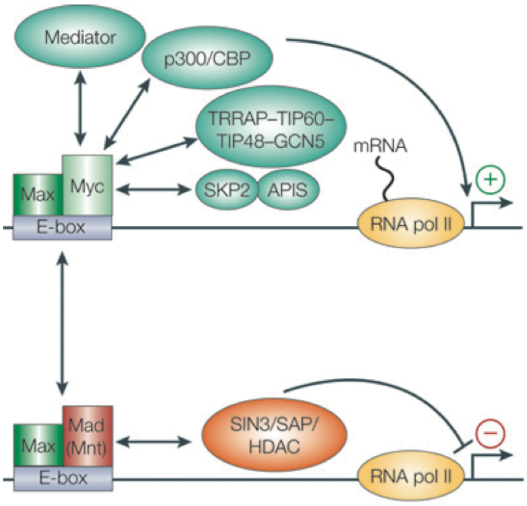

Figure 1.6 MYC/MAD/MAX and transcription regulation

Myc transcription factor forms heterodimer complex with Max and then the Max-Myc complex bind to E-box DNA sequence 'CACGTG'. Their binding could recruit other transcription coactivator and drive the target gene expression. However, Mad/Mnt can also form heterodimer with Max and bind to E-box sequence in a competition manner with Myc. When Max-Mad bind to E-box, they will recruit transcription repressor factors and turn down the target gene expression. Figure is adapted from [199].

Interestingly in addition to its ability to advance cell growth and proliferation, c-MYC also has anti-apoptotic activity[207,208]. c-MYC itself is insufficient to initiate tumorigenesis but it could cooperate with TGFα to accelerate such process by activating the survival pathway and increasing the expression of anti-apoptic factor Bcl-xL[209]. In addition, dysregulated c-MYC promotes the expression of tumor repressor gene BAX and p53 and causes apoptosis[210](**Figure 1.7**).

The most widely accepted view for this contradictory processes (oncoprotein-induced apoptosis) is that they are coupled: the prominent outcome of the synergy depends on the availability of the survival factor[211].

Dysregulation of MYC gene has been detected in many human cancers with poor prognosis, indicating its central role as an oncogene. Constitutively over expressed c-Myc could immortalize rat fibroblasts and cause uncontrolled proliferation. Declined c-Myc activity would limit the neo-angiogenesis, resume differentiation, and thus cause regression of the neoplasm, indicating that steady expression and activation of MYC is required to maintain a tumor[213].

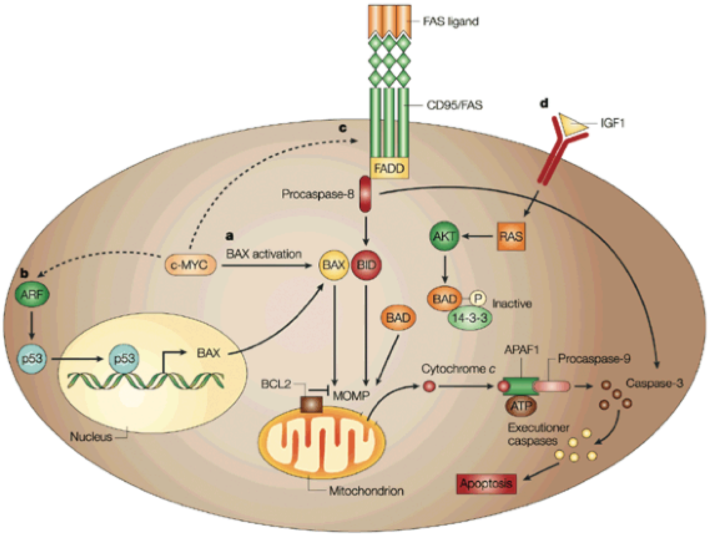

Figure 1.7 Different pathways of c-MYC involved in apoptosis

During apoptosis, c-Myc induces the release of cytochrome c from mitochondria through CD95/FAS-FADD-Caspase-8-BID-MOMP pathways (**c**). Then cytochrome c could activate caspase-9-Apaf-1 complex and cleave caspase-3 to cause apoptosis. In addition, c-Myc can also induce cytochrome c release by driving BAX expression through ARF-p53 pathway (**b**) or directly activate BAX (**a**). Figure is adapted from [212].

Very recent study showed that rather than directly regulating the target genes, c-Myc is a universal amplifyer just to increase the transcriptional level of the existing transcripts in cancer cells[214]. Thus over expression of c-Myc will also enhance the sensitivity of the cells to a broad range of pro-apoptotic signals. This phenomenon could explain why different tumor types displayed distinct sets of targets of c-Myc with little overlap and elevated c-Myc in tumor cells affects a broad spectrum of cellular events.

1.3 MOUSE MODELS OF CRC

Since CRC is one of the most common death causing diseases all over the world, studying the molecular mechanism and finding the drug targets for cancer therapies are becoming urgent. Several mouse models have been generated to study familial adenomatous polyposis (FAP).

FAP patients develop a large number of adenomas in their late teens or early twenties, some of which are capable of progressing to carcinomas, becoming invasive and producing liver metastasis. There are several FAP mouse models with different mutations in APC gene, showing more or less different phenotypes. The first and most canonical APC mutant FAP mouse model was multiple intestinal neoplasia (Min), carrying a nonsense mutation at the codon 850. The truncated APC protein product was 850 aa in long, missing the most carboxyl-terminal domains required for β-catenin and other partners binding. The homozygous APCmin mice died before birth, and the heterozygous mice have an average lifespan of 150 days compared with the average lifespan of 750-800 days for wild type mice (The Jackson Laboratory), and could develop over 100 adenomas and infrequent carcinomas in their small intestines and rare adenomas and carcinomas in colons depending on their genetic background[215,216].

In order to reconstitute the human APC mutations that are mostly detected in the last exon of the gene, another APC mutant strain, designated Apc1638N, was generated with an insertion of neomycin cassette into codon 1638, resulting in an unstable 183 kD truncation[217,218]. Similar to APCmin, the homozygous Apc1638N mice are embryonic lethal, and the heterozygous mice develop a few adenomas or carcinomas. However, heterozygous Apc1638N mice have a longer lifespan than 1 year though one case is observed that the mouse develops a liver metastasis[217]. Other APC mutant mouse models are also available, listed in **Table 1.5**.

Table 1.5 Mouse models for FAP with mutant *APC* gene

Name of the model	***Truncated Apc protein***	***Homozygous Phenotype***	***Heterozygous Phenotype***	***Lifespan***
APCmin[216]	1-850 aa	Embryonic mortality	Multiple adenomas in gastrointestinal tract	4-6 mon
Apc1638N[217-219]	Unstable 1-1638 aa	Embryonic mortality	3-4 adenomas in gastrointestinal tract, desmoid tumors, retinal epithelium abnormalities	>1 year
Apc1638T[220]	Stable 1-1638 aa	No preputial gland, nipple-associated cysts	Normal (indistinguishable from wild type)	Normal
Apc716[221]	1-716 aa	Emberyonic mortality	Multiple adenomas in gastrointestinal tract	Reduced
Apc580[222]	1-580 aa (conditional)		Multiple adenomas in 4 weeks after deletion	

In addition to various *APC* mutant FAP models which generally show adenoma phenotype in small intestines, there are some mice models with higher frequency of colon adenomas or carcinomas, e.g. mutated Pten tumor repressors, or parahox gene Cdx2, show the significantly increased number of colonic adenomas or carcinomas indicating their function in tumor suppression[223-225].

Cdx2 is a homeodomain transcription factor, highly and specifically expressed in colon and rectal tissues. The target genes of Cdx2 include the genes involved in intestinal and colonic differentiation[226]. The heterozygous mutation of Cdx2 causes the development of multiple proximal colonic polyps in the mice within 3 months[225]. The expression of Cdx2 is not detectable in the heterozygous tumor cells, suggesting that Cdx2 could be the culprit of tumorigenesis. Cdx2 has been identified as a colon specific tumor repressor gene and loss of Cdx2 has been reported in a subset of CRCs[227]. Some evidences suggest that the oncogenic role of Cdx2 mutation might result from the dysregulation of the iron transport protein expression as the cellular iron concentration is important for cell cycle progression[228].

All these models discussed above have been used to link the molecular function of the mutator genes to the CRC, and making great progress. However, lots of mutations in human CRC patients are at low penetrance and happen in the non-coding regions, the molecular significance has yet to be elucidated.

2 AIMS OF THE STUDY

The aim of this study is to understand the second genetic code, how TFs bind to the genome. We aim to advance our knowledge of the landscape of the transcriptional regulation network in colorectal cancer cells and find potential targets for cancer therapies. More specifically, three goals were to be fulfilled:

1. To find out *in vitro* DNA-binding specificities for most human TFs;

2. To map the majority of TF binding sites and establish the transcriptional regulation landscape in human CRC cells;

3. To test the finding in a mouse model.

3 MATERIALS AND METHODS

3.1 MATERIALS

3.1.1 TF full-length and DBD cDNA collection

In total, there were 984 clones of human TF full-length and 891 clones of human TF DBD used in the study. The clone collection was accomplished by PCR amplification from Mammalian gene collection (Thermo Scientific, USA), ORFeome (Thermo Scientific, USA), Megaman cDNA library (Agilent Technologies, USA), or by direct commercial synthesis (Genescript, USA). For protein expression, the cDNA of all TFs were cloned into the mammalian expression vector pDEST40-Gau-SBP[26] or pcDNA3.1-3xFLAG, or bacterial expression vector pETG20A[229], pDEST15-MAGIC[25] and pETG20A-SBP which was C-terminally tagged with SBP.

To compare the human data with mouse TF paralogs binding specificities, 444 clones of mouse TF DBD were generated by PCR from templates used in other studies[24,230]. A full list of all the clones and sequence information could be found in **Paper I, Supplemental Information, Table S1**. All constructs were verified by Sanger-sequencing in National Public Health Institute, Helsinki, Finland and MWG, Ebersberg, Germany.

3.1.2 Antibodies

With regards to the large number of chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq), 239 antibodies were used in the main batch of LoVo TF set experiments and 83 antibodies were used in the GP5d experiments. All the TF antibodies were purchased from different vendors and most of them were ChIP grade specified by manufacturers.

The antibody information, including manufacturer, catalog number and antigen TF name, were listed in the **Paper II, Supplemental Information, Table S5**.

3.1.3 Cell lines

Two human CRC cell lines, LoVo and GP5D were included in the study. LoVo cell line is derived from a 56-year old Caucasian male colorectal adenocarconima patient, and was purchased from ATCC, USA, catalog number: CCL-229;

GP5D cell line is originally derived from poorly differentiated carcinoma of the colon resected from a 71-year old Caucasian female CRC patient; the GP5D cell line that was used in this study was acquired from European Collection of Cell Culture (ECACC, UK), catalog number: 95090715.

Two human embryonic kidney derived cell lines, HEK293T and HEK293FT, were used for protein production in Study I. HEK293T cell line was purchased from ATCC, catalog number: CRL-11268; HEK293FT cell line was obtained from Invitrogen Life Technologies, catalog number: R700-07.

3.1.4 Mice and ethical issues

To generate the conditional knock out (cKO) of Myc-335 enhancer element in mice, we introduced two LoxP sites flanking the sequence chr15: 61449842-61451581 (NCBI37: mm9) in C57Bl/6 embryonic stem cell genome. The mouse derived from the modified ES cell was crossed to EIIa-Cre strain (C57Bl/6 background, Jackson Laboratory) to obtain the offspring carrying the Myc-335 null allele. All the mice used in the study were maintained in the C57Bl/6 background.

PCR was applied to genotype the mice. The primer sequence is listed below: forward primer was the same for both alleles, 5´-TAT CTG CGG GTA GTA CAC CTG T-3´; reverse primer for wild typel allel, 5´-GCT GAC AGA GAT TGC TGA CAT AA-3´, and for Myc-335 null allele, 5´-TAG TGA TTG GGT AAT AAA GAA TGA GGT C-3´. PCR conditions were denaturation 98 ºC for 15s; annealing 66 ºC for 15s; extension 72 ºC for 30s; in total for 35 cycles with Phusion High-Fidelity DNA Polymerase (Thermo Scientific, catalog number F-530L).

All animal experiments were designed and performed in accordance with Swedish laws and European Union’s ethical guideline and approved by Swedish Ministry of Agriculture (Jordbruksverket). The ethical permit number for this specific study is S111-10.

3.2 METHODS

3.2.1 Cell culture and Transfection

All four mammalian cell lines were cultured under the same conditions. The complete culture medium includes DMEM (Life Technologies, catalog number: 31885-049), 10% fetal bovine serum (FBS, Life Technologies, catalog number: 10270-106), 100 U/mL of Penicillin and 100 µg/mL of Streptomycin (Life Technologies, catalog number: 15140-122), and supplemented with 2 mM L-glutamine (Life Technologies, catalog number: 25030-024). The incubation was performed at 37 ºC with 5% CO2.

For plasmid transfection, cells were washed with 10 mL of PBS (Life Technologies, catalog number: 14190-094) and trypsinized with 3 mL of Trypsin-EDTA (Life Technologies, catalog number: 25300-054) for 5 min at 37 ºC. The reaction was quenched with 7 mL of culture medium and 500 000 cells of HEK293T and HEK293FT cells were subseeded one day in advance to a well of 6-well plate. On the day of transfection, 3 µg of plasmid DNA was diluted in 150 mM of NaCl followed by adding 2 µL of 4.5 ng/µL Polyethylenimine (PEI, Polysciences, catalog number: 23966-2) into the dilution. After vigorously vortexed, the tube was incubated at room temperature (RT) for 15 min to allow the formation of transfection complex. The complex was then applied drop-wise to the top of the cultured cells and the cells were continued for 48-hour culture before harvesting.

With regard to siRNA transfection, reverse transfection was applied. In detail, the transfection complex was prepared: 20 µL of 20 µM siRNA was diluted in 1500 µL of serum free Opti-MEM (Life Technologies, catalog number: 31985) followed by addition of 80 µL of HiPerfect transfection reagent (Qiagen, catalog number: 301707). After vigorous vortexing, the transfection complex which was applied to the top of a new 10-cm culture dish. The dish was then incubated at RT for 20 min in order for the formation of transfection complex. During the incubation time, GP5D cells were washed with PBS, trypsinized off the dish and resuspended in the complete culture medium. 2×10^6 cells were seeded on top of the transfection complex, and cultured for additional 72 hours before harvesting.

3.2.2 ChIP-seq

ChIP-seq was performed in the two CRC cell lines, LoVo and GP5D, and the mouse colon tissues, using a protocol modified from previous studies[27,51,52].

The specific part of ChIP-seq for cell line and mouse tissue is the step of chromatin preparation. Regarding cell line, 1% of formaldehyde (Sigma) was directly added into the warm cell culture medium in 15-cm dish for 10 min and the reaction was quenched by adding 2.5 M of Glycine in PBS to the final concentration of 125 mM for another 5 min incubation. Cells were collected by scraping off the dish and washed twice with ice cold PBS and spun down by centrifugation at 800 x g at 4 ºC for 10 min. Cells were suspended in 15 mL hypotonic buffer (1 mM EDTA, 10 mM KCl, 20 mM Hepes pH=7.9, 10% Glycerol, 1 mM DTT with Roche cOmplete Protease inhibitors, catalog number: 11836145001) and rotated at 4 ºC for 20 min. Nuclei would be in the pellet after centrifugation and then lysed with 3 mL of RIPA buffer (150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% Sodium deoxycholate and protease inhibitors (Roche) in 10 mM Tris-Cl, pH 8.0). The nuclear lysate was applied for sonication with Bioruptor (high power, 60 cycles of 30 s/60 s working/pause, Diagnode, B01010002 UCD-200 TO) to shear the chromatin to fragments of 200-500 bp in length and pre-cleared with Protein G Sepharose beads (GE Healthcare, catalog number: 17-0618-02) at 4 ºC for 3 hours. The chromatin was then ready for ChIP-seq or could be stored at -80 ºC.

For tissue chromatin preparation, colon was cut to fine pieces of smaller than 1 mm x 1 mm and cross-linked with 1.5% formaldehyde in PBS for 15 min at RT. Similarly, the reaction was quenched with 125 mM of Glycine at RT for additional 5 min with rotation. The tissue was then washed with ice old PBS and homogenized in 2 mL of RIPA buffer with tissue ruptor (Qiagen, catalog number: 9001271) followed by 60 strokes with tissue grinder (Wheaton, USA, catalog number: 357422) on ice. The homogenized lysate was directly applied to sonication with the same condition as above. The tissue debris was removed by 15 min centrifugation at 16 000 x g and lysate was pre-cleared at 4 ºC for 3 hours. The chromatin was then ready for ChIP-seq or could be stored at -80 ºC.

For ChIP-seq, 5 μg of antibody was added to the chromatin prep from 1 million cells, and incubated with rotation at 4 ºC over night (o/n). Thirty microliters of o/n blocked Sepharose beads were mixed with the chromatin prep and antibody, and continued with rotation at 4 ºC for 2 hours. Beads were collected with centrifugation and washed with 1 mL of different buffers: 3 times with RIPA buffer, twice with high salt RIPA buffer (RIPA with 300 mM NaCl), once with Lithium chloride buffer (250 mM LiCl, 1 mM EDTA, 0.5% IGEPAL CA-630, 0.1% Sodium deoxycholate in 10 mM Tris-Cl pH 8.0) and twice with TE (1mM EDTA in 10 mM Tris-Cl pH 8.0). Chromatin was then eluted from beads with 150 μL 1% SDS in TE supplemented with

12 μL 5M NaCl and 15 μg DNase-free RNase A (Sigma-Aldrich, catalog number: R4642) at 37 ºC for 1 hour. The cross-links were then reversed with 30 μg of Proteinase K (Thermo Scientific, catalog number: EO0492) and at 65 ºC o/n. On the second day, DNA was purified with PCR purification kit (Qiagen, catalog number: 28106) and eluted to 40 μL of EB supplied in the kit.

The ChIPed DNA was then prepared for illumina sequencing: DNA double strand end polishing was performed with 3 µl of T4 polymerase (Thermo Scientific, catalog number: EP0062), 3 μL of T4 Polynucleotide kinase (NEB, catalog number: M0201L), 1 μL of Klenow fragment (Thermo Scientific, catalog number: EP0052), 4 μL of 10mM dNTP and 10 μL of NEB T4 ligation buffer in 100 μL reaction volume at 20 ºC for 30 min, and purified with Qiagen PCR purification kit and eluted in 30 μL of EB; Adenosine-addition was performed with 3 µl of Klenow Fragment Exo- (Thermo Scientific, catalog number: M0212L), 1 µl of 100 mM dATP, and 10 µl of 10 × Klenow Buffer supplied together with the Klenow enzyme in 50 µl reaction volume at 37 ºC for 30 min and purified with Qiagen PCR purification kit and eluted in 30 μL of EB; illumina HiSeq adapter was then ligated to the DNA fragments with T4 DNA ligase (NEB, catalog number M0202L) at 16 ºC o/n and purified with Qiagen PCR purification kit and eluted in 30 μL of EB. DNA size selection for 300-500 bp with 2% agarose gel was carried out for the ligation product. Purified DNA was enriched by 16 cycles of PCR and sequenced using illumina GAIIx or HiSeq 2000 sequencer (The Karolinska High Throughput Center, Huddinge, Sweden).

Duplicated reads were removed to avoid PCR bias and the reads were mapped to human reference genome (hg18) using the Burrows-Wheeler Alignment tool (BWA)[231] and the mapping quality threshold was set as 20. After mapping, peak and peak summit position calling was performed using MACS[232]. Peaks were filtered for downstream analysis based on the criteria: unadjusted $p < 10^{-5}$, fold change over IgG control ≥ 2 and false discovery rate (FDR) ≤ 5%.

3.2.3 RNA isolation and Microarray

To test the expression level of all TFs in LoVo cells and verify the knockdown efficiency in GP5D cells, mRNA extraction was performed with RNeasy kit (Qiagen, catalog number: 74106). For transcript analysis of mouse intestinal tissues, Trizol (Life Technologies, catalog number: 15596-026) was used to extract the total RNA according to the manufacturer protocol followed by purification with RNease kit (Qiagen).

For RNA microarray, mRNA was further cleaned up with RNeasy MinElute clean up kit (Qiagen, catalog number: 74204) and eluted in 15 μL of RNase free water. The mRNA was then sent to Karolinska Bioinformatics and Expression Analysis core facility (BEA, Huddinge, Sweden) for microarray analysis using Affymetrix HG-U133 Plus 2.0 array with triplicates.

Raw data was normalized using the robust multi-array average with GC-correction (GCRMA), implemented in the R package gcrma and using probeset definitions from a custom cdf based on ensemble gene annotations[233]. Genes with a log2 probeset intensity value of four or more were considered as potentially expressed, more than seven as highly expressed.

For checking the knock-down efficiency or c-Myc transcript level in mouse intestinal tissues, isolated RNA from GP5D cells or mouse tissues were used as the templates for reverse transcription with High capacity reverse transcription kit (Life Technologies, catalog number: 4368814) to obtain the cDNA. Quantitative PCR (qPCR, instrument: Roche LightCycler 480; reagent: Maxima, SYBR Green/ROX qPCR kit, Fisher Scientific, catalog number: 11591545) was performed to test the siRNA target transcript level in GP5D cells or c-Myc level in mouse intestinal tissues normalized with internal control (β-actin).

3.2.4 High throughput Systematic Evolution of Ligands by Exponential Enrichment (HT-SELEX)

In HT-SELEX[26], TF protein fused with Streptavidin Binding Protein (SBP) tag expressed in mammalian cells was immobilized in streptavidin-coated plate (Thermo Scientific, catalog number: 15502). The plate coated with TF proteins was washed with lysis buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1% Triton X-100), high salt buffer (50 mM Tris-HCl pH 7.5, 500 mM NaCl, 1% Triton X-100) and low stringent ligand binding buffer (20 mM HEPES pH 7.05, 140 mM KCl, 5 mM NaCl, 1 mM K_2HPO_4, 2 mM $MgSO_4$, 100 µM EGTA, 3 µM $ZnSO_4$). The plate wells were blocked with 0.5% bovine serum albumin (BSA; Sigma-Aldrich, catalog number: A9418) in low stringent ligand binding buffer at RT for 30 min. The plate wells were washed twice with low stringent ligand binding buffer before applying the DNA oligos for selection.

DNA double-stranded oligos with randomized 14N, 20N, 30N and 40N adapted to illumina GAIIx or HiSeq2000 system were diluted in low stringent ligand binding buffer and added to all wells for selection by TF proteins with vigorous shaking at RT for 2 h (oligo sequence information detail was listed in **Paper I, supplemental Table S1**). Barcode for each well was included in the DNA oligos flanking the randomized regions, with 5-6 bp and 0-3 bp barcode sequences before and after the regions.

After binding, the plate wells went through intensive washing steps with the low stringent ligand binding buffer: each well was washed once with 80 µL of low stringent ligand binding buffer, twice with 90 µL, 3 times with 110 µL, 3 times with 130 µL, 6 times with 300 µL, 4 times with 310 µL and finally 6 times with 410 µL. The bound DNA oligos were then eluted to 65 µL of elution buffer (10 mM Tris-Cl, pH 8.0 containing 1 mM EDTA, 0.05% Tween 20) by incubating the plate at 85 ºC for 15min. Thirteen microliters of eluted DNA was enriched by PCR using the following primer set: forward primer, 5’- TCC ATC ACG AAT GAT ACG GCG ACC ACC GAA CAC TCT TTC CCT ACA CGA CGC TCT TC, and reverse primer, 5’-CGG AGT CGG CAA GCA GAA GAC GGC ATA CG. PCR conditions were denaturation 98 ºC for 15s; annealing 66 ºC for 15s; extension 72 ºC for 30s; in total for 25 cycles with Phusion High-Fidelity DNA Polymerase.

Different PCR products were pooled together depending on the barcode sequence and subjected to illumina sequencing with GAIIx or HiSeq2000 in Karolinska

High Thtoughput Center (KHTC). Meanwhile the PCR products would be used as the input oligos for the next round of selection, which would be repeated three times (in total 4 cycles) for each TF protein in HT-SELEX followed by sequencing.

The solution adding and transferring was performed with a liquid handling workstation (Agilent Bravo and BenchCel, KHTC). The plate washing was implemented with two consecutive plate washers (BioTek Washer ELx405 Select CW, KHTC).

The sequencing reads were sorted based on the barcodes. To model TF binding, we generated position weight matrices using a multinomial method that yields profiles that are similar to those generated using maximum likelihood methods such as BEEML[234]. For each TF binding model, the most frequent k-mer sequence was detected by our custom based software IniMotif[26] and used as the seed for the multinomial model. For multinomial model, the seed was matched exactly (multinomial 1 model) outside the position considered and all the matched reads were used to count the weight of each nucleotide at that position. Putting together the nucleotide weight for all positions, position weight matrix (PWM) was generated. Use of the multinomial method was selected because: 1) it exactly corresponds to the PWM model representation, 2) the number of enriched sequences was very high, alleviating the need to analyze a large sequence space, and 3) use of larger sequence space commonly resulted in mixing of multiple different models.

In general, cycle used for PWM generation was selected from cycles 2 to 4 in such a way that more than 1000 subsequences were included to the model after background correction. After initial PWM generation using the most frequent k-mer as seed, the seed was made more redundant to accommodate more sequences at positions where the frequency of the most common base was < 0.5. At these positions, we used either N, or where the ratio between the second and third most frequent bases was > 2, we used the IUPAC symbol for the two most frequent bases (R, Y, M, K, S, or W). If the length of the seed was longer than 10 bp, a multinomial model allowing a single mismatch at any position was used. Seed sequences were further manually curated to prevent mixing of two distinct binding modes, and to distinguish between monomer and dimer models.

Multiple seeds were used for the same factor if the IniMotif analysis of 6 to 12-mer sequences, or if plotting of the observed k-mers versus those expected from the PWM revealed that the first model did not explain the most enriched k-mers, or if enrichment of dimers was observed. Models were corrected for background by subtracting normalized counts from the previous round as described in Jolma et al., (2010)[26] using the equation $M_{corrected}$=M_{k+1}- lambda *M_k, where lambda is the fraction of DNA carried over non-specifically estimated using 8-mer frequencies, and M_{k+1} and M_k are the uncorrected matrices normalized for number of input sequences from cycles k+1 and k, respectively.

3.2.5 DNase I Hypersensitivity Assay (DHS)

To test the DNA accessible for protein binding, DHS was performed in LoVo and GP5D cells using a protocol modified from previous study[235] and adapted to illumina sequencing system. Nuclei were isolated by washing the cells in 1ml ice-cold RSB buffer (10 mM Tris-Cl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$) and lysing the cell membrane with RSB Lysis Buffer (10 mM Tris-Cl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630) and continuing to incubate at 4°C for 10min. After incubation, centrifugation at 2000 x g at 4°C for 5 min was performed to collect the nuclei pellets. During the incubation time, 4 enzyme mix tubes were prepared and DNase I was diluted in 10 µL of the 1 x DNase Incubation Buffer (400 mM Tris-HCl pH 7.9, 100 mM NaCl, 60 mM $MgCl_2$, 10 mM $CaCl_2$): 0, 0.1, 0.25 and 1U of DNase I (Roche, catalog number: 04 716 728 001). The nuclei pellets were diluted with 400 µL of 1x DNase Incubation Buffer and mix by flicking the tube. A 100 ul aliquot of the re-suspension was carefully added to each enzyme mix tubes and incubated at 37 °C for 15 min. The reaction was quenched by adding 30 µL of 500 mM EDTA to each reaction tube. 20 µg of DNase free RNase A was added to treat the samples at 37 °C for 1 h followed by removing the DNase I, RNase A and other bound proteins by adding 40 µg of Proteinase K and incubate them at 56°C for 1 hour. The DNA was purified using phenol/chloroform extraction and dissolve the DNA in 50 µL of TE buffer (10 mM Tris-Cl, pH 8.0, 1 mM EDTA). To enrich the DNA fragment released by DNase I, all the DNA samples were run in 1.5% TAE agarose gel electrophoresis at 60 Volts for 2 h and collect the fragments around 100bp. The gel slice was purified with QIAquick Gel Extraction Kit (Qiagen, catalog number: 28704) and the DNA was eluted in 40 µL of elution buffer supplied with the kit.

The eluted DNA was subjected to illumina sequencing library preparation using the same protocol described for ChIP-seq earlier in this section. The DNA was sequenced with illumina HiSeq2000 sequencer and duplicate reads were removed as for ChIP-seq. The reads were mapped to human hg18 reference genome with BWA. DNase I cuts were indicated by the counts of 5' position of the raw reads.

3.2.6 Cell synchronization

Double thymidine block was applied to synchronize the cells in G1/S phase boundary and single thymidine followed by nocodazole block was used to synchronize the cells in early M phase[236]. For G1/S phase synchronization, LoVo cells were cultured to 30-40% of confluence and washed with pre-warmed PBS. The medium was changed to DMEM containing 10% FBS and 2 mM of thymidine (Sigma-Aldrich, catalog number: T1895) and the cells were continued to be cultured for 12 h (the first thymidine block). The cells were washed with pre-warmed PBS and the medium was changed to normal DMEM with 10% FBS and continue to culture for additional 8 h. The cells were washed with pre-warmed PBS and the medium was changed to DMEM containing 10% FBS and 2 mM of thymidine again (the second thymidine block). The cells were cultured for additional 12 h before harvesting.

For M phase synchronization, after the first thymidine block, LoVo cells were washed with pre-warmed PBS and changed to fresh DMEM medium containing 10% FBS and 200 ng/ml nacodazole (Sigma-Aldrich, catalog number: M1404) for 10 h before harvesting.

To monitor the cell cycle status, the cells were fixed with 70% pre-chilled (-20 ºC) ethanol o/n and stained with 20 µg/mL propidium iodide (PI; Sigma-Aldrich, catalog number: P4864) in 0.6% Triton-X100 in PBS. The stained cells were then analyzed with Flow Cytometry (FACSCalibur, BD Sciences).

3.2.7 Sister Chromosome Proximity Ligation (SCPL)

SCPL is based on detection of the ligation product of a single HindIII restricted fragment end to the HindIII end of the corresponding fragment in trans on the sister chromosome. Ligation of the HindIII sites to themselves is only possible if two sister chromosomes are located in close proximity to each other. As an internal control for ligation efficiency, we analyzed the ligation of the HindIII site in cis to the other end of the same fragment. SCPL experiments were performed on two cohesin proximal and two distal regions using chromatin from S and M phase arrested LoVo cells. DNA from exponentially growing cells was used as a control. Chromatin was digested with HindIII, diluted and ligated, followed by RAD21 ChIP and qPCR-based detection of trans (between sister chromatids) and cis (self) ligation products.

Primers were designed in a way that the primer closer to the HindIII site contains a GC-rich 50 bp non-genomic sequence that allows specific amplification of a two-primer product, and forms a stem-loop that suppresses the single primer product. The other possible single primer product is suppressed by formation of a very long stem loop. A ChIP-3C protocol was adapted for use in sister chromosome proximity ligation assay as follows[237]. Growing LoVo cells and S or M phase arrested LoVo cells were cultured to 60% - 70% confluence and crosslinked with 1% formaldehyde. Nuclei were extracted with ice-cold hypotonic lysis buffer (20 mM HEPES, pH 7.9, 1 mM EDTA, 10 mM KCl, 10% glycerol) and washed with 1.14 x NEB buffer 2 with protease inhibitors (Roche). Nuclear membrane was broken and chromatin was released with 0.3% SDS and 2% Triton X-100 in NEB buffer 2. Chromatin was digested by 1,000 Units of Hind III (NEB) at 37 °C overnight and inactivated at 65 °C for 30 min. Chromatin was diluted by adding 14 volumes of 1.15 x T4 ligation buffer (NEB) and ligated at 16 °C for 4 h and at room temperature for 30 min. 50 µL of lysate was saved as input and the rest diluted with 9 volumes of RIPA buffer for ChIP. Antibodies against RAD21, SMC1A and SMC3 were used to pull down the cohesin-DNA complex. The complexes were eluted and crosslinking was reversed as described before. Both input and cohesin bound DNA was extracted using phenol/chloroform/isoamylalcohol (25:24:1, v/v; Life Technology, catalog number 15593-031) and dissolved in 10 mM Tris-Cl buffer, pH 8.0.

The ratio of cis-ligation and trans-ligation between sister chromatids was determined by quantitative PCR (qPCR; Light Cycler 480, Roche). Input DNA was used as a template for cohesin distal regions and ChIPed DNA was used as template for cohesin bound regions. If the two sister chromatids are bound closeby, the trans-ligation will take place at a higher ratio than the *cis*-ligation. Two pairs of primers were used to detect the trans-ligation of two distinct cohesin bound regions (proximal-trans1-fw 5'-CACGGGGCTTTCACCTGAACTAACC, proximal-trans1-rev 5'-GGCGACTGGTGTACAACCTCAGAAGC; proximal-trans2-fw: 5'-ACCATGAGCCACCACTGGAAGGT, proximal-trans2-rev: 5'-GGCGACTGGAACAAAGTACCCAAAGC). Two pairs of primers were used to detect the *cis*-ligation of cohesin bound regions (proximal-cis1-fw: 5'-

GGAGTGCAAACTCCGCTCCTACCTA, proximal-cis1-rev: 5’-TCAGAGCTTTGGACTTGTGGTAGCC; proximal-cis2-fw: 5’-GGGGGCTGCCACAACAAAGTA, proximal-cis2-rev: 5’-GCCTCCTTCCTCTTCCATCATTGTC). Two pairs of primers were used to detect the *trans*-ligation of cohesin distal regions (distal-trans1-fw: 5’-CCAGTCGCCGACACAACACTAAAGC, distal-trans1-rev: 5’-TATACGGTGAGGTCACGCTTCATGC; distal-trans2-fw: 5’-CCAGTCGCCTCCAGCCAAACTAAGC, distal-trans2-rev: 5’-AAGCCAGAAGAGAGTGGGGGTCCAT). Two pairs of primers were used to detect the *cis*-ligation of cohesin distal regions (distal-cis1-fw: 5’-GCCCGTAGGGATTTACTGACACCTC, distal-cis1-rev: 5’-TGCCTACTCCCCTTTGACCTTCTTC; distal-cis2-fw: 5’-ATATCACCAGCAGAGGCTGCAGAAC, distal-cis2-rev: 5’-GAGTGGGGGTCCATATTCAACCTTC). Single product was detected based on the melting curves that showed a single peak.

3.2.8 Counting of polyps in APCmin mice

The APCmin and the APCmin; Myc-335-/- mice were analysed at 4 months of age. The mice were euthanised and intestinal tract was removed. It was flushed with cold PBS and opened. All visible polyps were counted and the intestines were fixed in 4% PFA/PBS solution overnight and then stored in 70% ethanol. The polyps were subsequently scored using a stereomicroscope.

3.2.9 Public data deposition

All the sequencing data and microarray data were deposited to a public server, listed below:

HT-SELEX sequencing data were deposited in European Nucleotide Archive (ENA) under the accession number ERP001824.

Human ChIP-seq data for verification of the HT-SELEX results were deposited in ENA under the accession number ERP001826; other Human ChIP-seq data were deposited in Gene Expression Omnibus (GEO) under accession number GSE49402; mouse ChIP-seq data were deposited in ENA under accession number ERP001919.

DHS and genomic DNA sequencing data for LoVo and GP5D cell lines were deposited in ENA under accession number ERP002229.

RNA microarray for human CRC expression data were deposited in GEO under accession number GSE48448; Exon array data for mouse colon tissues were deposited in ArrayExpress database under accession number E-GEOD-41219.

4 RESULTS

4.1 STUDY I (*IN VITRO* STUDY): DNA-BINDING SPECIFICITIES OF HUMAN TFS

4.1.1 TF binding profiles and comparison to the existing databases

In order to determine the DNA-binding specificities of mammalian TFs, we cloned 891 human and 444 mouse TF DBDs, and 984 human TF full-length cDNA into the mammalian expression vectors, fused with SBP-tag sequence and tried to express all the proteins in mammalian cell line HEK293FT.

After performing the HT-SELEX for all of the proteins, we acquired the robust enrichment of binding specificities for most of the TF families. In total, we obtained 830 binding models for 303 human DBDs, 84 mouse DBDs and 151 human TFs, of which 79 proteins have models for both DBD and full length, being the largest collection of mammalian TF binding specificities and covering more than 50% of all high-confidence TFs at a 90% similarity threshold.

Some TF families have prominently lower success rate than the others: there are two TF families with the low coverage: high-mobility-group (HMG) and C2H2 zinc finger proteins. Earlier studies showed that some of the HMG proteins did not bind to DNA in a sequence specific fashion or HMG domains were largely involved in protein-protein interaction[238]. For C2H2 zinc finger proteins, the domains were discovered with other function, such as protein-protein interaction and RNA binding in addition to DNA binding[239,240]. Consistent with our data, the low successful rate for C2H2 zinc finger proteins DNA binding specificities was also observed in other independent assays from independent group (unpublished from Tim Hughes Lab, University of Toronto, Canada).

In total, our data covered DNA binding models for 31 TF families including over 450 distinct mammalian TFs (**Table 4.1**). TFs from the same family tend to bind to identical or similar DNA sequence than those from other families, except that C2H2 zinc figure TFs exhibited the largest diversities within the family.

As we obtained a number of models for both DBD and full length of the same TF, it is possible for the first time to systematically compare the DNA binding specificities between them. From the 79 pairs of TFs, we found that most TF DBDs and full-length proteins showed the same binding specificities, with only one exceptional case of ETS factor ELK1. ELK1 DBD could only bind to the monomer site while the full length ELK1 can bind to both monomer and dimer sites. These results suggested that TF DBD essentially determined the DNA binding specificities.

Due to the innate advantage of SELEX in which long sequence ligands could be used, homodimer binding could be detected for many TFs that were known to bind to DNA as monomers.

Table 4.1 Coverage of TF families with PWM motifs[6,27,241-243]

TF family	*Number of TFs with model*	*Number of TFs in the family*	*Percentage of TFs with motifs*
ZnfC2H2	53	675	7.85
Homeodomain	137	257	53.31
bHLH	39	114	34.21
HMG	18	57	31.58
bZip	23	56	41.07
Forkhead	16	48	33.33
Nuclear Receptor	23	45	51.11
ETS	24	27	88.89
MYB	2	21	9.52
ZNF-GATA	3	19	15.79
POU	13	18	72.22
T-box	12	16	75.00
znfBED	1	14	7.14
E2F/TDP	6	11	54.55
CENPB	1	10	10.00
IRF	6	9	66.67
RHD	4	9	44.44
MAD	1	8	12.50
SAND	1	8	12.50
RFX	4	8	50.00
Heat_shock	4	7	57.14
CP2	2	5	40.00
AP-2	3	5	60.00
MADS-Box	4	5	80.00
CTF/NFI	3	4	75.00
EBF	1	4	25.00
P53	3	3	100.00
GCM	2	2	100.00

When we compared the HT-SELEX data with the existing databases, we found that the data from HT-SELEX greatly advanced the number of mammalian TF binding models especially for those with long binding sites. In more detail, we specifically compared our HT-SELEX data with the protein binding microarray (PBM) dataset[24,230] and found that for the eight TF families that primarily bind to DNA as monomers two methods displayed the similar number of models and almost identical binding profiles. For the other 23 families with both monomer and dimer binding sites, the dimer motifs were mostly missing from the PBM data but retained in HT-SELEX (**Figure 4.1A**).

In order to confirm the dimer binding in vivo, we performed ChIP-seq for these TFs and found that the dimer sites from HT-SELEX assay were also enriched in ChIP-seq from MEME analysis (**Figure 4.1B**).

We also found that there were some TFs for which more than one model existed to describe their binding specificities in PBM database. However, with HT-SELEX, some of the two PBM models existed just because they broke the long binding sites into two shorter sites (**Figure 4.1C**).

4.1.2 Classification of TFs based on their binding specificities

As we discovered that TFs from the same structure family tended to bind to similar DNA sequences, we wondered whether we could further divide them into subclasses by the DNA binding specificities.

Indeed, similar to the assay performed previously for ETS families[27], only by comparing the DNA sequence of the 24 ETS factors acquired from HT-SELEX, we were able to corroborate all of them to the four defined subclasses (**Figure 4.2A**). Other families displaying one-to-one relationship between protein and binding sequence were identified: for example, five subclasses of GLI-like C2H2 zinc finger proteins, four

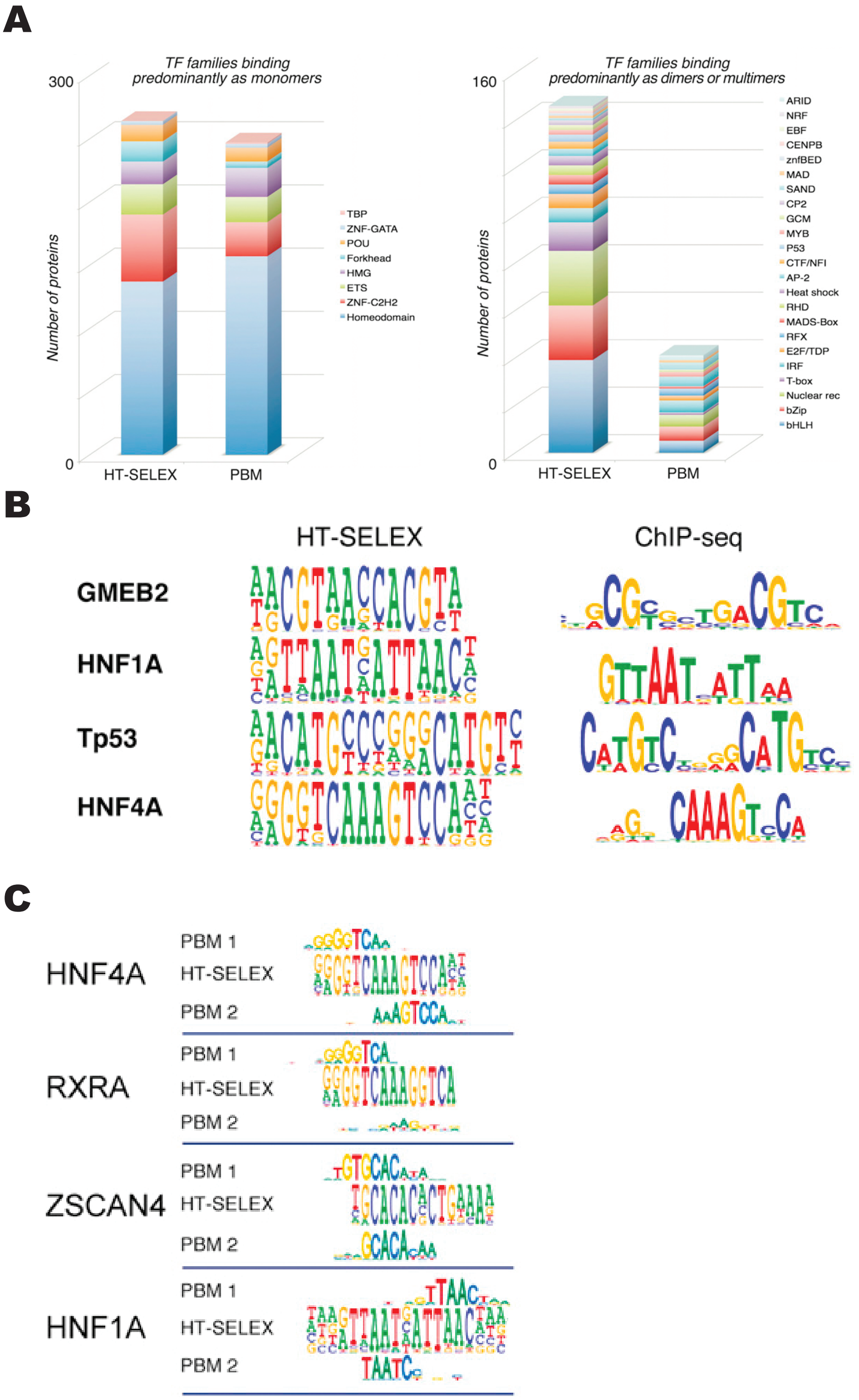

Figure 4.1 The homodimer binding

(A) Comparison of the number of TFs for which a model has been derived using PBM or HT-SELEX. Colors indicate different structural TF families that bind DNA primarily as monomers. Left panel shows the model of monomer and right panel shows the models of dimer or multimer.

(B) Examples of confirmation of dimer models derived from HT-SELEX with the *in vivo* data obtained from ChIP-seq.

(C) Examples showing that HT-SELEX could model the long binding specificities in homodimer binding while PBM has broken them into two models that together contribute to the whole binding motif. PBM1, and PBM2 are the two broken models.

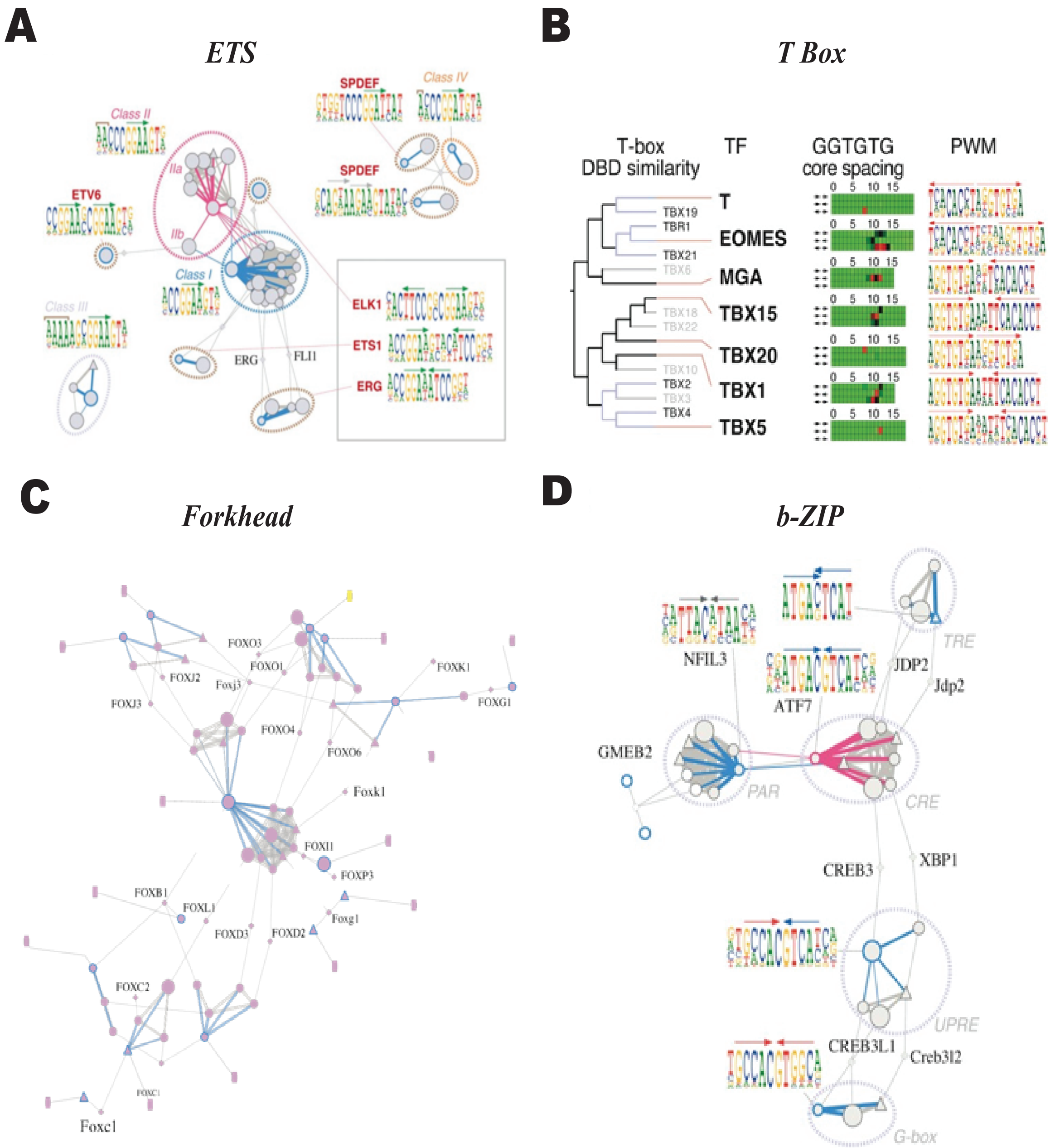

Figure 4.2 Classification of TFs according to the binding specificities

(A) ETS factors. HT-SELEX can accurately identify the four known ETS subclasses (indicated by colored ovals). Additional specificity determinants in classes II, III, and IV are indicated by brown brackets, and a novel dimer in ETV6 (class II) and two novel putative dimers in SPDEF (class IV) are indicated by brown dotted lines. Box indicates three different homodimeric sites within class I. Logos for representative PWM models are shown; green and gray arrows indicate GGA(A/T) and AGAA sequences, respectively.

(B) Classification of T box TFs based on dimer orientation and spacing. Left panel shows amino acid similarity dendrogram of T box DBDs. TFs for which models were not obtained are in gray. Middle panel shows heatmap displaying spacing and orientation (arrows) preferences of the enriched GGTGTG subsequences (red indicates max counts; green indicates 0); scale represents distances between the subsequence starting points. Right panel shows PWM describing most enriched dimeric binding site for each TF.

(C) Classification of forkhead TFs based on dimer orientation and spacing. All forkhead proteins show the similar monomer site which could be subdivided into three different classes based on their homodimer orientation and spacing.

(D) A subset of bZIP TFs recognizes two types of target sites in a tiled pattern, covering four site types. Arrows above the logos indicate half-sites; black specifies TTAC, blue designates ATGAC, and red shows GCCAC. Note that JDP2, CREB3, XBP1, CREB3L1, and Creb3l2 each can bind to two different site types, forming a tiled pattern ranging from TRE element (top) to G box.

subclasses of bHLH proteins, four subclasses of PAX, two classes of E2F, two classes of HSF, two classes of MADS, etc.

With the high resolution of HT-SELEX models, we could even further divide three out of the four well-studied ETS subclasses with additional new specificity determinants, the dimer spacing and orientation when they shared the same monomer site.

Not only for ETS family members, other TF structural families could also be sub-classified by the dimer spacing and orientation. Dimer spacing is defined as the number of nucleotides between the first nucleotides of the two monomer sites. Dimer orientation was defined as three types: head-to-head, the first monomer site was inverted; head-to-tail, two monomer sites were in the same direction; and tail-to-tail, with the second monomer site was inverted. T-box and forkhead proteins both displayed the same monomer binding specificities within their families. However, T-box could be subdivided into seven classes based on the dimer spacing and orientation and forkhead proteins could be further divided into three classes (**Figure 4.2B**, **C**).

Some cases are even more complicated, for example bZIP family. The bZIP proteins can bind to two different types of monomer sites. Different subclasses of proteins could be defined as a tiled pattern by their preference for different half sites or orientation and spacing of the two half sites. In detail, NFIL3 and GMEB2 bind to a dimer site called 'proline- and acidic-amino-acid-rich' binding site (PAR), with the head-to-head pattern of two half sites with 'TTAC' sequence. ATF prefers to bind to a dimer site called 'cAMP response elements' (CRE) composed of two head-to-head half sites 'ATGAC' with the spacing of 4 nucleotides. JDP2 can bind to this 'CRE' site too but it also binds to an additional dimer site called 'TRE' composed of two half sites with the same sequence 'ATGAC' but with the spacing of 3. Similarly, CREB3 and XBP1 can bind to 'CRE' and an additional dimer site 'UPR-dependent cis-acting element' (UPRE) of two different half sites, 'GCCAC' and 'ATGAC'. This 'UPRE' dimer site is shared by CREB3L1 and CREB3L2, although they can bind to 'G-box' dimer sites with two identical half sites of 'GCCAC' (**Figure 4.2D**).

4.1.3 Base pair interdependency in TF binding

To analyze how interdependency of two DNA sequence positions will affect TF binding, we compared the counts of all 16 nucleotide pairs at those positions with the expected counts from PWM models assuming they are independent. When we plotted the observed counts against the expected counts, we found that PWM was in general a good model for most TFs. If we calculated the correlation coefficient between the observed and expected counts of any two pairs, there were only less than 1% of all pairs in all models displayed the correlation coefficient lower than 0.9. We found the independence was even stronger when the two positions were three bases apart from each other (**Figure 4.3A**).

The main factor affecting the base pair interdependency was the base-stacking effect. It had been well established that different compositions of dinucleotides would affect the structure of DNA and could be used to predict a large number of properties of DNA, such as melting temperature and topology of the base pair[244,245]. Our results indicated that the base-stacking might affect the DNA structure for TF binding which should be considered in the quantitative analysis for TF-DNA binding. PWM was generally a good approximation for binding model and we could improve the quantitative prediction by considering the factor of adjacent base stacking.

Along with the general factor affecting base pair interdependency, there are some other types of interdependency for a specific group of TFs. We observed that a stretch of adenines or thymidines flanked the core motif of some TFs; for example, four 'A's preceding the core sequence 'AGCGGAAGTA' in the ETS factor SPI1 binding sites; or three 'T's preceding the core sequence 'TCCCGCCA' and three 'A's succeeding it in the E2F factor 'E2F7' binding motif. To exclude the 'A-stack' phenomenon from artifact and corroborate it *in vivo*, we performed ChIP-seq experiments for some of the factors, SPI1, MAFG and E2F7, and we did observe it in TF binding to genomic DNA as well. Among all the TF binding specificities we obtained, the A-stack was unveiled in several TF structural families, ETS, bZIP, E2F, CUT, Homeodomain, T-box and CP2 (**Figure 4.3B**).

Another type of base pair interdependency was only observed in posterior homeobox TFs, HOX9-13 and parahox TFs, CDX1 and CDX2. The paralogous posterior HOX proteins (clusters A, B, C and D), bound to the same sequence while the difference existed amongst HOXes 9, 10, 11, 12 and 13. All of them shared a partial binding site 'TAAAA' whereas the first 4 positions varied between them. However, each individual TF could not bind to all combinations of the four nucleotides at these positions, namely it preferred a few of the 4-mers. So the binding specificities could not easily be presented by PWM model (**Figure 4.3C**).

The fourth type of base pair interdependency we have discovered was in HMG TF, SOX9. It displayed the head-to-head pattern of pseudo-homodimer binding and the extremely strong correlation between the bases present in one half site and the corresponding bases in the other half site. This effect may not be mediated by the homodimer binding; instead it could come from that TF SOX9 binds to the stem loop formed from a single stranded DNA (**Figure 4.3D**).

The last type of base interdependency was identified in the homodimer binding. Some TFs bind to DNA as a very tightly packed homodimer, e.g. FLI1, MEIS2 and PKNOX2, which makes the overlapping sites show strong correlation. This type of binding could be modeled as an asymmetrical binding with non-palindromic PWMs **(Figure 4.3E).**

4.1.4 TF binding models

As we observed the strong base pair interdependency in some TFs discussed above, the simple PWMs could not be used to model the DNA binding. We needed special models for these TFs.

The first example is the A-stacking model. We introduced the first Markove chains to the 'Adjacent-dinucleotide-model' (ADM) in order to better describe the A-stacking effect on TF binding prediction. It allowed the scoring of k-mers that were shorter than the model itself. When we tested the 10-mer predictions with both the traditional PWM model and ADM model, the latter model showed the better prediction of the 10-mer for E2F3 binding with ‘A’ stretches flanking the core site.

The second model is for homodimer prediction. It considered the homodimer orientation and spacing weight in addition to the traditional PWM model. We tested the prediction of TBX20 that could bind to DNA as homodimer but with several different half site orientation and spacing with different affinities. We plotted the observed gapped 4-mers against the expected counts. The new model greatly advanced the prediction.

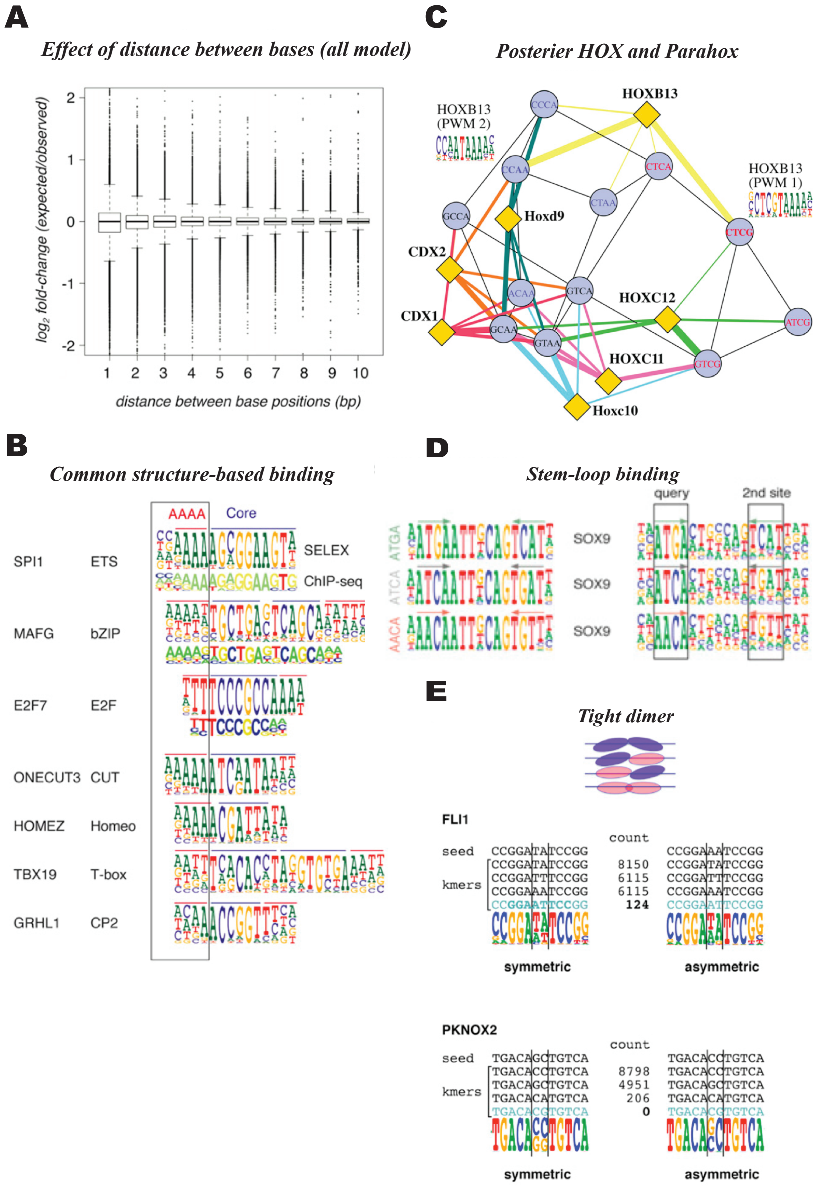

Figure 4.3 Dinucleotides interdependency

(A) Box plot shows the trend of interdependency of two positions. The x-axis shows the distance between two nucleotide positions; y-axis shows the effect of interdependency calculated from the log2 fold change between observed counts and expected counts from PWM model assuming each position is independently contributing to the binding.

(B) A stretch of A or T bases (box, red line above logos) is commonly observed adjacent to core TF binding sites (blue line). Models generated using ChIP-seq (short) followed by motif discovery are shown below HT-SELEX-generated models (tall). SPI1 motif is from [27].

(C) Posterior homeodomains exhibit strong correlations between bound positions. Diamonds represent the indicated posterior homeodomain proteins, and circles represent enriched 9-mer sequences (circles, first four bases shown, last five bases are TAAAA). Edges are drawn between kmer nodes if their Hamming distance is 1, and between a protein and a k-mer node if the k-mer is enriched by the protein. Edges between protein and k-mer nodes are colored for clarity, and their thickness represents the extent of the enrichment. Logos indicate two different PWM models for HOXB13 that are built using non-overlapping sequences (blue and red).

(D) HT-SELEX detects SOX binding to inverted repeat sequences that apparently represent a stem-looped single-stranded DNA. Left: three different apparent dimers are bound by SOX9. Right: sequences flanking ATGA (top), ATCA (middle) and AACA (bottom) query matches reveal that in each case, an inverted repeat of the query sequence appears 3’ to the query after a 7 bp gap (2nd site), suggesting that the bound sequence is a stem-loop formed from a single-stranded DNA. This interpretation is also consistent with the preferential presence of such matches in only one strand of the selection ligand (not shown).

(E) Asymmetric binding of the monomers is observed when monomer sites are located close together. Top: Close packing of target sites can affect monomer specificity. Protein can bind to an optimal site (pink oval) or to a weaker site (blue oval). Middle: The consensus sequence of FLI1 dimer expected by monomer specificity GGAATTCC (bottom, gray) is very weakly enriched, whereas sites where one or both monomers bind to a GGAT core are strongly enriched. Note that the asymmetric PWM (right) correctly describes lack of enrichment of the GGAATTCC site, whereas the symmetric PWM (left) predicts much higher enrichment for this sequence. Bottom: Similar effect is observed in a PKNOX2 dimer.

4.2 STUDY II (*EX VIVO* STUDY): TF BINDING IN HUMAN CELLS OCCURS IN DENSE CLUSTERS FORMED AROUND COHESIN ANCHOR SITES.

4.2.1 TF binding forms dense clusters

In order to query the binding sites of human TFs in CRC cells, we carried out a large scale of ChIP-seq experiments in two CRC cell lines, LoVo and GP5D. It covered the vast majority of highly expressed TFs in these two cell types. In addition, we also included enhancer, promoter and insulator marker proteins in the ChIP-seq, such as p300, MED12, monomethyl-lysine 4 of Histone H3 (H3K4me1), etc.

The LoVo experiments were carried out in two separate batches: the main batch contained 239 antibodies, and the control batch contained 322 antibodies which include TFs with lower expression level than the main batch TFs. Most of the downstream analyses utilized the data from the main batch, unless otherwise indicated. We also developed an automated quality control (QC) piple line to filter out any failed experiments. Finally, we had 112 successful TF ChIP-seq experiments from the main batch. The coverage of TFs in a single cell type exceeded the largest existing database of K562 cell line (44 TFs) published by ENCODE consortium.

It had been established that TF bound to genomes of human and other model organisms in restraint regions with high density of TF binding (hotspots[246,247]). We analyzed the successful ChIP-seq in LoVo and discovered a striking degree of clustering of the TF binding: over 75% of TF binding peaks occupied only 0.8% of human genome. It had been well known that TFs act in a combinatory fashion to regulate the target transcription[46]. The clustered binding of functionally related TFs is indeed widely observed[246,248,249]. To determine whether the higher degree of clustering results from the functional related TFs, we used a 2-kb cutoff to identify TF pairs that co-occurred. After the gene ontology (GO) analysis, the strong co-occurrence was observed for cohesin and mediator subunits that were reported to cooperate to regulate transcription by bridging enhancer and promoter[60], and for TFs sharing a few GO terms, e.g. 'promoter binding' and 'dimer formation'. However, a large number of TFs that do not share any GO term were also detected to co-occur.

For a broader view of the TF clusters, the cluster density was positively relevant to the gene density but not to the copy number of the genomic sequence. The expression level of genes was also higher when their promoters are close to (<2 kb) clusters than those that are not adjacent to any cluster (**Figure 4.4A**). The cluster size (number of TF bound to the cluster, thereafter) at the gene promoter was also positively correlated to the gene expression level (**Figure 4.4B**). We could even predict the cell type specific expression based on the cluster size in the promoter (**Figure 4.4C**).

The sequence in the center of cluster was also more conserved, especially for the clusters with more than 20 TF binding (defined as the large cluster, thereafter; **Figure 4.4D**). We also observed the enrichment of some enhancer or promoter markers within the cluster regions: mediator subunits MED1 and MED12 were enriched in clusters, histone H3K4me1, H3K4me3, H3K27ac, H3.3 and H2A.Z were also enriched in the flanking regions of the clusters (**Figure 4.4E**).

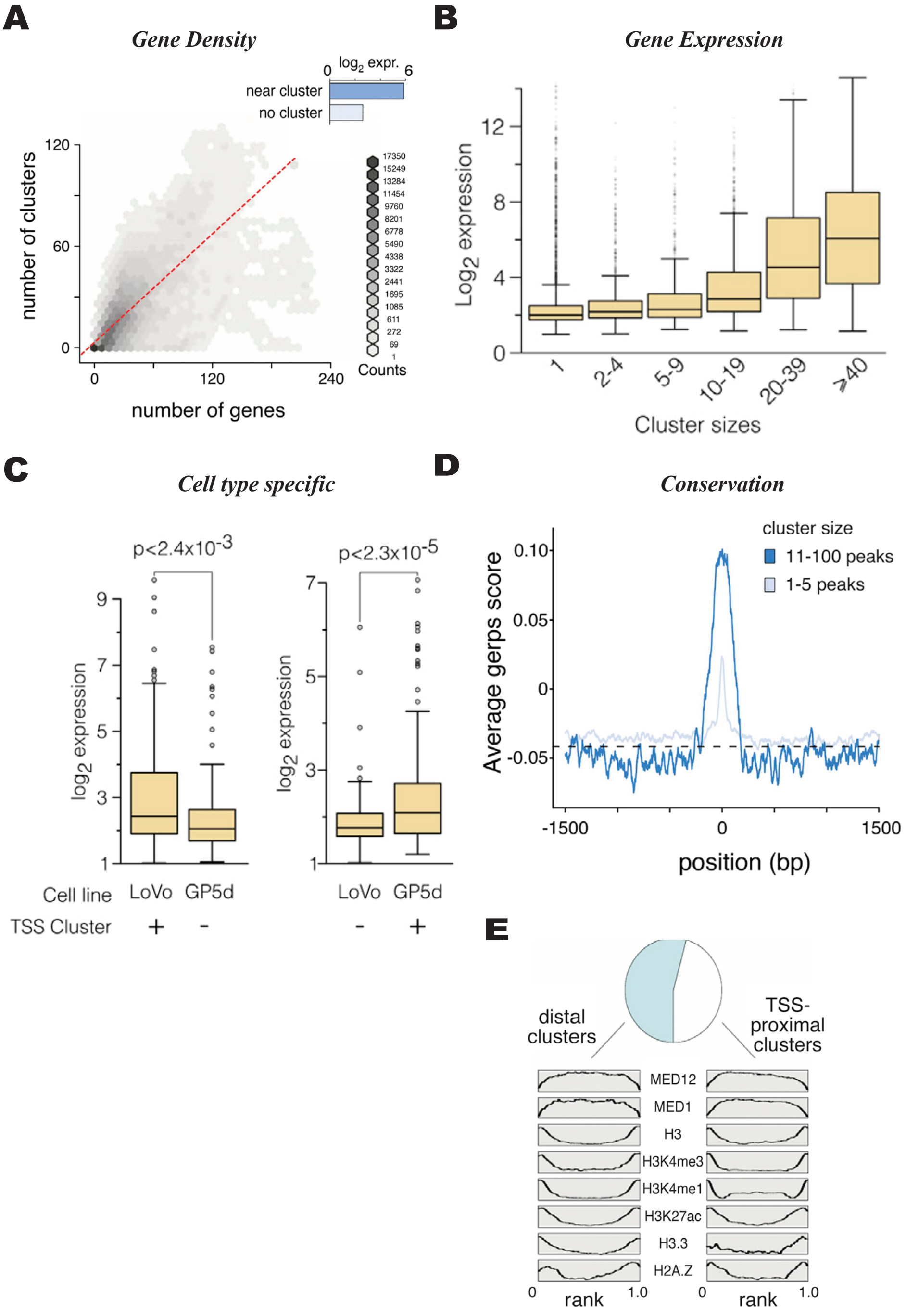

Figure 4.4 TF cluster and its characteristics

(A) Hexagonal bin density plot shows the number of clusters as a function of number of genes in 3 Mb genomic regions. Dashed red line indicates least-squares fit; slope is 0.5. Inset: expression level of genes with or without a cluster at their promoters (<2 kb of TSS).

(B) Cluster size at TSS predicts gene expression level (y axis). Boxes indicate the middle quartiles, separated by median line. Whiskers indicate last values within 1.5 times the interquartile range for the box.

(C) Clusters predict cell-type-specific gene expression. Boxplot of expression values for genes where one cell line has no promoter cluster (_), whereas the other has a large promoter cluster (+). Data from top 100 genes ranked by promoter cluster size, where the other cell-line has no promoter cluster are shown. Boxes indicate the middle quartiles separated by median line. Whiskers indicate last values within 1.5 times the interquartile range for the box.

(D) Conservation of sequences (average genomic evolutionary rate profiling [GERP] scores from 17 mammalian species) calculated for TF clusters. Note that regions containing TF clusters (blue) are more conserved than regions having between one and five TF peaks (light blue). The dashed line shows the average conservation score (GERP) for the flanking region.

(E) Localization of mediator subunits MED12 and MED1, histone H3, promoter (H3K4me3) and enhancer (H3K4me1, H3K27Ac) chromatin marks, and the variant histones H2A.Z and H3.3 within clusters (peak positions are rank normalized within each cluster) located more (left) or less (right) than 2 kb upstream from a TSS. Pie chart (top) shows fraction of clusters in each class.

4.2.2 TF motifs are enriched in TF clusters.

In order to attest the TF clusters and identify the TFs that shaped the TF clusters, we performed the *de novo* motif discovery assay for the promoter distal clusters. The motifs are heavily enriched for different classes of TF families: CTCF, other C2H2 zinc finger proteins, ETS, bZIP, nuclear receptor, GATA, homeodomain, NFI and forkhead factors. We also discovered a novel motif with the sequence C(A/T)G frequently enriched in the clusters. The function of the motif and factors bound to it are still unclear.

Since the *de novo* motif discovery could only detect the motifs with strong enrichment, we also checked the enrichment of the motifs that were identified in the HT-SELEX. A large fraction of these HT-SELEX motifs were found enriched in the TF clusters, 84 out of 239 motifs with the p-value lower than 0.01. Besides the motifs disclosed by *de novo* method, the motifs for NRF1, HINFP, TFAP2 and GLI-like C2H2 zinc finger proteins GLIS2 and ZIC1 were also significantly enriched.

As the motifs were heavily enriched in the TF clusters, we also tried to predict the TF clusters based on the enrichment of TF binding motifs obtained from HT-SELEX. A relatively low stringency with 1 site per 10 kb resulted in the highest accuracy of cluster prediction. With these criteria, we could only predict about 34% of the TF clusters. It indicated that weak binding of TFs is responsible for a substantial fraction of TF binding within the clusters.

4.2.3 Function of cohesin in TF clusters

The network analysis for TF pairs revealed that only a single connected network exist for all TFs tested in our ChIP-seq experiments. No major subnetwork was observed; instead one or more cohesin subunits were connected to all TFs in the network except two (E2F8 and DLX1).

Cohesin was found bound to almost all large clusters. Cohesin proximal DNA was more sensitive to DNase I and depleted of nucleosome. When we compared cohesin bound sites around the TF clusters and outside TF clusters, we found that both regions were hypersensitive to DNase I and depleted of nucleosomes. The only difference we discovered was that the TF clusters were enriched in TF binding motifs while the cohesin binding sites outside TF clusters were depleted of TF binding motifs. This could explain the reason why TF did not bind to the cohesin proximal DNA although it is accessible for protein binding. To more directly establish that cohesin binding could serve as a causative factor to increase the DNA accessibility and facilitate TF binding, we carried out siRNA to knock down cohesin subunit RAD21. After RAD21 depletion, genes with large clusters at their promoters tended to be down regulated. Both the DNase I hypersensitivity and nucleosome depletion decreased alongside RAD21 knock down. To exclude the global non-specific effect of RAD21 depletion for DNase I hypersensitivity (DHS), we compared the normalized DNase I cut number between cohesin proximal CTCF sites and cohesin distal REST sites. The DHS of cohesin proximal regions significantly decreased when RAD21 level was lowered but the difference of DHS in cohesin distal regions could not be distinguished. With regard to TF binding, cohesin loss resulted in decreased binding for TFs closely bound to cohesin sites but did not affect for the TFs generally bound to DNA farther from cohesin binding.

As we know that cohesin functions on facilitating TF binding, we tried to predict the TF binding with cohesin binding sites. Actually, the cohesin predictor worked very well, much better than any other predictor existing that includes prediction of TF binding sites with close proximity to DNase I hypersensitivity, FAIRE and H3K27ac regions (**Figure 4.5**).

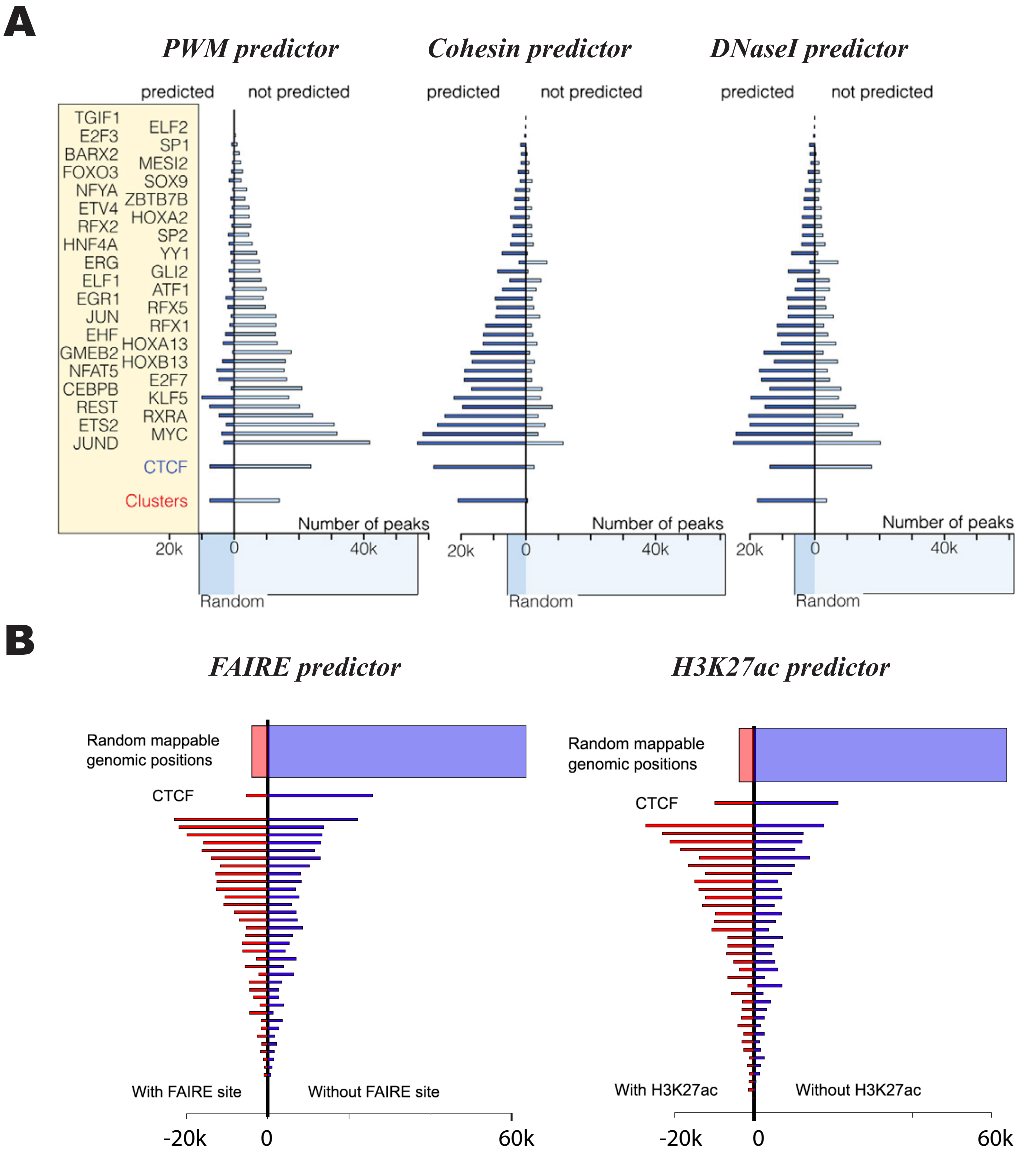

Figure 4.5 Different TF binding predictors

(A) Prediction of TF peaks, CTCF peaks and large TF clusters (Clusters) using a specific PWM model for each TF (left), proximity to cohesin (middle) or DNase I hypersensitive site (right). Peaks or clusters whose summits are in the predicted regions are in dark blue on the left side of the vertical lines. Peaks or clusters that could not be predicted are in light blue on the right side. Total length of each bar indicates the total number of peaks or clusters for the indicated experiment. Note that cohesin is the most sensitive predictor for individual TFs, CTCF, and TF clusters. Thick bars below the x-axes indicate false positive rate (prediction of random genomic regions).

(B) Prediction of TF peaks using proximity to FAIRE regions, or ChIP-seq peaks for lysine 27 acetylated histone H3 (H3K27Ac). Peaks whose summits are in the predicted regions are in red on the left side of the vertical lines, and peaks that are not located in the predicted regions are in blue (right side of vertical lines). Width of the bars is proportional to number of peaks for each TF. Random genomic positions (thick bar) and predictions for CTCF are shown for comparison. Order of the TF bars is the same as shown on (A). Note that for both FAIRE regions and acetylated H3K27 peaks the sensitivity is only approximately 50%.

4.2.4 TF clusters in cell cycle

Cohesin was initially discovered for its function in sister chromatid cohesion[128] and later found to have a role in regulating gene expression through bridging enhancer and promoter contact[60]. The two distinct roles actually indicated that cohesin might function on the inheritance of DNA accessibility and TF binding sites during cell cycle.

To establish such claim, cohesin should have the following characters: 1) cohesin proximal DNA is accessible; 2) cohesin could facilitate TF binding, both of which we already tested true for; 3) cohesin position remain constant through out the cell cyle; and 4) cohesin holds the sister chromatids at the TF cluster positions to mark the locations for both strands.

To test the cohesin position alongside the cell cycle, we synchronized the LoVo cells in G1/S by double thymidine block, and released the cells for 2, 4, 6, and 8 h. By checking the cohesin binding position with ChIP-seq in different time points, we discovered that cohesin bound to DNA constantly while the other TFs lost their binding in mitosis (**Figure 4.6A**, **B**). The loss of TF binding was not because of the depletion of DNA accessibility. Consistent with the constant binding of cohesin, the DNA remained hypersensitive to DNase I during the cell cycle (**Figure 4.6C**). The TF might be evicted by the strength from the chromosome condensation[250].

To test whether cohesin holds two sister chromatids at the TF cluster regions, we developed a 3C based method, Sister Chromosome Proximity Ligation. The ligation products of the sister strand ligation near the cohesin and TF clusters were significantly enriched in S and M phase, while the products far from cohesin sites remained the same. These results indicated that the cohesin held the sister chromosome at the TF cluster position, which could assist the inheritance of the TF cluster position to the newly synthesized chromosome.

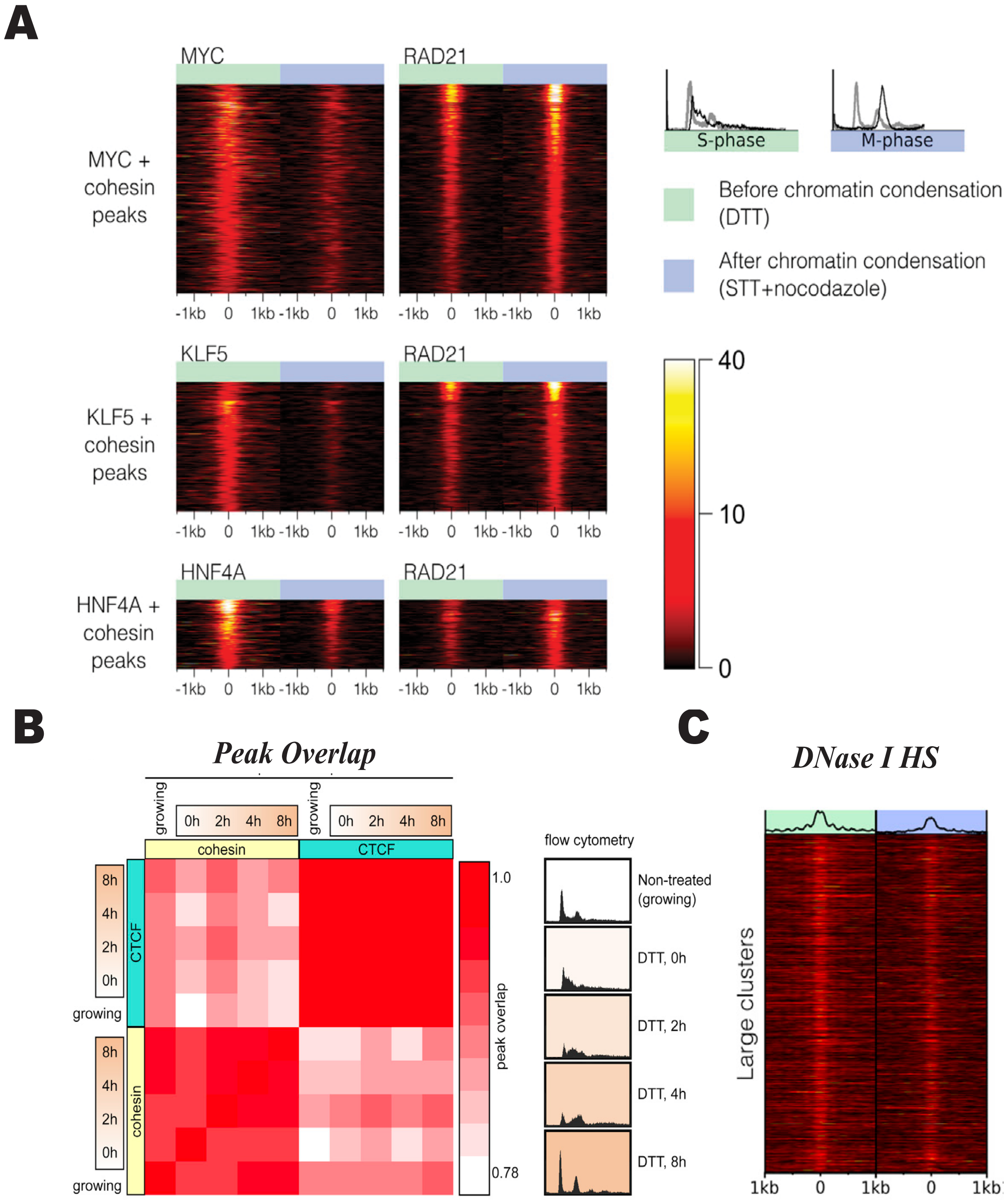

Figure 4.6 Cohesin remains bound to chromatin during cell cycle

(A) TFs (left), are bound to chromatin in S phase (green) but cleared from DNA in early M-phase-arrested cells (blue). Cohesin (RAD21) however, remains bound to chromatin also in M phase (blue). Heatmaps show sites where peaks for both a TF and RAD21 are found within 1 kb, sorted by maximum intensity. Color scale indicates peak height, coordinates are relative to the RAD21 peak. Top right: DNA content flow cytometry profiles for control (gray) and arrested (black) cell cultures.

(B) Cohesin and CTCF peak positions remain constant throughout cell division. LoVo cells were arrested in early S phase by double thymidine block (DTT). Cells were harvested 0, 2, 4 and 8 hr after DTT release and subjected to ChIP-seq analysis for CTCF and the cohesin subunit RAD21 (left), as well as flow cytometry analysis (right). The heatmap shows the fraction of pairwise overlap of the top 5,000 peaks from one condition to the top 10,000 peaks of the other. Note that the overlap is maintained throughout all time points analyzed, and that although there is considerable overlap between cohesin and CTCF, this overlap is not increased in Mphase (4 and 8 h), suggesting that both cohesin associated with CTCF and cohesin that is not associated with CTCF remain bound to DNA.

(C) Cohesin-associated sites remain accessible in M phase arrested cells. No significant change is observed in DNase I hypersensitivity, which appears to be centered at cohesin sites (right). Heatmaps are row-normalized, and sorted according to interhistone distance in S phase.

4.3 STUDY III (*IN VIVO* STUDY): MICE LACKING A MYC ENHANCER THAT INCLUDE HUMAN SNP RS6983267 ARE RESISTANT TO INTESTINAL TUMORS.

4.3.1 Generation of the Myc335-null mouse

As is well established, Myc is one of the key transcription regulators in intestinal epithelial proliferation. In the previous study, the software called 'Enhancer Element Locator' (EEL) was developed to search for enhancer elements by comparing the conserved binding sites of different TFs among different species[251]. EEL was used to detect the potential enhancers at *MYC* loci and discovered an enhancer element located 335 kb upstream of *MYC* gene transcription starting site and highly conserved between human and mouse. The enhancer element, designated Myc-335, harbors several high confidential CRC related SNPs[51,52,179,180]. The physical contact between the Myc-335 and *MYC* gene body was also verified in cultured cells[183,252,253]. Later in our Study II, we found that tens of TFs bound to this enhancer element including the TFs that were involved in gut tissue development such as TCF7L2, HNF4A, TFAP2A, HOXA13 and MYC (**Figure 4.7A**).

Although the direct effect of the SNP on c-Myc expression is inconclusive[51,254], it provides a plausible mechanism on the action of the SNP and the enhancer element. The lack of prominent effect of the single nucleotide polymorphism on Myc expression could be because it only alters the binding affinity of one single TF TCF7L2. In order to evaluate the role of the SNP and the enhancer function, we generated a mouse model lacking the whole cluster region Myc-335.

The mice carrying the conditional knock out allele were generated by flanking the 1740 bp Myc-335 cluster with two LoxP sites (mice strain designated: Myc-335 cKO). Then the Myc-335 cKO mice were crossed to the deletor strain EIIa-Cre that expressed Cre-recombinase to obtain the Myc-335 null mice (**Figure 4.7B**). The genotype was determined by PCR analysis using genomic DNA as templates.

4.3.2 Myc transcription in the Myc335-null mouse

The Myc-335 null mice (Myc335-/-, thereafter) were viable and fertile, and did not show any overt phenotype in normal conditions. Such non-prominent was distinct with the phenotype of Myc gene knock out mice; it is embryonic lethal, and even with heterozygous loss of Myc the newborn mice are abnormally small in body size, due to the placental dysfunction[255,256].

As the Myc-335 harbors SNP rs6983267 that is related to CRC, we analyzed the morphology of the intestinal tissues. The histological examination of the Myc335-/- at the postnatal day 1 (p1) exhibited comparable phenotype with their wild type siblings. Then we tested the proliferation of the Myc335-/- intestines with immunohistochemistry analysis of the proliferative cell marker Ki-67 and could not distinguish any difference with their littermates. Furthermore, immunohistochemistry analysis of Myc in the Myc335-/- intestinal crypts also detected the same normal expression pattern as wild type mice (**Figure 4.7C**).

When we performed the similar analyses in Myc335-/- adult mice, they all showed the same phenotypes as their contemporary siblings. The loss of Myc-335 did not have any major impact on the intestinal differentiation, which was consistent with the lack of effect of Myc loss on intestinal epithelial homeostasis[257-259]. The result indicated that Myc-335 was dispensable for normal function of mouse intestinal tissues under the standard condition.

Despite that the mRNA level of c-Myc was indistinguishable between Myc335-/- and wild-type p1 duodenum, a significant decrease of c-Myc RNA could be detected in Myc335-/- colon with both quantitative PCR (qPCR) and exon microarray analysis (Figure 4.7D). The ChIP-seq experiment of Tcf7l2 in mice colon revealed that the highest peak around Myc TSS at Myc-335 was lost due to the deletion. We did not observe any compensatory binding of Tcf7l2 within 1 Mb of Myc gene, which indicated that Myc-335 included a major binding site for Tcf7l2 and its loss caused a moderate decrease of Myc transcription.

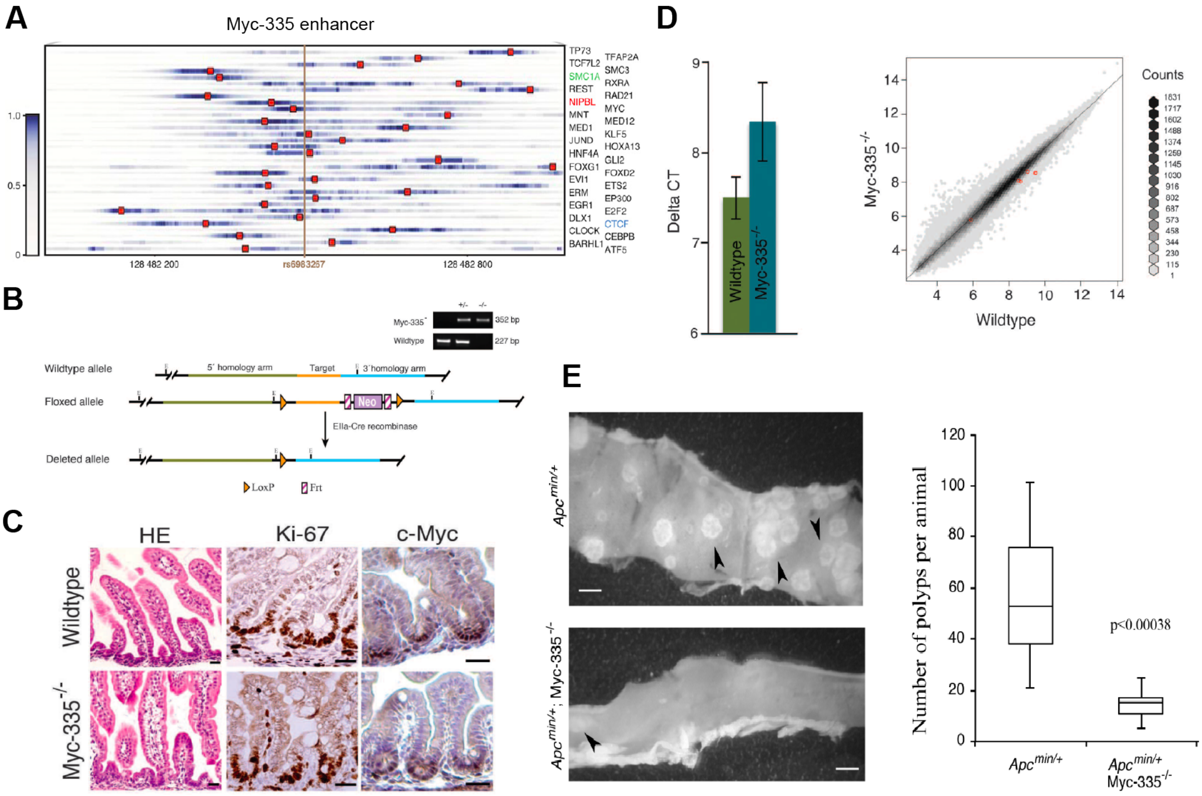

Figure 4.7 Myc-335 function in mice

- **(A)** ChIP-seq shows that many TFs bind to Myc-335 element in LoVo. Vertical blue lines show binding sites of such TF and red box indicate the highest binding summit within Myc-335. The brown vertical line marks the CRC related SNP rs6983267.
- **(B)** Generation of Myc-335 cKO and Myc-335 null mice. Inset shows the genotyping result.
- **(C)** Immunohistology of Myc-335 null mice compared with wildtype mice at p1, showing that intestinal tissues from Myc-335 null mice develop similar morphology (HE), proliferation (Ki-67) and Myc expression (c-Myc) as those from the wildtype mice. Scale bar shows 10 µm.
- **(D)** Myc expression tested by qPCR (left) and exon array (right) shows that Myc-335 null mice express moderately lower amount of c-Myc in their intestines than the wildtype mice.
- **(E)** Reduced incidence of polyp formation indicated by arrows in Apcmin/+; Myc-335−/− mice at 4 months of age (left). The box plot (right) shows total number of polyps in both small intestines and colon per mouse (n=9 for both APCmin and APCmin; Myc-335-/-). Student t-test.

4.3.3 Tumorigenesis

Althought Myc-335 only displayed a moderate effect on Myc transcriptional regulation and was dispensable for normal intestinal function, it could still have significant impact on intestinal tumorigenesis as Myc is central for cell proliferation and harbors a conserved confidential CRC risk SNP rs6983267. APCmin mouse strain is a well established model for FAP that develops 50 to 100 tumors in the intestinal tissues at the age of 2-4 months depending on Myc[260-262] and TCF/LEF activity[263-265].

To test the effect of Myc-335 on tumorigenesis, we crossed the Myc-335 null strain with APCmin strain. We scored the polyps in small intestine and colon of the mice at the age of 4 months. We found that there were a significantly smaller number of polyps in APCmin; Myc335-/- intestines than the control APCmin (n=9; p<0.00038; **Figure 4.7E**).

5 DISCUSSION

5.1 IMPROVE THE MODEL OF TF-DNA BINDING SPECIFICITIES

The transcription factors in cells shape the tissue specific gene expression and determine the cell identity and cell fate[156,266-271]. To establish where and how TFs bind to the genome requires the knowledge of the TF-DNA binding in different levels including the biochemical affinity, genomic context for TF-DNA binding and protein-protein interaction and synergy. In order to first obtain the biochemical affinity of TF binding to DNA, we performed HT-SELEX for the vast majority of human and mouse TFs and acquired 830 binding profiles to describe 239 distinctly different binding specificities for over 400 TFs, describing the largest collection of mammalian TF DNA binding specificities. Considering the fact that proteins with the similar structure bind to similar DNA sequence, our dataset provides a precious resource for the majority of human TF binding specificities which will be widely applied to various researches on TFs such as TF binding *in silico* prediction, the functional study of disease related noncoding SNPs, TF gene evolution and so on.

Knowing the biochemical affinity of the TF to different DNA sequence could help to model where and how TF binds to the genome[251,272]. Despite the central importance of transcriptional regulation in different biological processes such as development, very little effort has been made to study the binding specificities of human TFs. Most of the existing databases had focused on the DNA binding specificities of TFs in lower and less complex model organisms, like yeast, *C.elegans* and *Drosophila*[273-275]. And the study on mammalian TF binding specificities are only concentrated on the DNA binding domain, with very few cases of analysis for full-length TF proteins[24,230]. The comprehensive study of human TF-DNA binding specificities especially the full-length TFs could improve our understanding of the landscape of transcriptional regulation network in human cells. This knowledge is very important for personal medicine as it can be used to predict the gene expression level affected by the mutation and variants in the genome[51,52].

Before our study, the largest collection of existing mammalian TF-DNA binding specificities were performed with protein binding microarray (PBM)[24,230]. PBM uses all combinations of 8-mer sequences for TF binding selection due to the technical basis of microarray. It provides the unbiased environment for all the DNA sequences for protein binding that HT-SELEX might lack since HT-SELEX uses DNA fragments that are amplified by PCR. The over amplification will quickly saturate the binding of TFs with the sequences that have high affinity to the TFs. However, PBM could only test the binding of TFs to 8-mer DNA sequences. We found in our study that most TFs bound to DNA sequences that were longer than 10 bp, especially for thoese TFs that could bind to DNA forming dimers. By comparing our HT-SELEX dataset with PBM datasets, we find that HT-SELEX models slightly more TF-DNA binding specificities for TFs which tend to bind to DNA as monomers. However, HT-SELEX dramatically increases the number of models for TFs that bind to DNA as dimers or multimers.

In HT-SELEX, we successfully modeled TF binding specificities for 79 TFs with both full length proteins and DBDs alone. By comparing the 79 pairs of binding specificities, we discovered only one case that had different binding specificities. ELK1 full length protein might bind to DNA with both monomeric site and homodimeric site, while ELK1 DBD could only bind to monomeric site. Based on the HT-SELEX data, we could claim that DBD alone determined the binding specificities of the TF.

Some previous study also pointed out that position weight matrix (PWM) might not be good enough to describe the binding specificities of TF-DNA binding as PWM was based on the hypothesis that each DNA position was independent in TF binding[24]. In our study, we found this hypothesis was generally true except for some exceptional cases: 1) adjacent base-stacking affected TF-DNA binding independency; 2) Posterior homeodomain TFs tended to bind only some specific k-mer sequences preceding 'TAAAA' sequence; 3) DNA structure affected TF-DNA binding as some TFs preferred to bind to DNA sequences with a stretch of 'A' or 'T' flanking the core motif. This information could help to better understand TF-DNA binding and improve the model for TF binding prediction.

5.2 MODEL OF THE INHERITANCE OF TF CLUSTERS

Although more and more functional information about the genome has been discovered, we still lack of enough knowledge to fully understand the transcriptional regulation network. Lots of efforts have been put world-wide to add up such information but they are not carried out in a systematic way making it more difficult to compare or incorporate the data. ENCODE consortium has performed the high throughput and more systematic analyses to query the TF binding sites, protein accessible regions and chromatin topology of the human genome in different tissue types. However, the largest coverage in a single cell type only accounts for a small fraction (50 TFs in K562 cell) of TFs, which is far from explaining the entire system of transcription in such cell. More comprehensive study needs to be introduced to the same cell type for better understanding the individual role and also the cooperation of regulatory elements in transcriptional regulation.

After obtaining the TF-DNA binding specificities, we launched an *in vivo* study using ChIP-seq to determine the binding sites for the majority of highly expressed TFs in CRC cell line LoVo. By putting together the ChIP-seq peaks, we found most TFs bound to a very limited fraction of the genome, forming highly dense clusters. TFs in the clusters are not functionally related. Strikingly, 75% of the TF peaks are localized within only 0.8% of the human genome.

Virtually all TF clusters contain cohesin-binding sites. By using siRNA to knock down cohesin, we demonstrated that cohesin causatively increased the adjacent DNA accessibility and hence facilitate TF binding. Furthermore, we found that cohesin binding position remained constant during cell cycle when TFs were cleared off chromosomes twice when DNA replication fork passed their binding sites and when chromosome got condensed in prometaphase. We also showed that cohesin held two sister chromosomes together at the TF cluster position in S and M phases, providing a mechanism for cohein function in the inheritance of TF clusters through cell cycle. Thus we proposed a model that cohesin binding played a role in cellular memory that promoted the fast recovery of TF clusters after DNA replication and chromatin condensation.

Based on all the facts about cohesin function on TF clusters: 1) cohesin binding excludes nucleosomes; 2) cohesin increases proximal DNA accessibility for proteins; 3) cohesin binding without TF cluster formation results from depletion of TF motifs; 4) cohesin facilitates TF binding; 5) cohesin positions remain constant across cell cycle while TFs are evicted; 6) cohesin holds sister chromosomes at the TF cluster sites, we proposed a model to describe the TF binding inheritance in cell cycle. In this model, TFs are cleared off chromatin by DNA polymerase while cohesin remains bound to both chromatids in S phase; cohesin could then reload TFs back to chromatin at its positions; TFs will be evicted again when chromosome gets condensed in early M phase while cohesin still binds at the same position; TFs would then in turn bind to cohesin sites after removal of cohesin from its positions in late M phase.

The epigenetic bookmark model could help to answer the genome-wide TF searching issue during cell cycle: it is estimated that a single TF needs to spend more than 1 minute searching the bacteria E.coli genome to find its binding sites[276]. Mammalian genome is about 1000 times larger than E.coli and encodes for many more TFs for the sophisticated regulation. And during cell cycle, all TFs will be kicked off the chromosomal DNA twice, respectively in S and M phases. Some cell types divide very fast, duplicating within a day. Such a rapid turnover of bound TFs on chromosome would challenge cell for time and energy that it has to spend on maintaining the TF binding patterns and hence transcription regulation network.

There must be some epigenetic marks that are stable during the two periods to bookmark the TF binding positions. The 2012 Nobel laureate John B. Gurdon found that histone variant H3.3 could serve as a mitotic bookmarking to remember the transcriptional regulatory landscape in mitosis[277]. DNA methylation was also proposed to maintain the heterochromatin region and works as a negative mitotic bookmark[278]. Other sequence specific transcription factors were also discovered as candidates for the cellular memory of TF binding in mitosis, such as GATA1[279] and TBP-PP2A complex[280].

In this study, cohesin could increase the DNA accessibility and remain bound to chromatin in mitosis. It thus been attested as another epigenetic factor working for the mitotic bookmarking. It is even more robust as it can remain bound not only in mitosis but also in S phase when all DNA binding proteins would be cleared off by the replication fork. The mechanism of how transcriptional regulation is maintained in S or M phase is not clear. Cohesin, serving as a transcription regulatory element bookmark, could thus help us better understand it.

Cohesin and mediator were also found involved in bridging the enhancer-promoter looping[60,62,281]. Cohesin is a ring shape protein complex and proposed to mediate the looping by encircling two chromatin strands[282]. However, it is still not clear whether one cohesin ring encircles both strands or two cohesin rings each traps one individual chromatin strand. When we deleted one cohesin binding TF cluster in mouse, we could only detect the cohesin-binding change at the very deleted site but nowhere else within 1 million nucleotides region. If only one cohesin molecule is involved in cis-looping, we should have also observed the cohesin-binding change in neighbouring region. Consistently in Bacterial, the stoichiometry study of Bacterial SMC homologue MukBEF using single-molecule millisecond multicolor fluorescence microscopy in live cell suggested that dimer is the minimal unit for functional cohesin complex[283]. The result suggested that there might be two or multiple cohesin molecules synergizing to mediate the enhancer-promoter *cis*-looping.

As had been well established before, cohesin functioned in the sister chromatid cohesion by encircling two sister chromosomes[284]. And in this study, we demonstrated that cohesin held the sister chromosomes at the TF cluster positions.

5.3 MYC-335 FUNCTION

In order to test the TF cluster function, we targeted and deleted one TF cluster that harbored a high confidential CRC related SNP 335 kb upstream of Myc gene. The Myc-335 null mice are indistinguishable from wild type animals under standard conditions. However, when they were crossed to the FAP mouse model APCmin strain mice, the Myc-335-/-, APCmin mice had moderately attenuated Myc expression than APCmin mice and developed significantly fewer polyps in small intestines. This result suggested that Myc-335 TF cluster was a tumor specific enhancer for Myc and dispensable for normal intestinal function.

This study is one of the pioneer studies that use a mouse model to validate the function of an enhancer element. The enhancer element Myc-335 has been previously identified by our early study[51] and the 1 Mb region around Myc-335 enhancer element also harbors several SNPs identified by genome-wide association studies (GWAS) linked to different diseases including CRC[157,179,285,286]. How these SNPs contribute to the CRC risk remains very promiscuous. The study added the evidence that these SNPs could be functional via Myc function. However, we also found that the enhancers were tumor specific and disposable for normal intestinal function. However the Myc-335 enhancer is highly conserved and positively selected during evolution, indicating the importance of its function. In that Myc-335 enhancer is required for tumor growth, it may contribute to the fast proliferation of intestinal tissues, as colon and intestine are among the tissues that go through fast self-renewal or it might be critical during tissue repairing process.

It would be interesting to test its function in the intestinal stem cell self-renewal. To specifically knock out the Myc-335 enhancer in intestinal stem cell, we could cross the mouse strain with two Lox-P sites sandwiching Myc-335 cluster with Lgr5-Cre knockin strain. The Cre recombinase is specifically expressed in Lgr5+ cells and Myc-335 will consequently be knocked out in the same cell. Since Cre-LoxP recombination efficiency is lower than absolute, the offsprings will bear mosaic genotypes in their intestinal stem cells. If Myc-335 plays a role in stem cell self-renewal, the proliferation will be imbalanced between wild type and Myc-335 knock out stem cells and the later stage embryo should bear more wild-type Lgr5+ cells than early embryos that start to express Lgr5.

6 CONCLUSIONS AND PROSPECTS

The entire studies comprehensively advance the knowledge of TF-DNA binding and its function: from the cell-free biochemical binding affinity, to the binding mechanism in the context of nucleosome-rich mammalian CRC genome, to how the binding is inherited during cell proliferation, and finally the validation of the function of TF binding in CRC animal model.

As we already found that TFs in cells worked in a combinatory fashion, namely a group of TFs bound to DNA next to each other within a region for several hundred base pairs long. One way of setting up the collaboration is through protein-protein interaction in which one factor can bind to DNA first and recruit the other. The most famous example is the transcription cofactor which itself could not bind directly to DNA but to other TFs. The TF binding alone does not drive the target gene expression, but with transcription co-activator the complex could boost the transcription[287].

Another way of TF collaboration depends on the pioneer factor. Pioneer factors are TFs that bind to the closed chromatin and open up the region to concomitantly facilitate other factors recruitment[288]. Several pioneer factors were reported with forkhead DNA binding domain which is similar to the DBD of histone H1. Such property could ease the pioneer factor to bind to the closed chromatin pre-occupied by nucleosomes[289]. After pioneer factor binding, the nucleosome is locally rearranged and more accessible for other TFs. The well established pioneer factors include TFs of forkhead factors, GATA, ETS, TFAP2, Goucho-related (Gro/TLE/Grg), etc[290-292].

The third type of TF collaboration is through the DNA mediated dimerization. In cells, there are often multiple TF binding sites in close proximity in a segment of DNA region. Some TF binding could increase other TF binding while some may exclude others from binding to the same region. TF can bind to DNA in homodimer formed with two identical TFs or heterodimer formed between two distinct TFs to cooperate the transcriptional regulation[293]. Some TF families are very common to form homodimers, such as ETS[27] and T-box[18]. The dimer binding sites originally derive from the two individual TF binding sites but evolve depending on the topology of the protein complex. However, the knowledge of dimer binding especially heterodimer binding is still lacking, which makes the prediction for some TF binding miserable.

HT-SELEX is successfully set up as a powerful tool to determine the long TF binding sites that are very prominent for TF dimer binding specificties. The binding specificities of TF homodimer have mostly been profiled in this study, whereas those of the TF heterodimer are still missing. As the cDNA clones or proteins of most TFs are available, we can tandemly purify the protein interaction complex of two TFs with different tagging and apply them to HT-SEXLEX to systematically profile the binding specificities of TF heterodimers. This should greatly increase our knowledge of the TF cooperation effect in the transcriptional regulation.

Another more important question arises: when we annotate the genome with regulatory elements, what are the targets of them? A large fraction of the regulatory elements are not immediately next to but physically contact the target genes via looping. Understanding the TF binding affinity and genome-wide TF binding sites does not necessarily tell how they work to regulate the transcription. Recently, chromosome confirmation capture (3C) and the derived technology Hi-C have been developed to investigate the topology of the chromatin looping. For further study, one could study the genome-wide physical interactions between regulatory elements and gene bodies in LoVo or GP5d, to more comprehensively understand the transcriptional regulation network in colorectal cancer cells.

To test the function of a broader range of TF clusters in cell or mouse model, it is feasible now with the development of genome editing tools, such as Zinc Finger Nuclease (ZFN), transcription activator-like effectors nuclease (TALEN) and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR/Cas)[294,295]. CRISPR/Cas system has been lately applied to genome editing[296,297]. Since it is less intricate and powerful especially for high-efficient multiple targeting[298], it immediately became popular and widely applied. With CRISPR/Cas system, we could specifically knock out the TF clusters in cell and test their functions in a systematic manner. The new editing tool also provides the possibility to study the function of TF cluster in animal models.

All such knowledge could greatly help to understand the regulatory code (also called the second genetic code) of human cells especially of cancer cells, which is very important for suggesting the potential targets for personal cancer therapies.

7 ACKNOWLEDGEMENTS

The whole thesis work was conducted in the laboratory of Prof. Jussi Taipale in the Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge campus, Sweden.

Above all, I'd like to express my greatest and deepest gratitude and respect to my supervisor **Prof. Jussi Taipale** from the bottom of my heart. You opened the window for my scientific views and introduced me to the research field of transcription and systems biology.

How time flies! I've been in Jussi's research group for over five years since June 2008 in Biomedicum Helsinki, Finland. I still remember the moment when I came for the job interview. I was almost like a piece of white paper in this field in the beginning. And till today under his supervision and direction, I managed to finish several projects and published the results in high profiled journals. I acquired not only the scientific thoughts from Jussi, but have also been affected by his enthusiasm in science. Jussi is always very busy but he never neglects the advice for me, from the study plan development, the meeting presentation rehearsal, to the future career plan. I also feel so lucky in Jussi's lab as to be able to attend many international academic meetings to set up communication with other young scientists and share my own work. None of my achievements could have been reached without you, Jussi.

My great gratitude will also be delivered to my other supervisor **Dr. Minna Taipale**, Jussi's wife. You are like a babysitter for my PhD life, helping me out of all kinds of troubles. Even before I started my work in Helsinki and Stockholm, Minna helped me arrange the flights and find my accommodation. To start my work, I could not have done all the paper work that smoothly without Minna's help. And the help never stops: suggesting an immense number of invaluable comments for all presentations and manuscripts, discussion and exchanging scientific ideas, help with experiment set up and trouble shooting, and tackling all kinds of issues in sequencing. Minna, without you, I could imagine a much more troublesome PhD life with no doubt.

I am so grateful to my three main scientific collaborators in Jussi's lab, **Arttu Jolma**, **Drs. Martin Enge** and **Inderpreet Sur**. I enjoy so much working with you, discussing the scientific problems and also sharing various experiences in life. Without your contribution, the works could have never been that beautiful.

Emma Inns, no words could express my gratitude to you. I am aware that you have worked lots of extra hours because of me. You helped so much on the flight booking, hotel arrangement, seminar room booking, language polishing, letter preparation, party arrangement, ... I apologize for the limited space here to acknowledge all that you have done for me.

My respect and thanks will be dedicated to my previous colleague in Helsinki Drs. **Gong-Hong Wei**, **Song-Ping Li** and **Fuping Zhang**. Thank you very much for your help with my experiment and sharing the life experience in the foreign country. Also for Gong-Hong, congratulations to your great success as a young principal investigator leading your own research group. You set a great example for me.

Monica Ahlberg, tack så mycket for your help with all the tedious documents of my PhD registration, yearly follow-up, half-time control and dissertation to make my PhD offcially happen.

Sini Miettinen, **Ritva Nurmi**, **Anna Zetterlund**, **Dr. Lijuan Hu**, and **Sandra Augsten**, thank you very much for all technical assistance in my experiments to make them succeeded.

My colleagues in Stockholm, **Drs. Jianping Liu**, **Kazuhiro Nitta**, **Kashyap Dave**, **Ekaterina Morgunova**, **Bernhard Schmierer**, **Yimeng Yin**, **Åsa Kolterud**, **Anders Eriksson** and our new PhD student **Bei Wei**, we have had so much good time together in this fantastic lab and the amazing Scandinavian capital. I express my deep gratitude. I feel so lucky working with so many awesome persons. Without you, my life and work would have become much more boring.

My colleagues in Helsinki, **Mikko Turunen**, **Drs. Teemu Kivioja**, **Anna Vähärautio**, **Maria Sokolova**, thank you for helping my work and life in Helsinki. It was impressive. I am still missing Helsinki, the first European city where I lived and the lab where I started my study in genetics and transcription.

My previous colleagues, **Lin Feng**, **Ji Zhang**, **Thirupathi Pattipaka**, **Drs. Thomas Whitington**, **Outi Hallikas**, **Martin Bonke**, and **James Thompson**, I am so glad to have spent some time with you and get help from you.

Outside science and work, I made so many awesome friends in Helsinki and Stockholm. In Helsinki, **Yajing Gao**, **Yan Yan**, **Yizhou Hu**, **Ping Chen, You Zou** and **Drs. Kui Qian**, **Shentong Fang**, **Robert M. Badeau**, thank you so much for making my life in Helsinki exciting. In Stockholm, Drs. **Chenglin Wu**, **Ting Jia**, **Xiaohui Jia**, **Xiaoyan Huang** and also **Tianwei Gu**, **Erik Weifeng Lin**, **Sai Lu**, **Jeremy Iehl**, **Xiaozhen Li**, **Jia Sun**, **Bin Li**, **Meng Li**, **Xiaoyan Liu**, **Xin Wang**, **Min Wang**, **Wei Xiao**, **Chengjun Sun**, **Sichao Li**, **Jian Sun**, **Luming Ye**, **Alexander van Tilburg, Dequan Ning**, **Peter de Hval**, we have had so much fun time together and I could still remember almost every moment that I have spent with you, badminton, swimming, fika, gym, shopping, party, food, travel, etc. Thank you a lot for making my life interesting in Stockholm. And I also thank those that are not listed here due to the limited space.

My Mum and Dad, thank you so much for always patiently listening to me regardless of what it was, standing with me and comforting my heart. I am so proud of being your son. Meanwhile I would also express my deepest apologies that I spent so limited time with you every year during my PhD study. I appreciate your understanding and forgiveness. I love you!

Financially, the thesis work has been mainly supported by European Union Framework Project 7, SYSCOL led by Prof. Jussi Taipale.

Sincerely, Jian Yan

Huddinge,

December, 2013

8 REFERENCES

1 Visvader, J. E. Cells of origin in cancer. *Nature* **469**, 314-322, doi:10.1038/nature09781 (2011).

2 Mitchell, P. J. & Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245**, 371-378 (1989).

3 Ptashne, M. & Gann, A. Transcriptional activation by recruitment. *Nature* **386**, 569-577, doi:10.1038/386569a0 (1997).

4 Warnmark, A., Treuter, E., Wright, A. P. & Gustafsson, J. A. Activation functions 1 and 2 of nuclear receptors: molecular strategies for transcriptional activation. *Molecular endocrinology* **17**, 1901-1909, doi:10.1210/me.2002-0384 (2003).

5 Orphanides, G., Lagrange, T. & Reinberg, D. The general transcription factors of RNA polymerase II. *Genes & development* **10**, 2657-2683 (1996).

6 Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics* **10**, 252-263, doi:10.1038/nrg2538 (2009).

7 Stegmaier, P., Kel, A. E. & Wingender, E. Systematic DNA-binding domain classification of transcription factors. *Genome informatics. International Conference on Genome Informatics* **15**, 276-286 (2004).

8 Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research* **34**, D108-110, doi:10.1093/nar/gkj143 (2006).

9 Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic acids research*, doi:10.1093/nar/gkt997 (2013).

10 Wolfe, S. A., Nekludova, L. & Pabo, C. O. DNA recognition by Cys2His2 zinc finger proteins. *Annual review of biophysics and biomolecular structure* **29**, 183-212, doi:10.1146/annurev.biophys.29.1.183 (2000).

11 Garvie, C. W. & Wolberger, C. Recognition of specific DNA sequences. *Molecular cell* **8**, 937-946 (2001).

12 Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic acids research* **29**, 2860-2874 (2001).

13 Rohs, R. *et al.* The role of DNA shape in protein-DNA recognition. *Nature* **461**, 1248-1253, doi:10.1038/nature08473 (2009).

14 Stefflova, K. *et al.* Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell* **154**, 530-540, doi:10.1016/j.cell.2013.07.007 (2013).

15 Heinz, S. *et al.* Effect of natural genetic variation on enhancer selection and function. *Nature* **503**, 487-492, doi:10.1038/nature12615 (2013).

16 De Val, S. *et al.* Combinatorial regulation of endothelial gene expression by ets and forkhead transcription factors. *Cell* **135**, 1053-1064, doi:10.1016/j.cell.2008.10.049 (2008).

17 Siebenlist, U. Nucleotide sequence of the three major early promoters of bacteriophage T7. *Nucleic acids research* **6**, 1895-1907 (1979).

18 Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327-339, doi:10.1016/j.cell.2012.12.009 (2013).

19 Stormo, G. D. & Fields, D. S. Specificity, free energy and information content in protein-DNA interactions. *Trends in biochemical sciences* **23**, 109-113 (1998).

20 Zhao, Y., Ruan, S., Pandey, M. & Stormo, G. D. Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics* **191**, 781-790, doi:10.1534/genetics.112.138685 (2012).

21 Mathelier, A. & Wasserman, W. W. The next generation of transcription factor binding site prediction. *PLoS computational biology* **9**, e1003214, doi:10.1371/journal.pcbi.1003214 (2013).

22 Garner, M. M. & Revzin, A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic acids research* **9**, 3047-3060 (1981).

23 Fried, M. & Crothers, D. M. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic acids research* **9**, 6505-6525 (1981).

24 Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720-1723, doi:10.1126/science.1162327 (2009).

25 Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology* **24**, 1429-1435, doi:10.1038/nbt1246 (2006).

26 Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome research* **20**, 861-873, doi:10.1101/gr.100552.109 (2010).

27 Wei, G. H. *et al.* Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *The EMBO journal* **29**, 2147-2160, doi:10.1038/emboj.2010.106 (2010).

28 Maerkl, S. J. & Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233-237, doi:10.1126/science.1131007 (2007).

29 Meng, X., Brodsky, M. H. & Wolfe, S. A. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nature biotechnology* **23**, 988-994, doi:10.1038/nbt1120 (2005).

30 Berger, M. F. & Bulyk, M. L. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature protocols* **4**, 393-411, doi:10.1038/nprot.2008.195 (2009).

31 Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

32 A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology* **9**, e1001046, doi:10.1371/journal.pbio.1001046 (2011).

33 Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108, doi:10.1038/nature11233 (2012).

34 Doolittle, W. F. Is junk DNA bunk? A critique of ENCODE. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 5294-5300, doi:10.1073/pnas.1221376110 (2013).

35 Graur, D. *et al.* On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome biology and evolution* **5**, 578-590, doi:10.1093/gbe/evt028 (2013).

36 Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annual review of genomics and human genetics* **7**, 29-59, doi:10.1146/annurev.genom.7.080505.115623 (2006).

37 Zwir, I., Latifi, T., Perez, J. C., Huang, H. & Groisman, E. A. The promoter architectural landscape of the Salmonella PhoP regulon. *Molecular microbiology* **84**, 463-485, doi:10.1111/j.1365-2958.2012.08036.x (2012).

38 Estrem, S. T. *et al.* Bacterial promoter architecture: subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase alpha subunit. *Genes & development* **13**, 2134-2147 (1999).

39 Smale, S. T. & Kadonaga, J. T. The RNA polymerase II core promoter. *Annual review of biochemistry* **72**, 449-479, doi:10.1146/annurev.biochem.72.121801.161520 (2003).

40 Vannini, A. & Cramer, P. Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Molecular cell* **45**, 439-446, doi:10.1016/j.molcel.2012.01.023 (2012).

41 Yang, C., Bolotin, E., Jiang, T., Sladek, F. M. & Martinez, E. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene* **389**, 52-65, doi:10.1016/j.gene.2006.09.029 (2007).

42 Anish, R., Hossain, M. B., Jacobson, R. H. & Takada, S. Characterization of transcription from TATA-less promoters: identification of a new core promoter element XCPE2 and analysis of factor requirements. *PloS one* **4**, e5103, doi:10.1371/journal.pone.0005103 (2009).

43 Ong, C. T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature reviews. Genetics* **12**, 283-293, doi:10.1038/nrg2957 (2011).

44 Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-112, doi:10.1038/nature07829 (2009).

45 Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82, doi:10.1038/nature11232 (2012).

46 Levine, M. Transcriptional enhancers in animal development and evolution. *Current biology : CB* **20**, R754-763, doi:10.1016/j.cub.2010.06.070 (2010).

47 Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-1117, doi:10.1016/j.cell.2008.04.043 (2008).

48 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).

49 Carter, A. J. & Wagner, G. P. Evolution of functionally conserved enhancers can be accelerated in large populations: a population-genetic model. *Proceedings. Biological sciences / The Royal Society* **269**, 953-960, doi:10.1098/rspb.2002.1968 (2002).

50 Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947, doi:10.1016/j.cell.2013.09.053 (2013).

51 Tuupanen, S. *et al.* The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nature genetics* **41**, 885-890, doi:10.1038/ng.406 (2009).

52 Tuupanen, S. *et al.* Characterization of the colorectal cancer-associated enhancer MYC-335 at 8q24: the role of rs67491583. *Cancer genetics* **205**, 25-33, doi:10.1016/j.cancergen.2012.01.005 (2012).

53 Heinz, S. *et al.* Effect of natural genetic variation on enhancer selection and function. *Nature*, doi:10.1038/nature12615 (2013).

54 Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837, doi:10.1016/j.cell.2007.05.009 (2007).

55 Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics* **39**, 311-318, doi:10.1038/ng1966 (2007).

56 Jin, C. *et al.* H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nature genetics* **41**, 941-945, doi:10.1038/ng.409 (2009).

57 Lai, F. *et al.* Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* **494**, 497-501, doi:10.1038/nature11884 (2013).

58 Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187, doi:10.1038/nature09033 (2010).

59 Natoli, G. & Andrau, J. C. Noncoding transcription at enhancers: general principles and functional models. *Annual review of genetics* **46**, 1-19, doi:10.1146/annurev-genet-110711-155459 (2012).

60 Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430-435, doi:10.1038/nature09380 (2010).

61 Handoko, L. *et al.* CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature genetics* **43**, 630-638, doi:10.1038/ng.857 (2011).

62 Seitan, V. C. *et al.* Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome research*, doi:10.1101/gr.161620.113 (2013).

63 Orom, U. A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58, doi:10.1016/j.cell.2010.09.001 (2010).

64 Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-1311, doi:10.1126/science.1067799 (2002).

65 Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics* **38**, 1348-1354, doi:10.1038/ng1896 (2006).

66 Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature genetics* **38**, 1341-1347, doi:10.1038/ng1891 (2006).

67 Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research* **16**, 1299-1309, doi:10.1101/gr.5571506 (2006).

68 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

69 Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58-64, doi:10.1038/nature08497 (2009).

70 Fanucchi, S., Shibayama, Y., Burd, S., Weinberg, M. S. & Mhlanga, M. M. Chromosomal Contact Permits Transcription between Coregulated Genes. *Cell* **155**, 606-620, doi:10.1016/j.cell.2013.09.051 (2013).

71 Burgess-Beusse, B. *et al.* The insulation of genes from external enhancers and silencing chromatin. *Proceedings of the National Academy of Sciences of the United States of America* **99 Suppl 4**, 16433-16437, doi:10.1073/pnas.162342499 (2002).

72 Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**, 1194-1211, doi:10.1016/j.cell.2009.06.001 (2009).

73 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

74 Ogbourne, S. & Antalis, T. M. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *The Biochemical journal* **331 (Pt 1)**, 1-14 (1998).

75 Bode, J. *et al.* Biological significance of unwinding capability of nuclear matrix-associating DNAs. *Science* **255**, 195-197 (1992).

76 Kim, M. K., Lesoon-Wood, L. A., Weintraub, B. D. & Chung, J. H. A soluble transcription factor, Oct-1, is also found in the insoluble nuclear matrix and possesses silencing activity in its alanine-rich domain. *Molecular and cellular biology* **16**, 4366-4377 (1996).

77 Olguin, P., Oteiza, P., Gamboa, E., Gomez-Skarmeta, J. L. & Kukuljan, M. RE-1 silencer of transcription/neural restrictive silencer factor modulates ectodermal patterning during Xenopus development. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **26**, 2820-2829, doi:10.1523/JNEUROSCI.5037-05.2006 (2006).

78 Haecker, S. A., Muramatsu, T., Sensenbaugh, K. R. & Sanders, M. M. Repression of the ovalbumin gene involves multiple negative elements including a ubiquitous transcriptional silencer. *Molecular endocrinology* **9**, 1113-1126 (1995).

79 Yeo, M. *et al.* Small CTD phosphatases function in silencing neuronal gene expression. *Science* **307**, 596-600, doi:10.1126/science.1100801 (2005).

80 Loven, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320-334, doi:10.1016/j.cell.2013.03.036 (2013).

81 Su, A. I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 4465-4470, doi:10.1073/pnas.012025199 (2002).

82 Zheng, Y., Thomas, P. M. & Kelleher, N. L. Measurement of acetylation turnover at distinct lysines in human histones identifies long-lived acetylation sites. *Nature communications* **4**, 2203, doi:10.1038/ncomms3203 (2013).

83 Waterborg, J. H. Dynamics of histone acetylation in vivo. A function for acetylation turnover? *Biochemistry and cell biology = Biochimie et biologie cellulaire* **80**, 363-378 (2002).

84 Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *Journal of molecular biology* **196**, 261-282 (1987).

85 Hinoue, T. *et al.* Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome research* **22**, 271-282, doi:10.1101/gr.117523.110 (2012).

86 Hon, G. C. *et al.* Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nature genetics* **45**, 1198-1206, doi:10.1038/ng.2746 (2013).

87 Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477-481, doi:10.1038/nature12433 (2013).

88 Xie, W. *et al.* Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148**, 816-831, doi:10.1016/j.cell.2011.12.035 (2012).

89 Wang, Z. *et al.* Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* **138**, 1019-1031, doi:10.1016/j.cell.2009.06.049 (2009).

90 Ropero, S. & Esteller, M. The role of histone deacetylases (HDACs) in human cancer. *Molecular oncology* **1**, 19-25, doi:10.1016/j.molonc.2007.01.001 (2007).

91 Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858, doi:10.1038/nature07730 (2009).

92 Zhang, B. *et al.* A dynamic H3K27ac signature identifies VEGFA-stimulated endothelial enhancers and requires EP300 activity. *Genome research* **23**, 917-927, doi:10.1101/gr.149674.112 (2013).

93 Cotney, J. *et al.* Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. *Genome research* **22**, 1069-1080, doi:10.1101/gr.129817.111 (2012).

94 Szerlong, H. J., Prenni, J. E., Nyborg, J. K. & Hansen, J. C. Activator-dependent p300 acetylation of chromatin in vitro: enhancement of transcription by disruption of repressive nucleosome-nucleosome interactions. *The Journal of biological chemistry* **285**, 31954-31964, doi:10.1074/jbc.M110.148718 (2010).

95 Graff, J. & Tsai, L. H. Histone acetylation: molecular mnemonics on the chromatin. *Nature reviews. Neuroscience* **14**, 97-111, doi:10.1038/nrn3427 (2013).

96 Ghisletti, S. *et al.* Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity* **32**, 317-328, doi:10.1016/j.immuni.2010.02.008 (2010).

97 Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 21931-21936, doi:10.1073/pnas.1016071107 (2010).

98 Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283, doi:10.1038/nature09692 (2011).

99 Briggs, S. D. *et al.* Histone H3 lysine 4 methylation is mediated by Set1 and required for cell growth and rDNA silencing in Saccharomyces cerevisiae. *Genes & development* **15**, 3286-3295, doi:10.1101/gad.940201 (2001).

100 Kim, J. *et al.* RAD6-Mediated transcription-coupled H2B ubiquitylation directly stimulates H3K4 methylation in human cells. *Cell* **137**, 459-471, doi:10.1016/j.cell.2009.02.027 (2009).

101 Wysocka, J. *et al.* WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. *Cell* **121**, 859-872, doi:10.1016/j.cell.2005.03.036 (2005).

102 Shilatifard, A. Molecular implementation and physiological roles for histone H3 lysine 4 (H3K4) methylation. *Current opinion in cell biology* **20**, 341-348, doi:10.1016/j.ceb.2008.03.019 (2008).

103 Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116-120, doi:10.1038/35065132 (2001).

104 Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120-124, doi:10.1038/35065138 (2001).

105 Karimi, M. M. *et al.* DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell stem cell* **8**, 676-687, doi:10.1016/j.stem.2011.04.004 (2011).

106 Wei, Y., Mizzen, C. A., Cook, R. G., Gorovsky, M. A. & Allis, C. D. Phosphorylation of histone H3 at serine 10 is correlated with chromosome condensation during mitosis and meiosis in Tetrahymena. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 7480-7484 (1998).

107 Sauve, D. M., Anderson, H. J., Ray, J. M., James, W. M. & Roberge, M. Phosphorylation-induced rearrangement of the histone H3 NH2-terminal domain during mitotic chromosome condensation. *The Journal of cell biology* **145**, 225-235 (1999).

108 Wen, W. *et al.* MST1 promotes apoptosis through phosphorylation of histone H2AX. *The Journal of biological chemistry* **285**, 39108-39116, doi:10.1074/jbc.M110.151753 (2010).

109 Downs, J. A., Lowndes, N. F. & Jackson, S. P. A role for Saccharomyces cerevisiae histone H2A in DNA repair. *Nature* **408**, 1001-1004, doi:10.1038/35050000 (2000).

110 Shroff, R. *et al.* Distribution and dynamics of chromatin modification induced by a defined DNA double-strand break. *Current biology : CB* **14**, 1703-1711, doi:10.1016/j.cub.2004.09.047 (2004).

111 Chowdhury, D. *et al.* A PP4-phosphatase complex dephosphorylates gamma-H2AX generated during DNA replication. *Molecular cell* **31**, 33-46, doi:10.1016/j.molcel.2008.05.016 (2008).

112 Douglas, P. *et al.* Protein phosphatase 6 interacts with the DNA-dependent protein kinase catalytic subunit and dephosphorylates gamma-H2AX. *Molecular and cellular biology* **30**, 1368-1381, doi:10.1128/MCB.00741-09 (2010).

113 Chowdhury, D. *et al.* gamma-H2AX dephosphorylation by protein phosphatase 2A facilitates DNA double-strand break repair. *Molecular cell* **20**, 801-809, doi:10.1016/j.molcel.2005.10.003 (2005).

114 Lo, W. S. *et al.* Phosphorylation of serine 10 in histone H3 is functionally linked in vitro and in vivo to Gcn5-mediated acetylation at lysine 14. *Molecular cell* **5**, 917-926 (2000).

115 Clements, A. *et al.* Structural basis for histone and phosphohistone binding by the GCN5 histone acetyltransferase. *Molecular cell* **12**, 461-473 (2003).

116 Zhong, S., Goto, H., Inagaki, M. & Dong, Z. Phosphorylation at serine 28 and acetylation at lysine 9 of histone H3 induced by trichostatin A. *Oncogene* **22**, 5291-5297, doi:10.1038/sj.onc.1206507 (2003).

117 Mizuguchi, G. *et al.* ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* **303**, 343-348, doi:10.1126/science.1090701 (2004).

118 Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887-898, doi:10.1016/j.cell.2008.02.022 (2008).

119 Henikoff, S., Henikoff, J. G., Sakai, A., Loeb, G. B. & Ahmad, K. Genome-wide profiling of salt fractions maps physical properties of chromatin. *Genome research* **19**, 460-469, doi:10.1101/gr.087619.108 (2009).

120 Du, J. *et al.* Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. *Cell* **151**, 167-180, doi:10.1016/j.cell.2012.07.034 (2012).

121 Fischle, W. *et al.* Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. *Nature* **438**, 1116-1122, doi:10.1038/nature04219 (2005).

122 Vermeulen, M. *et al.* Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell* **142**, 967-980, doi:10.1016/j.cell.2010.08.020 (2010).

123 Rosenfeld, J. A. *et al.* Determination of enriched histone modifications in non-genic portions of the human genome. *BMC genomics* **10**, 143, doi:10.1186/1471-2164-10-143 (2009).

124 Benevolenskaya, E. V. Histone H3K4 demethylases are essential in development and differentiation. *Biochemistry and cell biology = Biochimie et biologie cellulaire* **85**, 435-443, doi:10.1139/O07-057 (2007).

125 Steger, D. J. *et al.* DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells. *Molecular and cellular biology* **28**, 2825-2839, doi:10.1128/MCB.02076-07 (2008).

126 Koch, C. M. *et al.* The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome research* **17**, 691-707, doi:10.1101/gr.5704207 (2007).

127 Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41-45, doi:10.1038/47412 (2000).

128 Nasmyth, K. Cohesin: a catenase with separate entry and exit gates? *Nature cell biology* **13**, 1170-1177, doi:10.1038/ncb2349 (2011).

129 Horsfield, J. A., Print, C. G. & Monnich, M. Diverse developmental disorders from the one ring: distinct molecular pathways underlie the cohesinopathies. *Frontiers in genetics* **3**, 171, doi:10.3389/fgene.2012.00171 (2012).

130 Schleiffer, A. *et al.* Kleisins: a superfamily of bacterial and eukaryotic SMC protein partners. *Molecular cell* **11**, 571-575 (2003).

131 Tedeschi, A. *et al.* Wapl is an essential regulator of chromatin structure and chromosome segregation. *Nature* **501**, 564-568, doi:10.1038/nature12471 (2013).

132 Gerlich, D., Koch, B., Dupeux, F., Peters, J. M. & Ellenberg, J. Live-cell imaging reveals a stable cohesin-chromatin interaction after but not before DNA replication. *Current biology : CB* **16**, 1571-1578, doi:10.1016/j.cub.2006.06.068 (2006).

133 Rolef Ben-Shahar, T. *et al.* Eco1-dependent cohesin acetylation during establishment of sister chromatid cohesion. *Science* **321**, 563-566, doi:10.1126/science.1157774 (2008).

134 Chan, K. L. *et al.* Cohesin's DNA exit gate is distinct from its entrance gate and is regulated by acetylation. *Cell* **150**, 961-974, doi:10.1016/j.cell.2012.07.028 (2012).

135 Remeseiro, S. & Losada, A. Cohesin, a chromatin engagement ring. *Current opinion in cell biology* **25**, 63-71, doi:10.1016/j.ceb.2012.10.013 (2013).

136 Deardorff, M. A. *et al.* HDAC8 mutations in Cornelia de Lange syndrome affect the cohesin acetylation cycle. *Nature* **489**, 313-317, doi:10.1038/nature11316 (2012).

137 Zhang, N. *et al.* A handcuff model for the cohesin complex. *The Journal of cell biology* **183**, 1019-1031, doi:10.1083/jcb.200801157 (2008).

138 Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281-1295, doi:10.1016/j.cell.2013.04.053 (2013).

139 Sofueva, S. *et al.* Cohesin-mediated interactions organize chromosomal domain architecture. *The EMBO journal*, doi:10.1038/emboj.2013.237 (2013).

140 Hou, C., Dale, R. & Dean, A. Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 3651-3656, doi:10.1073/pnas.0912087107 (2010).

141 Guillou, E. *et al.* Cohesin organizes chromatin loops at DNA replication factories. *Genes & development* **24**, 2812-2822, doi:10.1101/gad.608210 (2010).

142 Li, Y. *et al.* Characterization of constitutive CTCF/cohesin loci: a possible role in establishing topological domains in mammalian genomes. *BMC genomics* **14**, 553, doi:10.1186/1471-2164-14-553 (2013).

143 Remeseiro, S., Cuadrado, A., Gomez-Lopez, G., Pisano, D. G. & Losada, A. A unique role of cohesin-SA1 in gene regulation and development. *The EMBO journal* **31**, 2090-2102, doi:10.1038/emboj.2012.60 (2012).

144 Vega, H. *et al.* Roberts syndrome is caused by mutations in ESCO2, a human homolog of yeast ECO1 that is essential for the establishment of sister chromatid cohesion. *Nature genetics* **37**, 468-470, doi:10.1038/ng1548 (2005).

145 Solomon, D. A. *et al.* Mutational inactivation of STAG2 causes aneuploidy in human cancer. *Science* **333**, 1039-1043, doi:10.1126/science.1203619 (2011).

146 Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264-278, doi:10.1016/j.cell.2012.06.023 (2012).

147 Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).

148 Mantovani, A. Cancer: Inflaming metastasis. *Nature* **457**, 36-37, doi:10.1038/457036b (2009).

149 Kolonel, L. N., Altshuler, D. & Henderson, B. E. The multiethnic cohort study: exploring genes, lifestyle and cancer risk. *Nature reviews. Cancer* **4**, 519-527, doi:10.1038/nrc1389 (2004).

150 Wakeford, R. The cancer epidemiology of radiation. *Oncogene* **23**, 6404-6428, doi:10.1038/sj.onc.1207896 (2004).

151 Piazuelo, M. B., Epplein, M. & Correa, P. Gastric cancer: an infectious disease. *Infectious disease clinics of North America* **24**, 853-869, vii, doi:10.1016/j.idc.2010.07.010 (2010).

152 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

153 Boland, C. R. & Goel, A. Microsatellite instability in colorectal cancer. *Gastroenterology* **138**, 2073-2087 e2073, doi:10.1053/j.gastro.2009.12.064 (2010).

154 Vilar, E. & Gruber, S. B. Microsatellite instability in colorectal cancer-the stable evidence. *Nature reviews. Clinical oncology* **7**, 153-162, doi:10.1038/nrclinonc.2009.237 (2010).

155 Jass, J. R. HNPCC and sporadic MSI-H colorectal cancer: a review of the morphological similarities and differences. *Familial cancer* **3**, 93-100, doi:10.1023/B:FAME.0000039849.86008.b7 (2004).

156 Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676, doi:10.1016/j.cell.2006.07.024 (2006).

157 Zanke, B. W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature genetics* **39**, 989-994, doi:10.1038/ng2089 (2007).

158 Berns, A. Cancer genetics. Is p53 the only real tumor suppressor gene? *Current biology : CB* **4**, 137-139 (1994).

159 Ahmadian, M. *et al.* PPARgamma signaling and metabolism: the good, the bad and the future. *Nature medicine* **19**, 557-566, doi:10.1038/nm.3159 (2013).

160 Humphries, A. & Wright, N. A. Colonic crypt organization and tumorigenesis. *Nature reviews. Cancer* **8**, 415-424, doi:10.1038/nrc2392 (2008).

161 Potten, C. S., Kellett, M., Roberts, S. A., Rew, D. A. & Wilson, G. D. Measurement of in vivo proliferation in human colorectal mucosa using bromodeoxyuridine. *Gut* **33**, 71-78 (1992).

162 Taylor, R. W. *et al.* Mitochondrial DNA mutations in human colonic crypt stem cells. *The Journal of clinical investigation* **112**, 1351-1360, doi:10.1172/JCI19435 (2003).

163 Novelli, M. R. *et al.* Polyclonal origin of colonic adenomas in an XO/XY patient with FAP. *Science* **272**, 1187-1190 (1996).

164 Fodde, R. & Brabletz, T. Wnt/beta-catenin signaling in cancer stemness and malignant behavior. *Current opinion in cell biology* **19**, 150-158, doi:10.1016/j.ceb.2007.02.007 (2007).

165 He, X. A Wnt-Wnt situation. *Developmental cell* **4**, 791-797 (2003).

166 Korinek, V. *et al.* Depletion of epithelial stem-cell compartments in the small intestine of mice lacking Tcf-4. *Nature genetics* **19**, 379-383, doi:10.1038/1270 (1998).

167 Pinto, D., Gregorieff, A., Begthel, H. & Clevers, H. Canonical Wnt signals are essential for homeostasis of the intestinal epithelium. *Genes & development* **17**, 1709-1713, doi:10.1101/gad.267103 (2003).

168 Kuhnert, F. *et al.* Essential requirement for Wnt signaling in proliferation of adult small intestine and colon revealed by adenoviral expression of Dickkopf-1. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 266-271, doi:10.1073/pnas.2536800100 (2004).

169 Groden, J. *et al.* Identification and characterization of the familial adenomatous polyposis coli gene. *Cell* **66**, 589-600 (1991).

170 Joslyn, G. *et al.* Identification of deletion mutations and three new genes at the familial polyposis locus. *Cell* **66**, 601-613 (1991).

171 Su, L. K., Vogelstein, B. & Kinzler, K. W. Association of the APC tumor suppressor protein with catenins. *Science* **262**, 1734-1737 (1993).

172 Mitin, N., Rossman, K. L. & Der, C. J. Signaling interplay in Ras superfamily function. *Current biology : CB* **15**, R563-574, doi:10.1016/j.cub.2005.07.010 (2005).

173 Kosinski, C. *et al.* Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 15418-15423, doi:10.1073/pnas.0707210104 (2007).

174 Haramis, A. P. *et al.* De novo crypt formation and juvenile polyposis on BMP inhibition in mouse intestine. *Science* **303**, 1684-1686, doi:10.1126/science.1093587 (2004).

175 Rajagopalan, H., Nowak, M. A., Vogelstein, B. & Lengauer, C. The significance of unstable chromosomes in colorectal cancer. *Nature reviews. Cancer* **3**, 695-701, doi:10.1038/nrc1165 (2003).

176 Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759-767 (1990).

177 Broderick, P. *et al.* A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nature genetics* **39**, 1315-1317, doi:10.1038/ng.2007.18 (2007).

178 Jaeger, E. *et al.* Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nature genetics* **40**, 26-28, doi:10.1038/ng.2007.41 (2008).

179 Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature genetics* **39**, 984-988, doi:10.1038/ng2085 (2007).

180 Houlston, R. S. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature genetics* **40**, 1426-1435, doi:10.1038/ng.262 (2008).

181 Tomlinson, I. P. *et al.* A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nature genetics* **40**, 623-630, doi:10.1038/ng.111 (2008).

182 Tenesa, A. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nature genetics* **40**, 631-637, doi:10.1038/ng.133 (2008).

183 Pomerantz, M. M. *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nature genetics* **41**, 882-884, doi:10.1038/ng.403 (2009).

184 Toyota, M. *et al.* CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 8681-8686 (1999).

185 Weisenberger, D. J. *et al.* DNA methylation analysis by digital bisulfite genomic sequencing and digital MethyLight. *Nucleic acids research* **36**, 4689-4698, doi:10.1093/nar/gkn455 (2008).

186 Ogino, S., Kawasaki, T., Kirkner, G. J., Loda, M. & Fuchs, C. S. CpG island methylator phenotype-low (CIMP-low) in colorectal cancer: possible associations with male sex and KRAS mutations. *The Journal of molecular diagnostics : JMD* **8**, 582-588, doi:10.2353/jmoldx.2006.060082 (2006).

187 Shen, L. *et al.* Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 18654-18659, doi:10.1073/pnas.0704652104 (2007).

188 Laird, P. W. & Jaenisch, R. The role of DNA methylation in cancer genetic and epigenetics. *Annual review of genetics* **30**, 441-464, doi:10.1146/annurev.genet.30.1.441 (1996).

189 Issa, J. P. *et al.* Increased cytosine DNA-methyltransferase activity during colon cancer progression. *Journal of the National Cancer Institute* **85**, 1235-1240 (1993).

190 Schmidt, W. M. *et al.* Progressive up-regulation of genes encoding DNA methyltransferases in the colorectal adenoma-carcinoma sequence. *Molecular carcinogenesis* **46**, 766-772, doi:10.1002/mc.20307 (2007).

191 Ibrahim, A. E. *et al.* Sequential DNA methylation changes are associated with DNMT3B overexpression in colorectal neoplastic progression. *Gut* **60**, 499-508, doi:10.1136/gut.2010.223602 (2011).

192 Lao, V. V. & Grady, W. M. Epigenetics and colorectal cancer. *Nature reviews. Gastroenterology & hepatology* **8**, 686-700, doi:10.1038/nrgastro.2011.173 (2011).

193 Blackwood, E. M. & Eisenman, R. N. Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. *Science* **251**, 1211-1217 (1991).

194 Luscher, B. Function and regulation of the transcription factors of the Myc/Max/Mad network. *Gene* **277**, 1-14 (2001).

195 Martinato, F., Cesaroni, M., Amati, B. & Guccione, E. Analysis of Myc-induced histone modifications on target chromatin. *PloS one* **3**, e3650, doi:10.1371/journal.pone.0003650 (2008).

196 Ayer, D. E., Lawrence, Q. A. & Eisenman, R. N. Mad-Max transcriptional repression is mediated by ternary complex formation with mammalian homologs of yeast repressor Sin3. *Cell* **80**, 767-776 (1995).

197 Schreiber-Agus, N. *et al.* An amino-terminal domain of Mxi1 mediates anti-Myc oncogenic activity and interacts with a homolog of the yeast transcriptional repressor SIN3. *Cell* **80**, 777-786 (1995).

198 Dotto, G. P., Gilman, M. Z., Maruyama, M. & Weinberg, R. A. c-myc and c-fos expression in differentiating mouse primary keratinocytes. *The EMBO journal* **5**, 2853-2857 (1986).

199 Adhikary, S. & Eilers, M. Transcriptional regulation and transformation by Myc proteins. *Nature reviews. Molecular cell biology* **6**, 635-645, doi:10.1038/nrm1703 (2005).

200 Rosenwald, I. B., Rhoads, D. B., Callanan, L. D., Isselbacher, K. J. & Schmidt, E. V. Increased expression of eukaryotic translation initiation factors eIF-4E and eIF-2 alpha in response to growth induction by c-myc. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 6175-6178 (1993).

201 Jones, R. M. *et al.* An essential E box in the promoter of the gene encoding the mRNA cap-binding protein (eukaryotic initiation factor 4E) is a target for activation by c-myc. *Molecular and cellular biology* **16**, 4754-4764 (1996).

202 Coller, H. A. *et al.* Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 3260-3265 (2000).

203 Gomez-Roman, N., Grandori, C., Eisenman, R. N. & White, R. J. Direct activation of RNA polymerase III transcription by c-Myc. *Nature* **421**, 290-294, doi:10.1038/nature01327 (2003).

204 Berns, K., Hijmans, E. M. & Bernards, R. Repression of c-Myc responsive genes in cycling cells causes G1 arrest through reduction of cyclin E/CDK2 kinase activity. *Oncogene* **15**, 1347-1356, doi:10.1038/sj.onc.1201280 (1997).

205 Steiner, P. *et al.* Identification of a Myc-dependent step during the formation of active G1 cyclin-cdk complexes. *The EMBO journal* **14**, 4814-4826 (1995).

206 Staller, P. *et al.* Repression of p15INK4b expression by Myc through association with Miz-1. *Nature cell biology* **3**, 392-399, doi:10.1038/35070076 (2001).

207 Evan, G. I. *et al.* Induction of apoptosis in fibroblasts by c-myc protein. *Cell* **69**, 119-128 (1992).

208 Askew, D. S., Ashmun, R. A., Simmons, B. C. & Cleveland, J. L. Constitutive c-myc expression in an IL-3-dependent myeloid cell line suppresses cell cycle arrest and accelerates apoptosis. *Oncogene* **6**, 1915-1922 (1991).

209 Nass, S. J., Li, M., Amundadottir, L. T., Furth, P. A. & Dickson, R. B. Role for Bcl-xL in the regulation of apoptosis by EGF and TGF beta 1 in c-myc overexpressing mammary epithelial cells. *Biochemical and biophysical research communications* **227**, 248-256, doi:10.1006/bbrc.1996.1497 (1996).

210 Miyashita, T. & Reed, J. C. Tumor suppressor p53 is a direct transcriptional activator of the human bax gene. *Cell* **80**, 293-299 (1995).

211 Jamerson, M. H., Johnson, M. D. & Dickson, R. B. Dual regulation of proliferation and apoptosis: c-myc in bitransgenic murine mammary tumor models. *Oncogene* **19**, 1065-1071 (2000).

212 Pelengaris, S., Khan, M. & Evan, G. c-MYC: more than just a matter of life and death. *Nature reviews. Cancer* **2**, 764-776, doi:10.1038/nrc904 (2002).

213 Pelengaris, S., Littlewood, T., Khan, M., Elia, G. & Evan, G. Reversible activation of c-Myc in skin: induction of a complex neoplastic phenotype by a single oncogenic lesion. *Molecular cell* **3**, 565-577 (1999).

214 Lin, C. Y. *et al.* Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**, 56-67, doi:10.1016/j.cell.2012.08.026 (2012).

215 Shoemaker, A. R., Gould, K. A., Luongo, C., Moser, A. R. & Dove, W. F. Studies of neoplasia in the Min mouse. *Biochimica et biophysica acta* **1332**, F25-48 (1997).

216 Su, L. K. *et al.* Multiple intestinal neoplasia caused by a mutation in the murine homolog of the APC gene. *Science* **256**, 668-670 (1992).

217 Fodde, R. *et al.* A targeted chain-termination mutation in the mouse Apc gene results in multiple intestinal tumors. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 8969-8973 (1994).

218 Yang, K. *et al.* A mouse model of human familial adenomatous polyposis. *The Journal of experimental zoology* **277**, 245-254 (1997).

219 Marcus, D. M. *et al.* Retinal pigment epithelium abnormalities in mice with adenomatous polyposis coli gene disruption. *Archives of ophthalmology* **115**, 645-650 (1997).

220 Smits, R. *et al.* Apc1638T: a mouse model delineating critical domains of the adenomatous polyposis coli protein involved in tumorigenesis and development. *Genes & development* **13**, 1309-1321 (1999).

221 Oshima, M. *et al.* Loss of Apc heterozygosity and abnormal tissue building in nascent intestinal polyps in mice carrying a truncated Apc gene. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 4482-4486 (1995).

222 Shibata, H. *et al.* Rapid colorectal adenoma formation initiated by conditional targeting of the Apc gene. *Science* **278**, 120-123 (1997).

223 Hahn, S. A. *et al.* DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1. *Science* **271**, 350-353 (1996).

224 Di Cristofano, A., Pesce, B., Cordon-Cardo, C. & Pandolfi, P. P. Pten is essential for embryonic development and tumour suppression. *Nature genetics* **19**, 348-355, doi:10.1038/1235 (1998).

225 Chawengsaksophak, K., James, R., Hammond, V. E., Kontgen, F. & Beck, F. Homeosis and intestinal tumours in Cdx2 mutant mice. *Nature* **386**, 84-87, doi:10.1038/386084a0 (1997).

226 Boyd, M. *et al.* Genome-wide analysis of CDX2 binding in intestinal epithelial cells (Caco-2). *The Journal of biological chemistry* **285**, 25115-25125, doi:10.1074/jbc.M109.089516 (2010).

227 Ee, H. C., Erler, T., Bhathal, P. S., Young, G. P. & James, R. J. Cdx-2 homeodomain protein expression in human and rat colorectal adenoma and carcinoma. *The American journal of pathology* **147**, 586-592 (1995).

228 Butterworth, J. R. Another important function for an old friend! The role of iron in colorectal carcinogenesis. *Gut* **55**, 1384-1386, doi:10.1136/gut.2006.098350 (2006).

229 Vincentelli, R. *et al.* High-throughput protein expression screening and purification in Escherichia coli. *Methods* **55**, 65-72, doi:10.1016/j.ymeth.2011.08.010 (2011).

230 Berger, M. F. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266-1276, doi:10.1016/j.cell.2008.05.024 (2008).

231 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

232 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).

233 Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic acids research* **33**, e175, doi:10.1093/nar/gni179 (2005).

234 Zhao, Y., Granas, D. & Stormo, G. D. Inferring binding energies from selected binding sites. *PLoS computational biology* **5**, e1000590, doi:10.1371/journal.pcbi.1000590 (2009).

235 Crawford, G. E. *et al.* DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature methods* **3**, 503-509, doi:10.1038/nmeth888 (2006).

236 Whitfield, M. L. *et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular biology of the cell* **13**, 1977-2000, doi:10.1091/mbc.02-02-0030. (2002).

237 Tiwari, V. K. & Baylin, S. B. Combined 3C-ChIP-cloning (6C) assay: a tool to unravel protein-mediated genome architecture. *Cold Spring Harbor protocols* **2009**, pdb prot5168, doi:10.1101/pdb.prot5168 (2009).

238 Stros, M., Launholt, D. & Grasser, K. D. The HMG-box: a versatile protein domain occurring in a wide variety of DNA-binding proteins. *Cellular and molecular life sciences : CMLS* **64**, 2590-2606, doi:10.1007/s00018-007-7162-3 (2007).

239 Brown, R. S. Zinc finger proteins: getting a grip on RNA. *Current opinion in structural biology* **15**, 94-98, doi:10.1016/j.sbi.2005.01.006 (2005).

240 Brayer, K. J. & Segal, D. J. Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. *Cell biochemistry and biophysics* **50**, 111-131, doi:10.1007/s12013-008-9008-5 (2008).

241 Deppmann, C. D., Alvania, R. S. & Taparowsky, E. J. Cross-species annotation of basic leucine zipper factor interactions: Insight into the evolution of closed interaction networks. *Molecular biology and evolution* **23**, 1480-1492, doi:10.1093/molbev/msl022 (2006).

242 Liberg, D., Sigvardsson, M. & Akerblad, P. The EBF/Olf/Collier family of transcription factors: regulators of differentiation in cells originating from all three embryonal germ layers. *Molecular and cellular biology* **22**, 8389-8397 (2002).

243 Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic acids research* **33**, W116-120, doi:10.1093/nar/gki442 (2005).

244 Geggier, S. & Vologodskii, A. Sequence dependence of DNA bending rigidity. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 15421-15426, doi:10.1073/pnas.1004809107 (2010).

245 Zheng, G., Colasanti, A. V., Lu, X. J. & Olson, W. K. 3DNALandscapes: a database for exploring the conformational features of DNA. *Nucleic acids research* **38**, D267-274, doi:10.1093/nar/gkp959 (2010).

246 Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947-956, doi:10.1016/j.cell.2005.08.020 (2005).

247 Moorman, C. *et al.* Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 12027-12032, doi:10.1073/pnas.0605003103 (2006).

248 Roy, S. *et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**, 1787-1797, doi:10.1126/science.1198374 (2010).

249 Stanojevic, D., Small, S. & Levine, M. Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. *Science* **254**, 1385-1387 (1991).

250 Martin, R. M. & Cardoso, M. C. Chromatin condensation modulates access and binding of nuclear proteins. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **24**, 1066-1072, doi:10.1096/fj.08-128959 (2010).

251 Hallikas, O. *et al.* Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47-59, doi:10.1016/j.cell.2005.10.042 (2006).

252 Ahmadiyeh, N. *et al.* 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 9742-9746, doi:10.1073/pnas.0910668107 (2010).

253 Wright, J. B., Brown, S. J. & Cole, M. D. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Molecular and cellular biology* **30**, 1411-1420, doi:10.1128/MCB.01384-09 (2010).

254 Pomerantz, M. M. *et al.* Evaluation of the 8q24 prostate cancer risk locus and MYC expression. *Cancer research* **69**, 5568-5574, doi:10.1158/0008-5472.CAN-09-0387 (2009).

255 Trumpp, A. *et al.* c-Myc regulates mammalian body size by controlling cell number but not cell size. *Nature* **414**, 768-773, doi:10.1038/414768a (2001).

256 Dubois, N. C. *et al.* Placental rescue reveals a sole requirement for c-Myc in embryonic erythroblast survival and hematopoietic stem cell function. *Development* **135**, 2455-2465, doi:10.1242/dev.022707 (2008).

257 Bettess, M. D. *et al.* c-Myc is required for the formation of intestinal crypts but dispensable for homeostasis of the adult intestinal epithelium. *Molecular and cellular biology* **25**, 7868-7878, doi:10.1128/MCB.25.17.7868-7878.2005 (2005).

258 Cohen, J. C. *et al.* Transient in utero knockout (TIUKO) of C-MYC affects late lung and intestinal development in the mouse. *BMC developmental biology* **4**, 4, doi:10.1186/1471-213X-4-4 (2004).

259 Muncan, V. *et al.* Rapid loss of intestinal crypts upon conditional deletion of the Wnt/Tcf-4 target gene c-Myc. *Molecular and cellular biology* **26**, 8418-8426, doi:10.1128/MCB.00821-06 (2006).

260 Athineos, D. & Sansom, O. J. Myc heterozygosity attenuates the phenotypes of APC deficiency in the small intestine. *Oncogene* **29**, 2585-2590, doi:10.1038/onc.2010.5 (2010).

261 Ignatenko, N. A. *et al.* Role of c-Myc in intestinal tumorigenesis of the ApcMin/+ mouse. *Cancer biology & therapy* **5**, 1658-1664 (2006).

262 Yekkala, K. & Baudino, T. A. Inhibition of intestinal polyposis with reduced angiogenesis in ApcMin/+ mice due to decreases in c-Myc expression. *Molecular cancer research : MCR* **5**, 1296-1303, doi:10.1158/1541-7786.MCR-07-0232 (2007).

263 He, T. C. *et al.* Identification of c-MYC as a target of the APC pathway. *Science* **281**, 1509-1512 (1998).

264 Korinek, V. *et al.* Constitutive transcriptional activation by a beta-catenin-Tcf complex in APC-/- colon carcinoma. *Science* **275**, 1784-1787 (1997).

265 Morin, P. J. *et al.* Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. *Science* **275**, 1787-1790 (1997).

266 Tapscott, S. J. *et al.* MyoD1: a nuclear phosphoprotein requiring a Myc homology region to convert fibroblasts to myoblasts. *Science* **242**, 405-411 (1988).

267 Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51**, 987-1000 (1987).

268 Weintraub, H. *et al.* Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 5434-5438 (1989).

269 Lassar, A. B., Paterson, B. M. & Weintraub, H. Transfection of a DNA locus that mediates the conversion of 10T1/2 fibroblasts to myoblasts. *Cell* **47**, 649-656 (1986).

270 Porcher, C. *et al.* The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell* **86**, 47-57 (1996).

271 Robb, L. *et al.* The scl gene product is required for the generation of all hematopoietic lineages in the adult mouse. *The EMBO journal* **15**, 4123-4129 (1996).

272 Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. & Gaul, U. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* **451**, 535-540, doi:10.1038/nature06496 (2008).

273 Grove, C. A. *et al.* A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. *Cell* **138**, 314-327, doi:10.1016/j.cell.2009.04.058 (2009).

274 Zhu, C. *et al.* High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome research* **19**, 556-566, doi:10.1101/gr.090233.108 (2009).

275 Noyes, M. B. *et al.* Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **133**, 1277-1289, doi:10.1016/j.cell.2008.05.023 (2008).

276 Hammar, P. *et al.* The lac repressor displays facilitated diffusion in living cells. *Science* **336**, 1595-1598, doi:10.1126/science.1221648 (2012).

277 Ng, R. K. & Gurdon, J. B. Epigenetic memory of an active gene state depends on histone H3.3 incorporation into chromatin in the absence of transcription. *Nature cell biology* **10**, 102-109, doi:10.1038/ncb1674 (2008).

278 Rando, O. J. & Verstrepen, K. J. Timescales of genetic and epigenetic inheritance. *Cell* **128**, 655-668, doi:10.1016/j.cell.2007.01.023 (2007).

279 Kadauke, S. *et al.* Tissue-specific mitotic bookmarking by hematopoietic transcription factor GATA1. *Cell* **150**, 725-737, doi:10.1016/j.cell.2012.06.038 (2012).

280 Xing, H., Vanderford, N. L. & Sarge, K. D. The TBP-PP2A mitotic complex bookmarks genes by preventing condensin action. *Nature cell biology* **10**, 1318-1323, doi:10.1038/ncb1790 (2008).

281 Merkenschlager, M. & Odom, D. T. CTCF and cohesin: linking gene regulatory elements with their targets. *Cell* **152**, 1285-1297, doi:10.1016/j.cell.2013.02.029 (2013).

282 Uhlmann, F. & Hopfner, K. P. Chromosome biology: the crux of the ring. *Current biology : CB* **16**, R102-105, doi:10.1016/j.cub.2006.01.023 (2006).

283 Badrinarayanan, A., Reyes-Lamothe, R., Uphoff, S., Leake, M. C. & Sherratt, D. J. In vivo architecture and action of bacterial structural maintenance of chromosome proteins. *Science* **338**, 528-531, doi:10.1126/science.1227126 (2012).

284 Peters, J. M., Tedeschi, A. & Schmitz, J. The cohesin complex and its roles in chromosome biology. *Genes & development* **22**, 3089-3114, doi:10.1101/gad.1724308 (2008).

285 Easton, D. F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087-1093, doi:10.1038/nature05887 (2007).

286 Jones, A. M. *et al.* Thyroid cancer susceptibility polymorphisms: confirmation of loci on chromosomes 9q22 and 14q13, validation of a recessive 8q24 locus and failure to replicate a locus on 5q24. *Journal of medical genetics* **49**, 158-163, doi:10.1136/jmedgenet-2011-100586 (2012).

287 Naar, A. M., Lemon, B. D. & Tjian, R. Transcriptional coactivator complexes. *Annual review of biochemistry* **70**, 475-501, doi:10.1146/annurev.biochem.70.1.475 (2001).

288 Cirillo, L. A. *et al.* Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Molecular cell* **9**, 279-289 (2002).

289 Clark, K. L., Halay, E. D., Lai, E. & Burley, S. K. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **364**, 412-420, doi:10.1038/364412a0 (1993).

290 Sekiya, T. & Zaret, K. S. Repression by Groucho/TLE/Grg proteins: genomic site recruitment generates compacted chromatin in vitro and impairs activator binding in vivo. *Molecular cell* **28**, 291-303, doi:10.1016/j.molcel.2007.10.002 (2007).

291 Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes & development* **25**, 2227-2241, doi:10.1101/gad.176826.111 (2011).

292 Jozwik, K. M. & Carroll, J. S. Pioneer factors in hormone-dependent cancers. *Nature reviews. Cancer* **12**, 381-385, doi:10.1038/nrc3263 (2012).

293 Funnell, A. P. & Crossley, M. Homo- and heterodimerization in transcriptional regulation. *Advances in experimental medicine and biology* **747**, 105-121, doi:10.1007/978-1-4614-3229-6_7 (2012).

294 Perez-Pinera, P., Ousterout, D. G. & Gersbach, C. A. Advances in targeted genome editing. *Current opinion in chemical biology* **16**, 268-277, doi:10.1016/j.cbpa.2012.06.007 (2012).

295 Burgess, D. J. Technology: a CRISPR genome-editing tool. *Nature reviews. Genetics* **14**, 80, doi:10.1038/nrg3409 (2013).

296 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823, doi:10.1126/science.1231143 (2013).

297 Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823-826, doi:10.1126/science.1232033 (2013).

298 Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910-918, doi:10.1016/j.cell.2013.04.025 (2013).