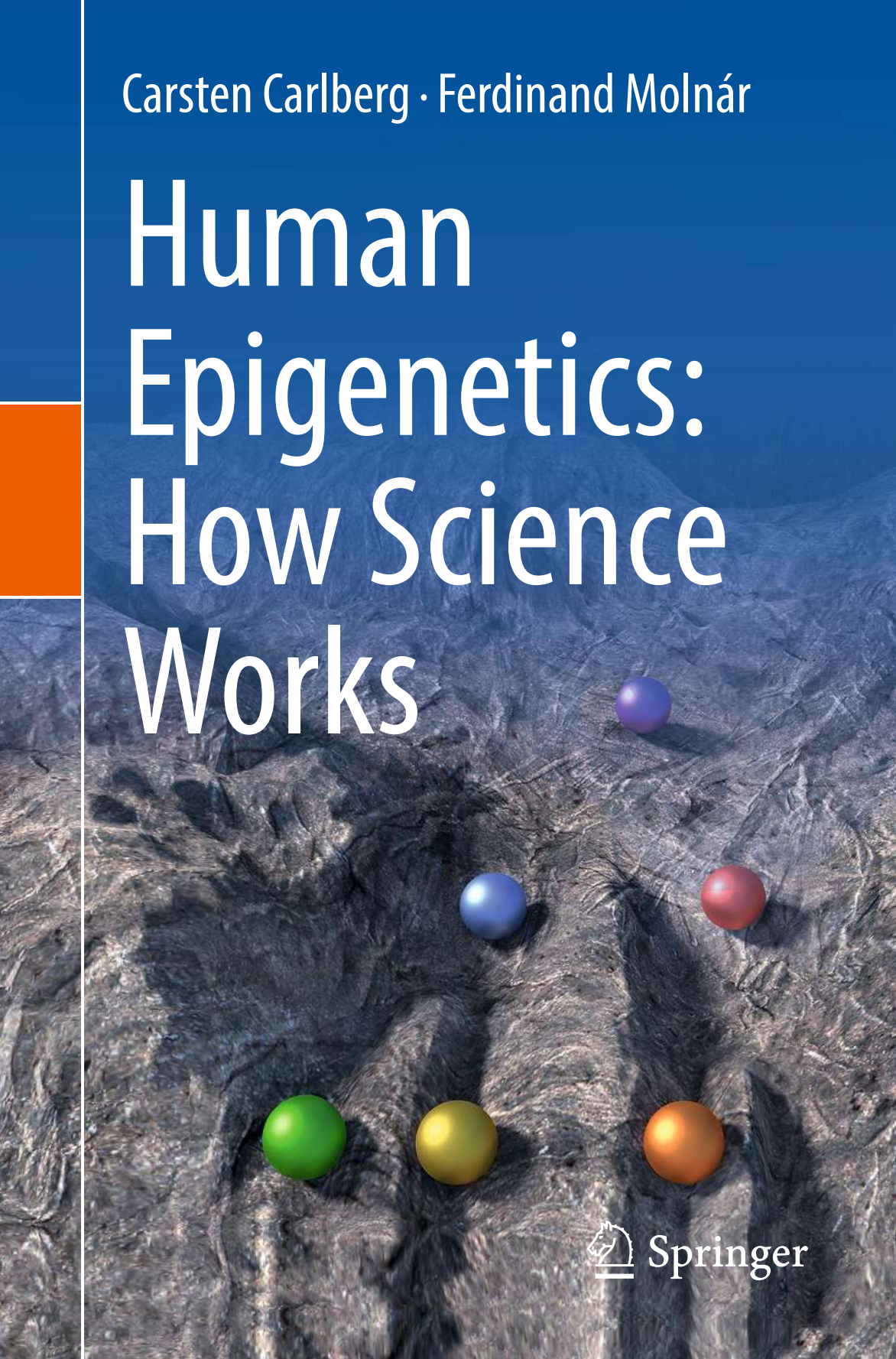
Carsten Carlberg · Ferdinand Molnár

Human Epigenetics: How Science Works



Springer

Human Epigenetics: How Science Works

Carsten Carlberg • Ferdinand Molnár

Human Epigenetics: How Science Works

Springer

Carsten Carlberg
Institute of Biomedicine
University of Eastern Finland
Kuopio, Finland

Ferdinand Molnár
Department of Biology
Nazarbayev University
Nur-Sultan, Kazakhstan

Cover figure legend: Waddington's epigenetic landscape as representation of the terrain rendering. The spheres represent cells at various developmental stages of their differentiation. The stem cell on the top of the hill (purple colored sphere) undergoes differentiation to tissue cell precursors (blue and red spheres) finally reaching the terminal differentiation at tissue level (green, yellow and orange spheres).

ISBN 978-3-030-22906-1 ISBN 978-3-030-22907-8 (eBook)
https://doi.org/10.1007/978-3-030-22907-8

Cover illustration: Alen German/Ferdinand Molnár

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Epigenetics describes the packaging and accessibility of the genome that we carry in each of the trillions of cells forming our body. The prefix "epi" (meaning "upon," "above," or "beyond") indicates that, in contrast to genetics, epigenetic processes do not affect the DNA sequence of our genome. This means that there exists a layer of information beyond that encoded in our genome. Catalyzed by the sequencing of our genome (completed in 2001), new methods for a genome-wide assessment of epigenetics have been developed. This opened the field of epigenomics describing epigenetics on a whole-genome view.

Genomic DNA is wrapped around complexes of histone proteins that help to fit the genome into a cell nucleus with a diameter of less than 10 μm. This protein-DNA complex is referred to as chromatin. The differentiation of embryonic stem (ES) cells into specialized cell types happens during embryogenesis within the first weeks of the life of a fetus. Moreover, in every moment of an adult's life stem cells in the bone marrow, the skin and the intestine are growing and differentiating into specialized cells that replace cell loss, such as of our immune system or of our body's outer and inner surface. The underlying mechanism of all these differentiation processes is epigenetic programing of chromatin structure, i.e., a change of the so-called epigenetic landscape. Thus, the most important function of chromatin is to keep in a cell- and tissue-specific manner some 90% of our genome inaccessible to transcription factors and polymerases. In other words, chromatin acts as a gatekeeper for undesired gene activation. Therefore, each of the 400 tissues and cell types forming our body uses a different subset of the 20,000 protein-coding genes of our genome. Thus, *each of us has only one genome but at least 400 different epigenomes*. Epigenetics prevents that, for example, a kidney cell changes overnight into a neuron or vice versa. In this way, *epigenetics provides terminally differentiated cells with permanent memory about their identity*.

In addition to its static function, epigenetics has also a dynamic aspect, where the activation of intracellular signal transduction cascades via extracellular signals, such as peptide hormones, cytokines, or growth factors, results in the activation of transcription factors and chromatin-modifying enzymes. Probably, the most dominant external signal that we are exposed to is our diet. Many dietary compounds, such as unsaturated fatty acids, directly activate transcription factors or cause changes in the levels of metabolites, such as α-ketoglutarate, which modulate the activity of chromatin-modifying enzymes. The actions of these nuclear proteins

cause local changes of the epigenome, which enable and modulate the transcription of specific target genes of the different signals affecting a cell. Most of these changes are transient, but some may leave permanent marks on the epigenome. In this way, *our epigenome can memorize environmental events*, such as what and how much we have eaten, whether we had been in contact with microbes, or whether we had been stressed in any other way.

Epigenetics provides a molecular explanation how our genome is connected with environmental signals. It associates with our lifestyle and environment during intrauterine as well as postnatal conditions. The dynamic component of epigenetics implies that *some epigenetic programing events are reversible*. For example, if an unhealthy lifestyle paired with food excess and physical inactivity over many years causes epigenetic programing of metabolic tissues that results in insulin resistance, this process may be reversed by significant lifestyle changes leading to a *reprograming of the dynamic component of the epigenome*. This reprograming of our epigenome can have significant consequences for our health, i.e., as long as no irreversible tissue damage has happened, we may have it in our own hands to reverse a disease condition. Thus, not only our mental memory, such as memories of our childhood, is a learning process that is based on epigenetic programing of neurons, but each cell of our body has an epigenetic memory recording the perturbations that the cell had been exposed to.

Most noncommunicable human diseases have a genetic, inherited component as well as an epigenetic component, which is based on our lifestyle choices and environmental exposures. Many common diseases, such as type 2 diabetes, can be explained only to some 20% via a genetic predisposition. *We cannot change the genes that we are born with, but we can take care of the remaining 80% being primarily based on our epigenome*. This means that the genetic predisposition for a disease can be counterbalanced by an appropriate healthy lifestyle that modulates the epigenome of the affected tissues. It is well known that there is a high level of individual responsibility for staying healthy, but a detailed understanding of epigenetics provides a molecular explanation for this life philosophy.

These first examples indicate that *epigenetics affects diverse aspects in health and disease*. This book will discuss the central importance of epigenetics during embryogenesis and cellular differentiation as well as in the process of aging and the risk for the development of cancer. Moreover, the role of the epigenome as a molecular storage of cellular events, not only in the brain but also in metabolic organs and in the immune system, will be described. In this context, epigenetic effects on neurodegenerative diseases as well as autism; metabolic diseases, such as type 2 diabetes; and disorders based on a malfunctional immune system, such as autoimmunity, will be discussed.

The content of the book is linked to a series of lecture courses in “Molecular Medicine and Genetics,” “Molecular Immunology,” “Cancer Biology,” and “Nutrigenomics” that were given by one of us (C. Carlberg) in different forms since 2002 at the University of Eastern Finland in Kuopio. This book represents an updated version of our textbook *Human Epigenomics* (ISBN 978-981-10-7614-8). However, we shortened and simplified the content in order to give also the

undergraduate students and other people engaged in life sciences an easier start into the topic. This book also relates to our textbooks *Mechanisms of Gene Regulation* (ISBN 978-94-017-7741-4) and *Nutrigenomics* (ISBN 978-3-319-30415-1), the studying of which may be interesting to the readers who like to get more detailed information. The first five chapters of this book will explain the molecular basis of epigenetics, while the following seven chapters will provide examples for the impact of epigenetics in human health and disease. A glossary in the appendix will explain the major specialist's terms (highlighted in orange at their first time use).

We hope that the readers will enjoy this rather visual book and get as enthusiastic about epigenetics as the authors are.

Kuopio, Finland — Carsten Carlberg
Nur-Sultan, Kazakhstan — Ferdinand Molnár
April 2019

Acknowledgments

The authors would like to thank Reinhard Bornemann, MD, DrPH, PhD, Andrea Hanel, and Gregory Suszko for their extensive proofreading and constructive criticism.

Contents

Abbreviations

3C	chromosome conformation capture
3D	three-dimensional
5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5hmU	5-hydroxyuracil
5mC	5-methylcytosine
β-OHB	D-β-hydroxybutyrate
AMPK	AMP-activated protein kinase
ARID	AT-rich interaction domain
ARNTL	aryl hydrocarbon receptor nuclear translocator-like
ASH1L	ASH1 like histone lysine methyltransferase
ASIP	agouti-signaling protein
ATAC-seq	assay for transposase-accessible chromatin using sequencing
BDNF	brain-derived neurotrophic factor
BMI1	BMI1 proto-oncogene, polycomb ring finger
bp	base pair
BRD	bromodomain containing
CAGE	cap analysis of gene expression
CBFB	core-binding factor subunit β
CBX	chromobox
CDKN	cyclin-dependent kinase inhibitor
CEBP	CCAAT-enhancer-binding protein
ChIP	chromatin immunoprecipitation
CIMP	CpG island methylator phenotype
CLOCK	clock circadian regulator
CLP	common lymphoid progenitor
CMP	common myeloid progenitor
CREBBP	CREB-binding protein, also called KAT3A
CRY1	cryptochrome circadian clock 1
CTCF	CCCTC-binding factor
CXCR	C-X-C motif chemokine receptor
DACH1	dachshund family transcription factor 1
DAMP	damage-associated molecular pattern

DNMT	DNA methyltransferase
DOHaD	developmental origins of health and disease
DOT1L	DOT1-like histone lysine methyltransferase
EED	embryonic ectoderm development
EHMT2	euchromatic histone lysine methyltransferase 2
EMT	epithelial-to-mesenchymal transition
ENCODE	encyclopedia of DNA elements
EP300	E1A-binding protein p300, also called KAT3B
eRNA	enhancer RNA
ES	embryonic stem
EZH	enhancer of zeste homolog
FAD	flavin adenine dinucleotide
FAIRE	formaldehyde-assisted isolation of regulatory elements
FANTOM	functional annotation of the mammalian genome
FMR1	fragile X mental retardation 1
FOX	forkhead box
FXN	frataxin
GATA	GATA-binding protein
GMP	granulocyte-monocyte progenitor
GTEx	genotype-tissue expression
GWAS	genome-wide association study
HAT	histone acetyltransferase
HDAC	histone deacetylase
HGPS	Hutchinson-Gilford progeria syndrome
Hi-C	high-throughput chromosome capture
HNRNPU	heterogeneous nuclear ribonucleoprotein U
HOTAIR	HOX transcript antisense RNA
HP1	heterochromatin protein 1, official name CBX5
HSC	hematopoietic stem cell
IAP	intracisternal A particle
ICR	imprint control region
IDH	isocitrate dehydrogenase
IGF2	insulin-like growth factor 2
IHEC	international human epigenome consortium
IL	interleukin
INFG	interferon γ
INO80	INO80 complex subunit
iPOP	integrative personal omics profiling
iPS	induced pluripotent stem
ISWI	imitation SWI
kb	kilo base pairs (1,000 bp)
KCNQ1	potassium voltage-gated channel subfamily Q member 1
KDM	lysine demethylase
KLF4	Krüppel-like factor 4
KMT	lysine methyltransferase

LAD	lamin-associated domain
LCR	locus control region
LINE	long interspersed element
LOCK	large organized chromatin K9-modification
LSD1	lysine specific demethylase 1, also called KDM1A
LTR	long terminal repeat
Mb	mega base pairs (1,000,000 bp)
MBD	methyl-DNA-binding domain
mCH	non-CpG methylation
MECP2	methyl-CpG-binding protein 2
MeDIP-seq	methylated DNA immunoprecipitation sequencing
MEIS1	meis homeobox 1
MEP	megakaryocyte-erythrocyte progenitor
MEN1	menin 1
MGMT	O-6-methylguanine-DNA methyltransferase
MHC	major histocompatibility complex
MIC	MHC class I polypeptide-related sequence
miRNA	micro RNA
MLH1	MutL homolog 1
MPP	multipotent progenitor
mRNA	messenger RNA
MS	multiple sclerosis
MTHFR	methylenetetrahydrofolate reductase
NAD	nicotinamide adenine dinucleotide
NAMPT	mononucleotide phosphoribosyltransferase
NANOG	nanog homeobox
NCOR	nuclear receptor corepressor
ncRNA	noncoding RNA
NF-κB	nuclear factor κB
NK	natural killer
NSD	nuclear receptor-binding SET domain protein
nt	nucleotides
OCT4	octamer-binding transcription factor 4
PAMP	pathogen-associated molecular pattern
PBMC	peripheral blood mononuclear cell
PDCD1	programmed cell death 1, also called PD1
PDGFRA	platelet-derived growth factor receptor α
PER1	period circadian clock 1
PGC	primordial germ cell
Pol II	RNA polymerase II
PPARGC1A	proliferator-activated receptor gamma, coactivator 1 α
PRC	Polycomb repressive complex
PU.1	purine-rich box 1
RB1	RB transcriptional corepressor 1
RCOR	REST corepressor

REL	REL proto-oncogene, NF-κB subunit
REST	RE1-silencing transcription factor
RUNX1	runt-related transcription factor 1
SAH	S-adenosylhomocysteine
SAM	S-adenosyl-L-methionine
SCN	suprachiasmatic nucleus
SETD2	SET domain containing 2
SHARP	SMRT/HDAC1-associated repressor protein
SINE	short interspersed element
SIRT	sirtuin
SLE	systemic lupus erythematosus
SMARC	SWI/SNF-related matrix-associated actin-dependent regulators of chromatin
SNP	single-nucleotide polymorphism
SOX2	SRY-box 2
STAT	signal transducer and activator of transcription
SUV39H	suppressor of variegation 3-9 homolog
SWI/SNF	switching/sucrose non-fermenting
TAD	topologically associated domain
TCGA	The Cancer Genome Atlas
TDG	thymine-DNA glycosylase
TET	ten-eleven translocation
T_H	T helper
TP53	tumor protein p53
TRIM	tripartite motif-containing protein
tRNA	transfer RNA
TSS	transcription start site
UHRF1	ubiquitin-like with plant homeodomain and RING finger domain 1
UTR	untranslated region
WRN	Werner syndrome RecQ like helicase
Xist	X-inactive specific transcript

Introduction

1

Abstract

Chromatin is the physical representation of epigenetics. The degree of chromatin compaction is inversely correlated with gene expression. The epigenome responds to intra- and extra-cellular signals *via* changes in DNA methylation and histone modifications. Accordingly, the epigenetic landscape visualizes epigenomic changes during cellular differentiation, tumorigenesis and other cellular perturbations. Thus, **the proper action of the epigenome has a major impact on our health and disease**.

Keywords

Epigenetics · Chromatin · Euchromatin · Heterochromatin · Human genome · Gene · Central dogma of molecular biology · Epigenome · Epigenetic landscape · Big Biology projects

1.1 What is Epigenetics?

Most of us probably had the first contact with epigenetics, when we were looking at chromosomes, either directly through a microscope or in a textbook. Our cells have 22 pairs of autosomal chromosomes and either two X chromosomes (female) or one X and one Y chromosome (male). Chromosomes are formed of chromatin, which is the macromolecular complex of genomic DNA and nuclear proteins. Chromatin packs our genome, *i.e.,* the 16 to 85 mm long DNA molecules of each chromosome into the nucleus of a cell with a diameter of only 6–10 μm. However, chromosomes are only visible during a special phase of the cell cycle, referred to as the metaphase of mitosis (Sect. 2.2). During mitosis the prime importance is that the genome is divided equally to the two daughter cells. Therefore, the 46 DNA molecules are packed into the form of highest compaction, referred to as chromosomes. At that time all genes of our genome are temporally switched off for approximately only

C. Carlberg, F. Molnár, *Human Epigenetics: How Science Works*,
https://doi.org/10.1007/978-3-030-22907-8_1

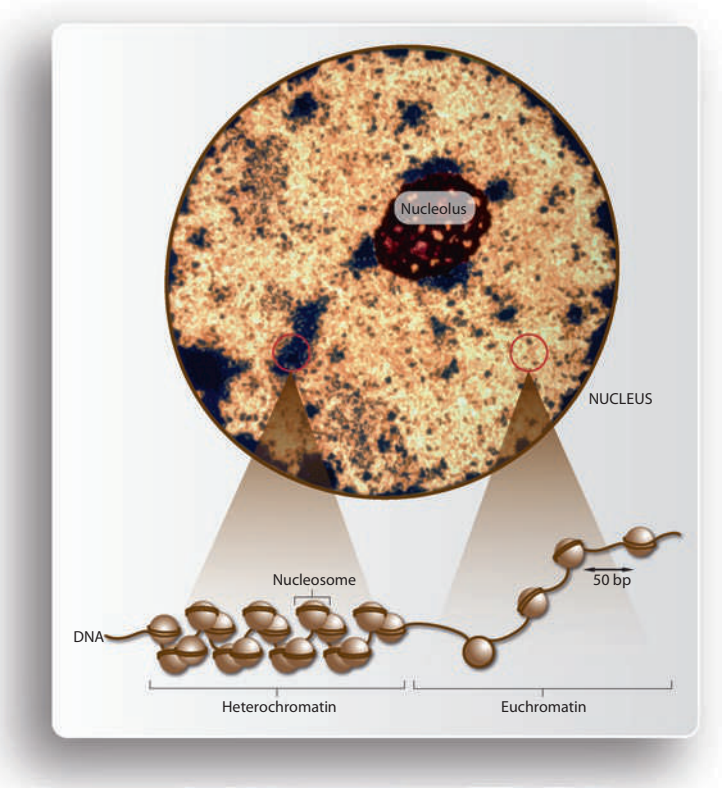


Fig. 1.1 **Eu- and heterochromatin**. An electron microscopic picture of a nucleus during interphase is shown. The darker areas located mostly in the periphery of the nucleus represent constitutive heterochromatin (inactive), whereas the lighter areas in the center are euchromatin (active). The nucleolus is a nuclear sub-structure, where ribosomal RNA genes are transcribed. A schematic drawing (**bottom**) monitors dense nucleosome packaging in heterochromatin (**left**, also referred to as closed chromatin) and loose nucleosome arrangement in euchromatin (**right**, open chromatin)

1 h, *i.e.,* **mitosis represents the most extreme case of epigenetic regulation of our genome**. However, the very most cells of our body are not dividing and are in the interphase. In this phase, within the nucleus only lighter and darker areas can be distinguished, which represent lightly packed euchromatin and tightly packed heterochromatin, respectively (Fig. 1.1).

Genes are defined as segments of a chromosome, *i.e.,* stretches of genomic DNA, encoding either for proteins or non-coding RNA (ncRNA). The phenomena of gene imprinting (Sect. 3.3) and X chromosome inactivation (Sect. 6.3) were the first indications that identical genetic material (from individual genes to whole chromosomes) can be in the same nucleus in an "on" as well as in an "off" state. Thus, genes can be either actively expressed, *i.e.,* their information is copied into RNA, or they are inactive, *i.e.,* not expressed. In analogy to the term "epigenesis" (*i.e.,* morphogenesis and development of an organism) Conrad Waddington

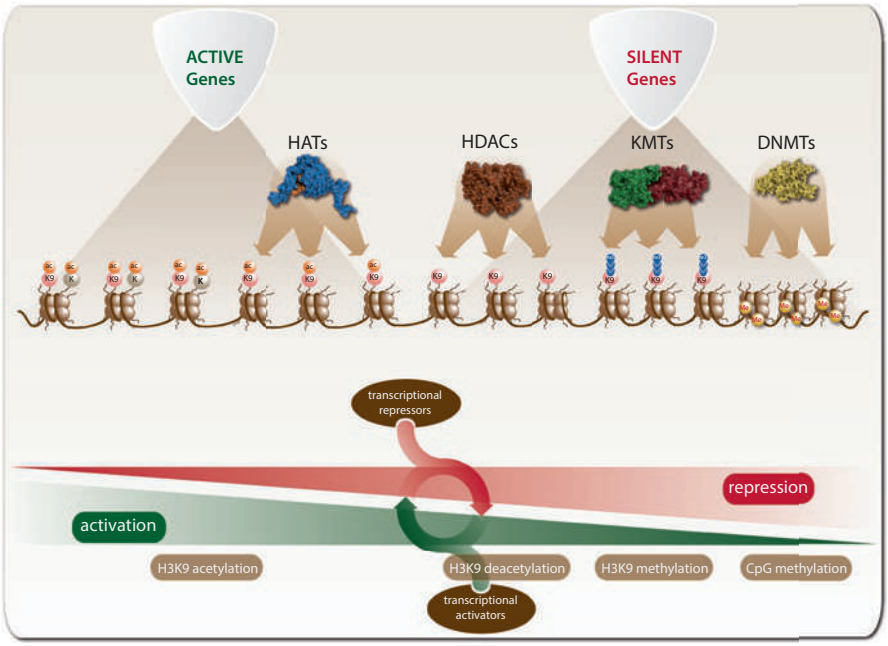


Fig. 1.2 **Active and silent chromatin stages**. DNA methylation and histone modification have different roles in gene silencing. While DNA methylation represents a very stable silencing mark (**right**) that is seldom reversed, histone modifications mostly lead to transient and reversible transcriptional activation and repression (**left**). In heterochromatin CpG islands (Sect. 3.1) are methylated, nucleosomes (Sect. 2.1) are arranged in a regular dense fashion and histone proteins are tri-methylated at position H3K9 (Sect. 4.1). K, lysine; ac, acetylated; me, methylated; HATs, histone acetyltransferases; HDACs; histone deacetylases; KMTs, lysine methyltransferases; DNMTs, DNA methyltransferases

proposed in 1942 the word "epigenetics" for describing changes in the phenotype that are not based on changes in the genotype. This definition was later extended to "**epigenetics is the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail change in DNA sequence**". In the following chapters, we will describe in more detail that epigenetics refers to modifications of the genome, such as DNA methylation (Chap. 3) and histone modifications (Chap. 4), but does not involve a change in the nucleotide sequence.

Epigenetic changes are functionally relevant when they result in changes in mRNA levels and initiate the production of proteins. Such gene expression patterns may stay persistent throughout the following cell divisions for the remainder of the cell's life and can even last for multiple generations (Sect. 7.1). This implies that **epigenetics can inherit the information which genes are expressed in which cells**.

In this book, we define epigenetics primarily *via* its molecular representation, *i.e.*, the different stages of chromatin accessibility and function (Fig. 1.2). In the past, the packaging of genomic DNA into higher density, such as the extreme

example of a metaphase chromosome, was considered to be the main function of chromatin. In fact, the small size of the nucleus requires a compaction of our genome (Box 1.1) by more than 200,000-fold. This is in part achieved by wrapping genomic DNA around histone octamers (Sect. 2.1). However, nowadays it is clear that

Box 1.1 The Human Genome

The human genome is the complete sequence of the anatomically modern human (*Homo sapiens*) and was obtained by the *Human Genome Project* (www.genome.gov/10001772) *via* whole genome sequencing. This reference sequence is accessible *via* different genome browsers (Box 2.2) and represents the assembly of the genomes of a few young healthy donors. With the exception of germ cells, *i.e.,* female oocytes and male sperm, each human cell contains a diploid genome formed by 2x 3.235 billion base pairs (bp), *i.e.,* 3235 mega bp (Mb), that is distributed on 2x 22 autosomal chromosomes and two X chromosomes for females and a XY chromosome set for males. In addition, every mitochondrium contains 16.6 kilo base pairs (kb) mitochondrial DNA. The haploid human genome encodes for some 20,000 protein-coding genes and about the same number of ncRNA genes. The protein-coding sequence covers less than 2% of our genome, *i.e.,* **the vast majority of the genome is non-coding and seems to have primarily regulatory function**

The *1000 Genomes Project* (www.internationalgenome.org) provided the sequence of in total 2504 individuals from 26 populations covering all five continents. In total, the project describes some 88 million genome variants, of which 84.7 million are single nucleotide polymorphisms (SNPs), 3.6 million short insertion/deletions (indels) and 60,000 larger copy number variants. In average, a typical genome contains some 150 variants resulting in protein truncation, 10,000 changing amino acids in translated proteins and 500,000 affecting transcription factor binding sites (Sect. 12.1).

Almost 50% of the sequence of our genome is formed by repetitive DNA, which is sorted into the following categories (by order of frequency):

Long interspersed elements (LINEs, 500–8000 bp)	21%
Short interspersed elements (SINEs, 100–300 bp)	11%
Retrotransposons, such as long terminal repeats (LTRs, 200–5000 bp)	8%
DNA transposons (200–2000 bp)	3%
Minisatellite, microsatellite or major satellite (2–100 bp)	3%

LINEs and SINEs are identical or nearly identical DNA sequences that are separated by large numbers of nucleotides, *i.e.*, the repeats are spread throughout the whole human genome. LTRs are characterized by sequences that are found at each end of retrotransposons. DNA transposons are full-length autonomous elements ("jumping DNA") that encode for a transposase, *i.e.*, an enzyme that transposes DNA from one to another position in the genome. Microsatellites are often associated with centromeric or peri-centromeric regions and are formed by tandem repeats of 2–10 bp in length. Minisatellites and major satellites are longer, with a length of 10–60 bp or up to 100 bp, respectively.

another important aspect of epigenetics is the regulation of chromatin accessibility. This is critical for determining whether transcription factors and associated nuclear proteins recognize their binding sites within enhancer and promoter regions (Sect. 2.3). Chromatin accessibility is regulated by methylation of genomic DNA at cytosines (Sect. 3.1) and post-translational modifications of histone proteins (Sect. 4.1). Furthermore, DNA looping and other 3D chromatin structures control chromatin activity within the nucleus (Sect. 2.4). Accordingly, **chromatin accessibility plays an important role in regulating gene expression**, *i.e.,* transcription factors and histone proteins compete for critical loci within the genome.

More then 99% of the approximately 30 trillion ($3x10^{13}$) cells of our body are terminally differentiated, *i.e.*, they do not divide anymore and they are in the interphase. The lighter staining of euchromatin reflects its less compact chromatin structure. In the "beads-on-a-string" model of euchromatin (Fig. 1.1, bottom right), nucleosomes, *i.e.*, assemblies of some 150 bp genomic DNA with a complex of eight histone proteins (Sect. 2.1), are regularly positioned every 200 bp leaving 50 bp-sized gaps of freely accessible genomic DNA between them. Euchromatin becomes condensed only during mitosis and has a higher gene density than heterochromatin. **Genes can only be transcribed into RNA, *i.e.*, they get expressed, when they are located within euchromatin**. While the euchromatin fiber has a diameter of 11 nm, more compacted heterochromatin forms a 30 nm fiber (Fig. 1.1, bottom left) or even higher order structures of 100 nm in diameter. For comparison, the diameter of a chromosome is even 700 nm.

During cellular differentiation, such as embryogenesis (Sect. 6.1), regions of genomic DNA end up either in a stage of permanent quiescence, referred to as constitutive heterochromatin, which is often highly methylated (Fig. 1.2, bottom right), or in facultative heterochromatin, which is characterized by silenced ("poised") genes that remain their potential to be activated by appropriate signals (Fig. 1.2, bottom central). Constitutive heterochromatin is found preferentially on repetitive genome sequences (Box 1.1), such as centromeres and telomeres, where gene density is low. Moreover, genes that should not be active in a given cell type, for example, embryonal pluripotent transcription factors (Sect. 6.2) in adult cells, are found in constitutive heterochromatin. In contrast, **facultative heterochromatin can reversibly transform into euchromatin**. Thus, facultative heterochromatin represents the most dynamic component of the epigenome, since its status depends on extracellular signals. The inactive second X chromosome in female cells is an example of facultative heterochromatin and can be observed as a Barr body in interphase nuclei. This means that female cells have the potential to reactivate the genes of the second X chromosome, in case the respective gene of the first X chromosome is defective.

As indicated by the central dogma of molecular biology (Box 1.2) the initial step of gene expression is the transcription of the genomic DNA of the gene body (*i.e.*, the DNA sequence between the transcription start site (TSS) and the transcription termination site) into mRNA, which after splicing and transport from the nucleus to the cytoplasm is translated into protein (Fig. 1.3). Proteins are the "workers" within a cell and basically mediate all functions therein, such as signal transduction, catalysis and

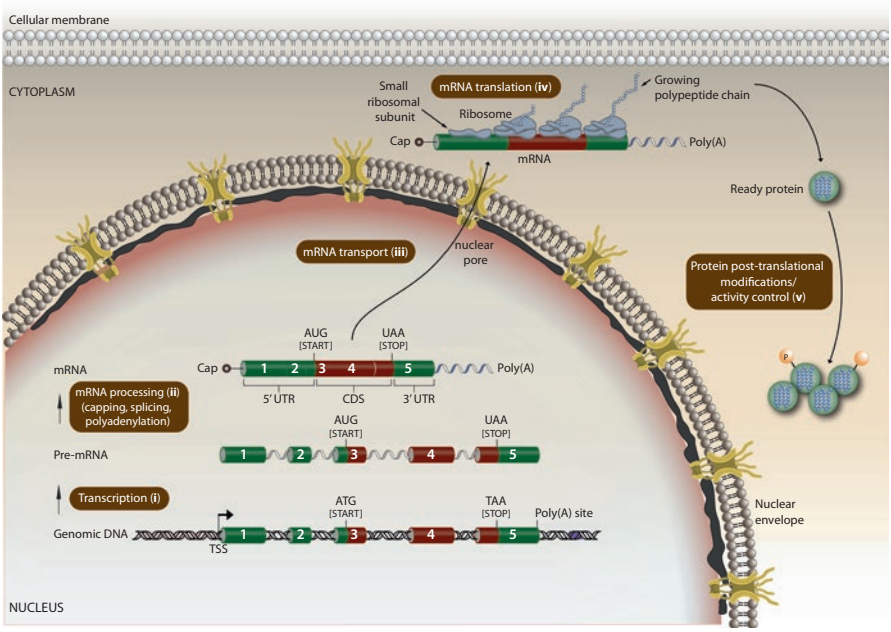


Fig. 1.3 **Flow of information from DNA to RNA**. The TSS of a gene is the first nucleotide that is transcribed into mRNA, *i.e.*, it defines the "start" of a gene body but has no defined sequence. In analogy, the "end" of a gene body is the position where RNA polymerase II (Pol II) dissociates from the genomic DNA template. The gene body is entirely transcribed into single-stranded pre-mRNA, which is composed of exons (numbered green and brown cylinders) and intervening introns (**i**). The introns are removed by splicing and the 5′- and 3′-end of the mRNA molecule are protected against digestion by exonucleases through a nucleotide cap and the addition of hundreds of adenines (polyadenylation (poly(A)), respectively (**ii**). Mature mRNA is then exported by an active, *i.e.*, ATP consuming, process from the nucleus through nuclear pores into the cytoplasm (**iii**). Small ribosome subunits scan the mRNA molecule from its 5′-end for the first available AUG (the "start codon"), assemble then with large subunits and perform protein translation process until they reach the sequence UAA, UAG or UGA (the "stop codons") (**iv**). The mRNA sequences upstream of the start codon and downstream of the stop codon are not translated and referred to as 5′- and 3′-untranslated regions (UTRs). The resulting polypeptide chains fold into proteins, most of which are further post-translationally modified in order to reach their full functional profile (**v**). Please note that for simplicity in this and in all following figures the nuclear envelope is drawn as single lipid bilayer and not as a double lipid bilayer. A, adenine; C, cytosine; G, guanine; T, thymine (occurs only in DNA); U, uracil (occurs only in RNA)

control of metabolic reactions, molecule transport and many more. In addition, proteins contribute to the structure and stability of cells and intra-cellular matrices. Therefore, **gene expression determines the phenotype, function and developmental state of cell types and tissues**. Gene expression patterns are cell-specific, but can also drastically change after exposure to intra- and extra-cellular signals and in response to pathological conditions, such as microbe infection or cancer.

While the genome is identical in each cell of an individual and stays reasonably static over a person's lifetime, **a large part of the epigenome is very dynamic,**

Box 1.2 The Central Dogma of Molecular Biology
The dogma indicates a clear direction in the flow of information from DNA to RNA to protein (Fig. 1.3). This implies that, besides a few exceptions, such as reverse transcription of the RNA genome of retroviruses, genomic DNA stores the building plan of on organism (in this book we will see that in addition also the epigenome has a large contribution). Accordingly, genes are defined as regions of genomic DNA that can be transcribed into RNA. In the original formulation of the dogma with "RNA" only "mRNA" was meant, *i.e.*, the RNA template used for protein translation, but it applies also to ncRNAs, such as rRNA, tRNA and micro RNA (miRNA). Nevertheless, **the expression of the 20,000 protein-coding genes of our genome, *i.e.*, their transcription into mRNA and the following translation into protein, determines which proteins are found in a given cell**

differs from one cell type to the other, and can respond to various signaling pathways. Accordingly, also the transcriptome (*i.e.*, the set of all translated RNAs) and the proteome (*i.e.*, the set of all produced proteins) of a cell are dynamic and cell-specific. For example, at a given genomic locus the methylation of nucleosome-forming histones results in the easily reversible formation of heterochromatin, whereas DNA methylation mostly leads to stable long-term repression (Fig. 1.2, bottom right). At regions where histones are acetylated *via* the action of histone acetyltransferases (HATs), genomic DNA remains unmethylated. In contrast, in repressed regions histones had been deacetylated by histone deacetylases (HDACs) and methylated by lysine methyltransferases (KMTs). In these regions, genomic DNA also becomes methylated *via* the action of DNA methyltransferases (DNMTs). HATs, HDACs, KMTs, and DNMTs are chromatin modifying enzymes that will be discussed later (Sects. 3.1 and 5.1) in more detail.

There are more different epigenetic stages of chromatin than euchromatin and heterochromatin, since on the level of both histone modification and DNA methylation there is a gradual transition between these extremes (Fig. 1.2, bottom). This also implies that there is a close co-operation between both types of chromatin marks: histone methylation is involved in directing DNA methylation patterns, while DNA methylation can serve as a template for some histone modifications after DNA replication.

1.2 The Epigenetic Landscape

The epigenetic landscape is a very illustrative model for understanding the underlying molecular mechanisms of cell fate decisions during development (Fig. 1.4). Cellular differentiation happens along lineages and is (under natural conditions) an irreversible forward-moving process; it results in highly specialized, terminally differentiated cell types.

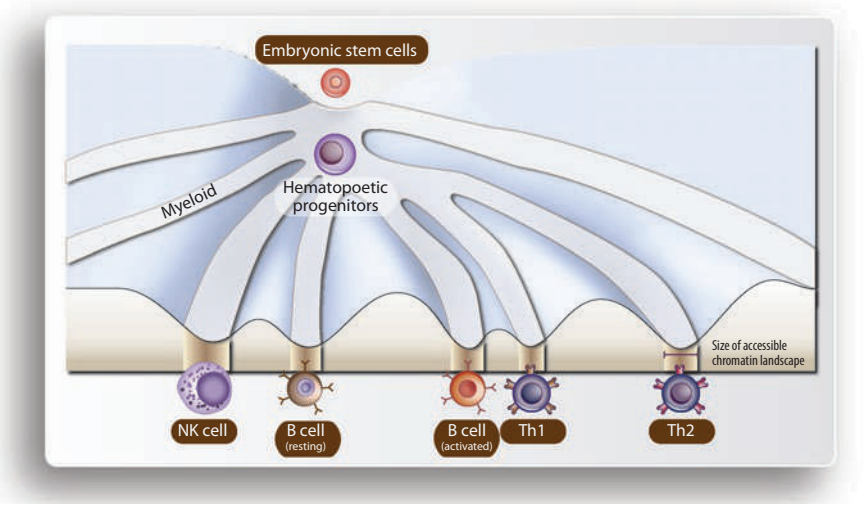


Fig. 1.4 **Epigenetic landscape model**. Cellular differentiation is accompanied by the progressive restriction of the epigenetic landscape indicated by the narrowing of valley floors. Here a subset of lymphoid differentiation is shown

In analogy, **cellular differentiation may be compared to a system of valleys of a mountain range**, where a cell (often represented by a ball, Fig. 1.5), for example, an embryonic stem (ES) cell, begins at the top and follows existing paths driven by gravitational force. The latter analogy should express that the path of differentiation has a clear direction. This directs the cell into one of several possible fates represented as valleys that get narrower in the trajectory toward terminally differentiated cell types (Fig. 1.5, bottom). Along the downhill path, cell fate decisions need to be taken at bifurcation points. These decisions often depend on the expression of lineage-determining transcription factors (Sect. 11.1). Once a cell has taken a decision, it is restricted in its subsequent decisions by the route it has taken.

The developmental potential of stem cells on top of the hill correlates with high entropy (*i.e.*, the potential to take a multitude of cellular stages), which declines during differentiation toward well defined cell types (Fig. 1.5, left). In contrast, when embryonal pluripotency transcription factors (Sect. 6.2), such as OCT4 (octamer-binding transcription factor 4) or NANOG (nanog homeobox), are reactivated in terminally differentiated cells, entropy can increase again and a cell may move uphill in the landscape (Fig. 1.5, center). This happens often during tumorigenesis, when (epi)mutations activate transcription factors or other nuclear proteins, such as chromatin modifiers, the activity of which results in gene expression heterogeneity (Sect. 8.1). The latter discontinues the cell fate choice and the transformed cells reach a state of higher entropy, in which they again proliferate and self-renew, *i.e.*, they are dedifferentiated compared with their normal counterparts (Fig. 1.5, right). The epigenetic landscape model is also used to illustrate the phenotypic plasticity of cells during the creation of induced pluripotent stem (iPS) cells (Box 1.3). Taken together, **the epigenetic landscape is an attractive,**

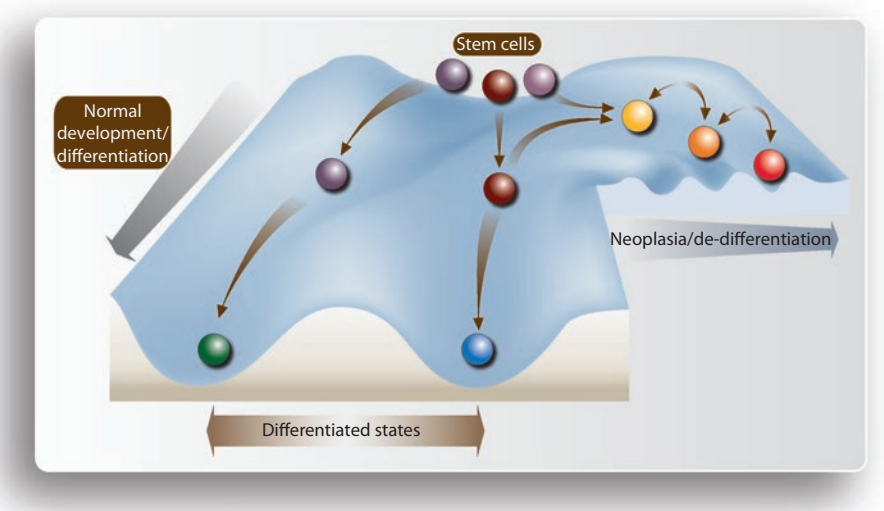


Fig. 1.5 **Phenotypic plasticity during cellular reprograming and neoplasia**. Waddington's landscape model can be used for the illustration of phenotypic plasticity of cells during normal development (**left**), the creation of iPS cells, *i.e.*, in cellular reprograming (**center**, Sect. 6.2), as well as during the induction of neoplasias, *i.e.*, in tumorigenesis (**right**, Sect. 8.1)

Box 1.3 Potency of Human Cells

The potency of a cell is its ability to differentiate into other cells. Totipotent cells can form all the cell types in a body including extra-embryonic placental cells. Only within the first 6–8 cell divisions after fertilization embryonic cells are totipotent. Pluripotent cells, such as ES cells, can give rise to all cell types forming our body. Multipotent cells, such as adult stem cells, can develop into more than one cell type but are more limited than pluripotent cells. In contrast, terminally differentiated cells, like more than 99% of those in our body, are unipotent. However, via the overexpression of pluripotency transcription factors unipotent cells can be induced to get pluripotent, *i.e.*, they transform into iPS cells

intuitively understandable model how the static information provided by the genome is translated dynamically into tissues and cell types.

1.3 Impact of Epigenetics on Health and Disease

For more than a decade it has been very popular to add the suffix "omics" to a molecular term in order to express that a set of molecules is investigated on a comprehensive and/or global level. After completing the sequence of our genome in 2001 (Box 1.1), "genomics" became the first omics discipline focusing on the study

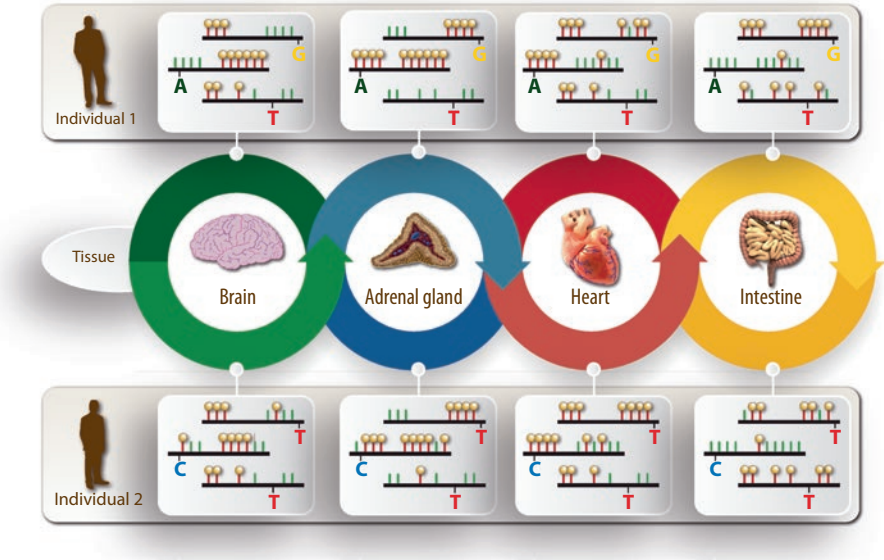


Fig. 1.6 **Individuals show epigenetic heterogeneity**. Tissue- and cell type-specific DNA methylations are displayed by clusters of methylated CG dinucleotides (CpGs, Sect. 3.1) that vary from tissue to tissue of the same individual. Filled circles illustrate methylated CpGs and lack of a circle unmethylated CpGs. SNPs are monitored by the corresponding base

of entire genomes in contrast to "genetics" that investigates individual genes. Genomics turned out to be the appropriate approach for the description and study of genetic variants, such as SNPs contributing to complex diseases like cancer, type 2 diabetes or Alzheimer's. Technological advances, in particular in next-generation sequencing methods (Sect. 2.5), led to the development of additional omics disciplines, such as epigenomics, transcriptomics, proteomics, metabolomics and nutrigenomics (Sect. 10.1).

Epigenomics is the global, comprehensive view of processes that modulate gene expression patterns in a cell independent from genome sequence. These patterns are primarily DNA methylation states (Sect. 3.1) and covalent modification of histone proteins (Sect. 4.1) that organize the architecture of the nucleus, restrict or facilitate transcription factor access to genomic DNA and preserve an epigenetic memory of past gene regulatory activities. **The epigenome can be considered as the "second dimension" of the genome**, which maintains cell-type specific gene expression patterns in normal processes, such as embryogenesis (Sect. 6.1) and aging (Sect. 7.3), as well as in cases of diseases, such as cancer (Chap. 8), autism, neurodegeneration (Chap. 9), type 2 diabetes (Chap. 10) or autoimmunity (Chap. 11).

In response to cellular perturbations by diet, microbe encounter, cellular stress or other environmental influences epigenomes vary a lot over time. Although different persons show consistency in the overall epigenome patterns of their tissues, **individuals vary far more on the level of their epigenomes than on the level of their genomes** (Fig. 1.6). This suggests that phenotypic differences between

individuals (as well as their pre-disposition for diseases) are rather based on the epigenome than on the genome (Sect. 12.1). For example, a study of the epigenome-wide patterns of five histone markers in 19 individuals indicated that the main inter-individual differences involve chromatin state transitions between active and repressive genomic regions and *vice versa*.

The epigenomic status of a tissue or a cell type depends on an effective interplay between environment and chromatin. In principle, **any perturbation of cellular homeostasis can result *via* epigenetic changes in long-lasting effects of the phenotype**, in particular when the perturbed cells are self-renewing stem cells or long-lived, terminally differentiated cells, such as neurons or memory T cells. These perturbations result in the activation of signal transduction pathways that often reach the nucleus. Within the nucleus activated transcription factors communicate with chromatin modifiers and remodelers (Chap. 5) and in this way create changes in epigenetic signatures, such as histone modifications, DNA methylation and 3D chromatin architecture. Some of these epigenetic changes are very transient (lasting from minutes to hours), while others can stay far longer (days, months or even years). Accordingly, the **epigenome serves as a storage facility of cellular perturbations of the past**. Speaking in numbers: each diploid cell contains approximately 30 million nucleosomes with more than 130 different possibilities for post-translational modifications (Sect. 3.1). With the approximately four billion (130 × 30 million) "letters" of the histone code, a single cell can store via histone modifications a tremendous amount of data.

The epigenome is able to memorize lifestyle events in basically every tissue or cell type. Thus, not only neurons store the memory of an individual (Sect. 9.2), but also the immune system memorizes encounters, for example, with microbes (Chap. 11), and metabolic organs, such as skeletal muscle, fat and liver, remember lifestyle choices on diet and physical activity (Chap. 10). In contrast, **a disruption of the epigenetic memory can lead to the onset of cancer** (Chap. 8).

With a delay of some 20 years molecular biologists followed the example of physicists and realized that some of their research aims can only be reached by international collaborations of dozens to hundreds of research teams and institutions in so-called "Big Biology" projects (Fig. 1.7). The *Human Genome Project*, which was launched in 1990 and completed in 2001 (Box 1.1), was the first example of a Big Biology project and has significantly changed the way of thinking in the bioscience community. As a consequence, more and more single gene studies shifted over to the genome scale.

The first Big Biology projects in the field of epigenomics and transcriptomics were the *ENCODE* (Encyclopedia of DNA elements, www.encodeproject.org) *Project* and the *FANTOM5* (Functional annotation of the mammalian genome, http://fantom.gsc.riken.jp) *Project*. The *ENCODE Project* systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. However, the project focused initially on some 150 human cell lines rather than on primary cells. The *FANTOM5 Project* used the method cap analysis of gene expression (CAGE) in order to map the 5′-end of *de novo* synthesized RNA at promoter and enhancer regions in some 750 primary tissues

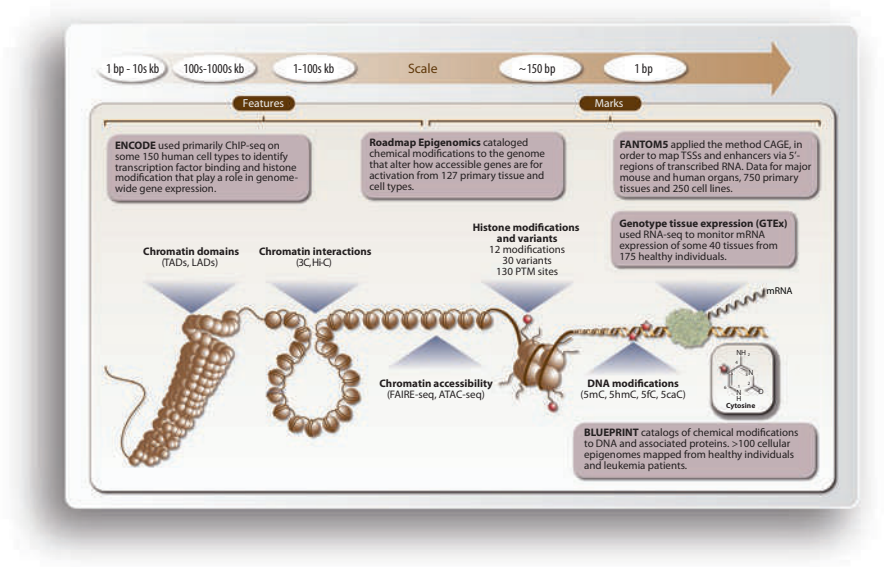


Fig. 1.7 **Epigenetic Big Biology projects**. Outline of the type of datasets collected by different Big Biology projects with an epigenetic focus. More details are provided in the text. bp, base pair; kb, kilo base pair; PTM, post-translational modification

and 250 cell lines. The *ENCODE* follow-up project *Roadmap Epigenomics* (http://www.roadmapepigenomics.org) provided human epigenome references from 111 primary human tissues and cell lines. Together with 16 samples that had already been provided by the *ENCODE Project*, 127 reference epigenomes were made publically available. Finally, *IHEC* (International Human Epigenome Consortium, http://ihec-epigenomes.org) became the umbrella organization under which national and international epigenome efforts are jointly coordinated. The key achievements of *IHEC* have been the introduction and the implementation of quality standards for harmonizing epigenomic data collection, management and analysis. For example, based on *IHEC* standards each reference epigenome needs to be composed of at least nine profiles and assays, but typically reference epigenomes are composed of 20–50 genome-wide profiles and represent a multi-dimensional data matrix.

Further Reading

Allis CD, Jenuwein T (2016) The molecular hallmarks of epigenetic control. Nat Rev Genet 17:487–500

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T et al (2014) An atlas of active enhancers across human cell types and tissues. Nature 507:455–461

Carlberg C, Molnár F (2018) Human Epigenomics. Springer Textbook *ISBN*, pp 978–981-10-7614-8

Deans C, Maggert KA (2015) What do you mean, “epigenetic”? Genetics 199:887–896
Encode-Project-Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74
Fantom-Consortium, Hoen PA, Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, Kulakovskiy IV et al (2014) A promoter-level mammalian expression atlas. Nature 507:462–470
Hasin Y, Seldin M, Lusis A (2017) Multi-omics approaches to disease. Genome Biol 18:83
Lappalainen T, Greally JM (2017) Associating cellular epigenetic models with human phenotypes. Nat Rev Genet 18:441–451
Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J et al (2015) Integrative analysis of 111 reference human epigenomes. Nature 518:317–330
Stricker SH, Koferle A, Beck S (2017) From profiles to function in epigenomics. Nat Rev Genet 18:51–66

Chromatin

2

Abstract

Nucleosomes are formed by histones, which can post-translationally become modified in more than 100 ways and in addition have many variant forms. The regulatory layers (*i.e.*, dimensions) of chromatin range from single nucleosomes *via* looping of genomic regions to large-scale folding of whole chromosomes into territories within the nucleus. The accessibility of the chromatin at enhancer and promoter regions determines whether a gene is expressed or not. Furthermore, chromatin architecture imposes which enhancers are able to regulate the expression of which genes. There are various epigenetic methods for the analysis of DNA methylation, transcription factor binding and histone modifications, chromatin accessibility as well as nuclear architecture. Using these approaches, Big Biology projects have already collected a large number of human epigenomes that are publically available.

Keywords

Chromatin · Histone proteins · Nucleosome · Histone variants · Post-translational histone modifications · Chromatin architecture · Enhancer · TSS · Gene expression · TADs · ChIP-seq · Single-cell assays · Genome browsers

2.1 Nucleosomes: Central Units of Chromatin

Due to its phosphate backbone genomic DNA is negatively charged (at physiologic pH). The electrostatic repulsion between adjacent DNA regions makes it impossible to fold the long DNA molecules (46 to 249 Mb, *i.e.*, 16 to 85 mm) of individual chromosomes into the limited space of the nucleus. Nature solved this problem by combining genomic DNA with histone proteins, which contain over proportional high amounts of the positively charged amino acids lysine (K) and arginine (R). Thus, histones are the main protein components of chromatin.

C. Carlberg, F. Molnár, *Human Epigenetics: How Science Works*,
https://doi.org/10.1007/978-3-030-22907-8_2

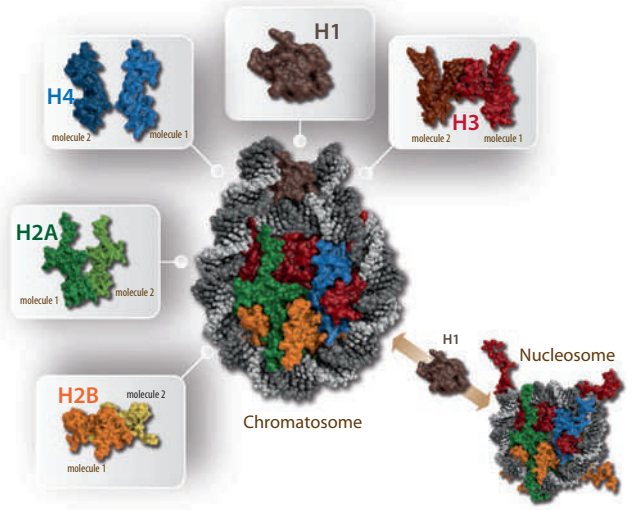


Fig. 2.1 **The nucleosome**. This space-filling surface representation of a nucleosome contains two copies each of the four core histone proteins H2A (green), H2B (orange), H3 (red) and H4 (blue) and 147 bp of genomic DNA (gray) being wrapped 1.8-times around the histone core. In complex with the linker histone H1 (brown) the nucleosome is referred to as chromatosome

Chromatin is the Physical Representation of the Epigenome Nucleosomes are the subunits of chromatin, and in every cell of our body our diploid genome is covered by approximately 30 million nucleosomes. Two copies each of the histone proteins H2A, H2B, H3 and H4, the so-called histone octamer, and 147 bp genomic DNA, which is wrapped nearly twice around the octamer, form the nucleosome (Fig. 2.1). The bending of genomic DNA is primarily enabled through the attraction between the positively charged histone tails and the negatively charged DNA backbone. In addition, at some genomic regions, the bending is supported by the natural curvature of DNA that is achieved by AA/TT dinucleotides repeating every 10 bp and a high CG content. Together with the linker histone H1 the nucleosome forms the so-called chromatosome. Each nucleosome is connected with the following one *via* linker DNA (20–80 bp). This forms a repetitive unit approximately every 200 bp of genomic DNA. The regular positioning of nucleosomes has the effect that the position of a given nucleosome determines the location of its neighbors. However, through the investment of ATP, *i.e.*, of energy, chromatin remodeling protein complexes are able modulate the position and composition of nucleosomes (Sect. 5.3).

The phosphate backbone of 200 bp genomic DNA carries 400 negative charges, which are in part neutralized by the approximately 220 positively charged lysine and arginine residues of the histone octamer. However, higher order folding of chromatin requires the neutralization of the remaining 180 negative charges by the

Box 2.1 Nomenclature of Histone Modifications
Histone modifications are named according to the following rule:

- the name of the histone protein (for example, H3)
- the single-letter amino acid abbreviation (for example, K for lysine) and the amino acid position in the protein
- the type of modification (ac: acetyl, me: methyl, P: phosphate, Ub: ubiquitin *etc*.)
- the number of modifications (only methylations are known to occur in more than one copy per residue, thus 1, 2 or 3 indicates mono-, di- or tri-methylation).

For example, H3K4me3 denotes the tri-methylation of the 4th residue (a lysine) from the amino-terminus of the protein histone 3. This type of histone modification serves as a general mark for active promoter regions.

positively charged linker histone H1 and also other positively charged nuclear proteins associating with chromatin.

Nucleosomes are the regularly repeating units of chromatin, but they can vary from one genomic region to the other by different post-translational modifications of the amino acid residues of their histones (Box 2.1, Sect. 4.1) and the introduction of histone variants (Box 4.1). These genomic locus-specific histone modifications are reversible and an important component of the epigenetic memory affecting transcription factor binding and differential gene expression between cell types. Thus, **nucleosomes are not simply barriers that block access to genomic DNA but serve as dynamic platforms linking and integrating many biological processes**, such as transcription and replication.

2.2 Chromatin Organization

Chromatin is organized into lower-order structures, such as the 11 nm fiber of euchromatin (Fig. 1.1, bottom) and higher-order structures, like the 30 nm fiber of heterochromatin or the 700 nm fibers of chromosomes. Accordingly, the most densely packaging of chromatin is found during the metaphase of mitosis, shortly before the chromosomes are distributed to both daughter cells. This phase has to be short, since at such dense chromatin packing there is no gene transcription possible, *i.e.*, no flexibility to respond to environmental signals. However, also during the interphase, 90% of the genomic DNA of terminally differentiated cells is not accessible to transcription factors. Thus, **heterochromatin is the default state of chromatin** and located preferentially within lamin-associated domains (LADs) close to the nuclear envelope (Fig. 2.2, bottom).

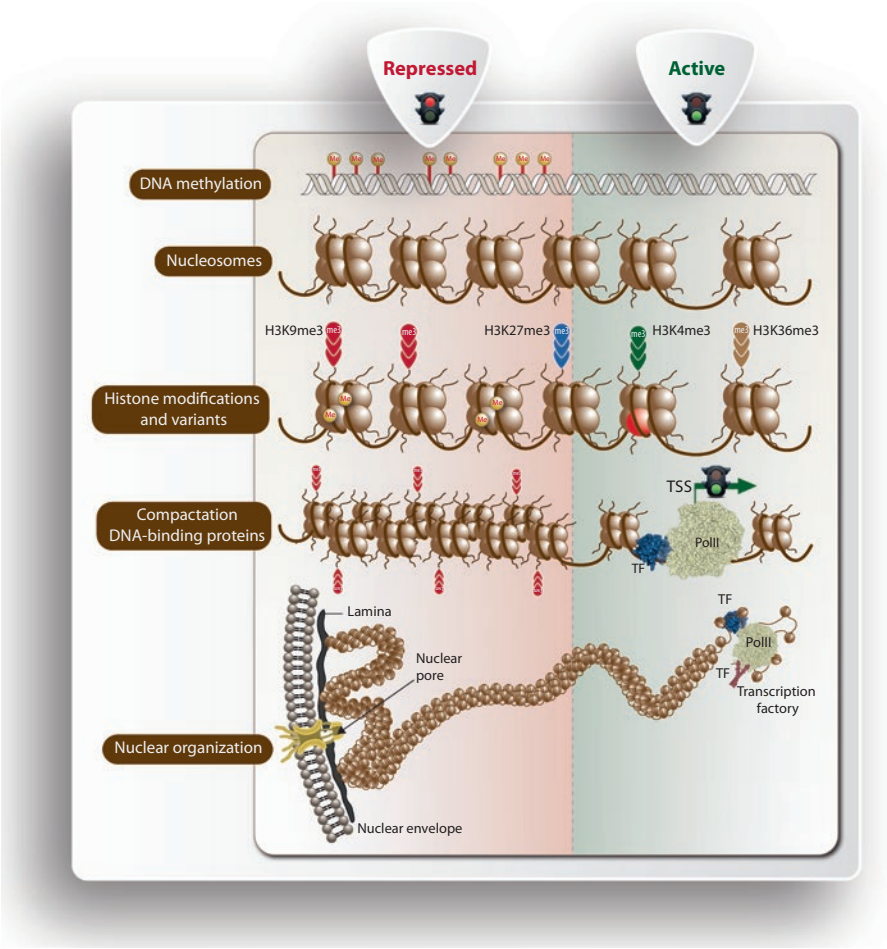


Fig. 2.2 **Epigenetic layers of chromatin organization**. There are at least five different layers of chromatin organization that are associated with inactive ("off", **left**) or active ("on", **right**) transcription
1: methylated versus unmethylated genomic DNA
2: nucleosome arrangement versus nucleosome-free regions
3: (exemplary) histone tri-methylation at positions H3K9 and H3K27 versus positions H3K4 and H3K36
4: dense nucleosome packaging in heterochromatin versus loose arrangement with transcription factors and Pol II binding in euchromatin
5: location within LADs close to the border of the nucleus versus at transcription factor assemblies, referred to as transcription factories, in the center of the nucleus.
LADs, lamin-associated domains

The structure and organization of chromatin can be interpreted as a number of superimposed epigenetic layers that lead either to open euchromatin and active gene expression ("on", Fig. 2.2, right) or to closed heterochromatin and no gene expression ("off", Fig. 2.2, left):

- The core of chromatin is genomic DNA that can be modified at cytosines, in particular at CG dinucleotides (CpGs, Sect. 3.1). Therefore, the first epigenetic layer is the DNA methylation status where hyper-methylation stimulates the formation of heterochromatin.
- The packaging of nucleosomes represents level 2 where more dense arrangements indicate heterochromatin.
- Histone modifications (Sect. 4.1) at specific positions are level 3 and mark for either active chromatin (mainly acetylated) or inactive chromatin (mainly methylated).
- The resulting accessibility of genomic DNA for the binding of transcription factors is considered as level 4.
- Finally, the complex formation and relative position of the chromatin, such as active transcription factories in the center of the nucleus and inactive chromatin in LADs attached to the nucleoskeleton at the nuclear periphery, represent level 5 (Sect. 2.4).

2.3 Epigenetics and Gene Expression

Chromatin acts as a filter for the access of DNA-binding proteins to functional elements of our genome, such as TSS regions, also referred to as core promoters, and enhancers. Genes can only be transcribed into mRNA, when their TSSs are accessible to the basal transcriptional machinery containing Pol II. However, even with given DNA access, mRNA transcription is often weak in the absence of stimulatory transcription factors (Fig. 2.3, top). Therefore, the second condition for efficient gene expression is that enhancer regions in relative vicinity to the TSS are not buried in heterochromatin and can be recognized by transcription factors. Thus, **in order to activate and transcribe a gene, the chromatin at both its TSS and at enhancer region(s) that control the gene's activity needs to be accessible**. Thus, in most cases, gene activation requires the transition from heterochromatin to euchromatin.

Enhancers are genomic regions that contain binding sites for sequence-specific transcription factors, which recruit co-activator and chromatin modifying proteins (Sect. 5.2) to the respective genomic loci. Thus, enhancers function *via* the cooperative binding of multiple proteins. Since this often happens in less than one nucleosome length, nucleosome eviction is not essential for enhancer function (Sect. 5.3). **Enhancer activity is determined by epigenome stages**, which is often recognized by histone markers of accessible chromatin, such as H3K4me1 and H3K27ac. When enhancers are close (+/− 100 bp) to the TSS, they are also often referred to as promoters. Thus, **there is no functional difference between enhancers and**

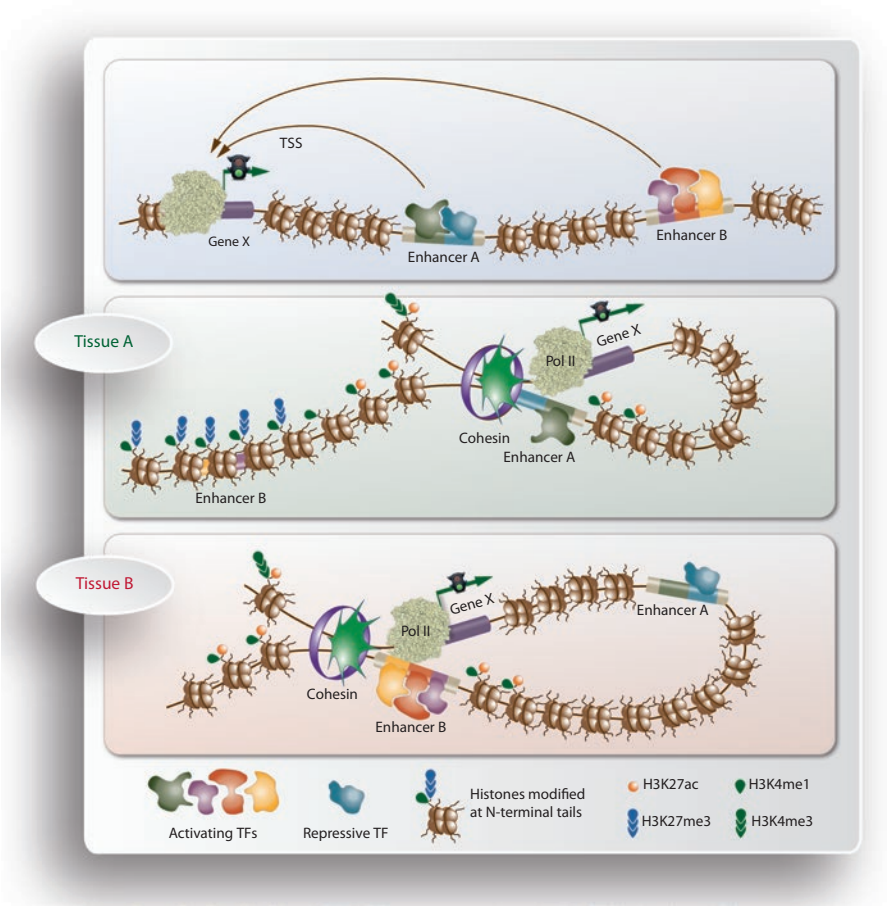


Fig. 2.3 **Enhancer function**. Enhancers are stretches of genomic DNA that contain binding sites for one or multiple transcription factors (TFs) stimulating the activity of the basal transcriptional machinery (Pol II and associated proteins) bound to the TSS of a target gene. Enhancers are located both upstream and downstream of their target genes in linear distances of up to 1 Mb (**top**). Transcription-factor-bound, active enhancers are brought into proximity of TSSs by DNA looping, which is mediated by a complex of cohesin, CTCF and other proteins. Active TSS regions and enhancers show depletion of nucleosomes while nucleosomes flanking active enhancers have specific histone modifications, such as H3K27ac and H3K4me1 (**center, tissue A**). In contrast, inactive enhancers are silenced by a number of mechanisms, such as repressing Polycomb proteins binding to H3K27me3 marks or by binding of repressive transcription factors (**bottom, tissue B**). More details are provided in the text.

promoters besides their distance relative to the TSS of the gene that they are regulating.

Enhancers that regulate the activity of a given gene should be located within the same topologically associated domain (TAD). Since TADs have an average size of 1 Mb (Sect. 2.4), this may be the maximal linear distance between an enhancer and the TSS(s) that it regulates (Fig. 2.3, top). Complexes of the proteins cohesin and CCCTC binding factor (CTCF) mediate these DNA looping events. These 3D structural arrangements bring transcription factors that bind to enhancers into close vicinity of TSS regions. In this way, transcription factors, which bind to distant enhancers, can contact and activate *via* intermediary complexes, such as Mediator, the basal transcriptional machinery. The looping mechanism also implies that enhancer regions are as likely upstream as downstream of TSS regions and may have tissue-specific usage and effects for transcription. For example, in tissue A enhancer A is used for activation, whereas in tissue B it mediates repression (Fig. 2.3, center and bottom). Results of the *ENCODE Project* (Sect. 1.3) demonstrated that basically all regulatory proteins have a Gaussian-type distribution pattern in relation to TSS regions, *i.e.*, the probability to find an active transcription factor binding site symmetrically declines both up- and downstream of the TSS. Thus, **the classical definition of a promoter as a sequence being located only upstream of the TSS is outdated**.

2.4 Chromatin Architecture: Epigenetics in 3D

The probability that two regions of a chromosome contact each other by chance *via* DNA looping rapidly decreases with the increase of their linear distance. However, when the contact between the two regions is stabilized, for example, by associated proteins, then architectural loops and regulatory loops are forming (Fig. 2.4). Most architectural loops are identical to TADs (also sometimes referred to as insulated neighborhoods), since they are anchored by CTCF-CTCF homodimers in complex with cohesin and carry at least one gene. Thus, **TADs are the units of chromosomal organization and segregate our genome into at least 2000 domains containing co-regulated genes**. Often TAD boundaries are identical with insulators (Sect. 3.3) and are bound by CTCF. Thus insulators are stretches of genomic DNA that separate functionally distinct regions of the genome from each other. Accordingly, neighboring TADs can differ significantly in their histone modification pattern, such as one TAD being in heterochromatic state containing silent genes and the other TAD being in euchromatin carrying transcriptionally active genes.

TADs are separated by boundaries for self-interacting chromatin and thus organize regulatory landscapes, *i.e.*, they define the genomic regions, in which enhancers can interact with TSS regions of their target gene(s) (Sect. 2.3). The linear size of TADs is in the range of 100 kb to 5 (median: 1) Mb and TADs contain 1–10 (median: 3) genes. Accordingly, most TADs contain a number of genes that may be regulated by the same set of enhancers, such as often observed for gene clusters. Regulatory loops are formed between enhancers and TSS regions that are located

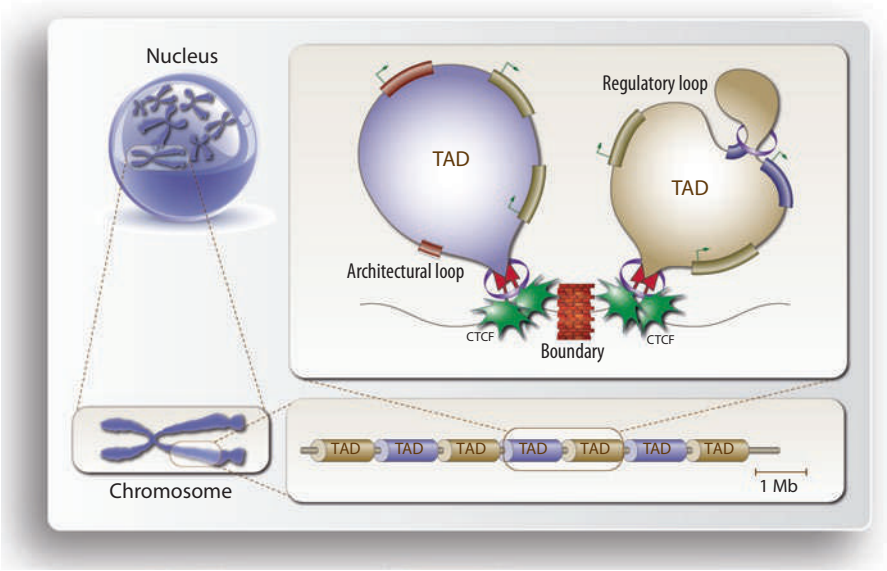


Fig. 2.4 **Organization of chromosomes into TADs**. Our genome is subdivided into a few thousand TADs defining genomic regions in which most genes have their specific regulatory elements, such as promoters and enhancers. TADs are architectural loops of chromatin that are insulated from each other by anchor regions binding complexes of CTCF and cohesin. Within TADs smaller regulatory loops between enhancers and promoters are formed.

within the same TAD, *i.e.*, they are smaller than TADs (Fig. 2.4). The formation of regulatory loops relies on the binding of transcription factors to the enhancer regions and its functional result is the stimulation of gene expression (Fig. 2.3).

The inner surface of the nuclear envelope is coated with nuclear lamina, which is a complex of lamins and a number of additional proteins (Fig. 2.5). Lamins maintain the shape and mechanical properties of the nucleus and serve as attachment points for LADs (Sect. 2.2). **LAD-lamin interactions form a nucleoskeleton**, *i.e.*, they serve as a structural backbone for the organization of interphase chromosomes. LADs vary in size from 0.1 to 10 Mb, cover up to 40% of our genome and are primarily formed by heterochromatin. LADs have a low gene density, but in total they still contain thousands of genes, most of which are not expressed. Accordingly, **the nuclear periphery is enriched for heterochromatin, whereas euchromatin is found more likely in the center of the nucleus** (Fig. 1.1). This suggests that the location of a gene within the nucleus is a functionally important epigenetic parameter.

The clustering of heterochromatin at the nuclear periphery creates silencing foci, so-called Polycomb bodies. These are complexes of members of the Polycomb family, such as the components of Polycomb repressive complex (PRC) 1 and 2 (Sect. 5.4). PRCs act as transcriptional repressors that are essential for maintaining

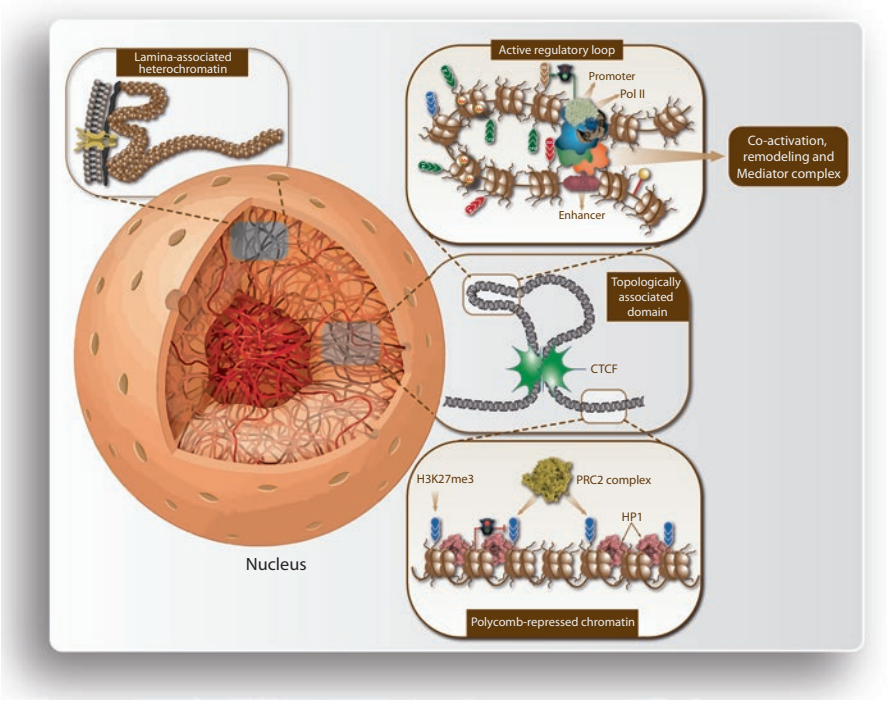


Fig. 2.5 **Chromatin architecture**. Mediated by structural proteins, chromatin within forms a 3D architecture the nucleus (**center left**). Heterochromatin is composed of stably repressed, inaccessible genomic elements and is located closer to the nuclear lamina (**top left**). Two CTCF proteins bound at adjacent chromatin boundaries form a complex with cohesin and other mediator proteins (**center right**). In this way, regulatory genomic regions, such as enhancers and promoters, which are separated by a genomic distance, can get into physical contact within DNA loops (**top right**, see also Fig. 2.3). TADs distinguish such genomic regions with active enhancers from chromatin tracts that are silenced by PRCs (**bottom right**)

tissue-specific gene expression programs, *i.e.*, they ensure the long-term repression of specific target genes.

The position of chromatin, and with that the position of genes, is not fixed, but there are dynamic changes in the contacts between the nucleoskeleton and genomic DNA involving single genes or small gene clusters. These changes are most pronounced during development. Of all human cell types, ES cells have the most accessible genome, *i.e.*, the chromatin of these cells is largely open. During the differentiation process, cells change their chromatin structure and larger compaction of their genome occurs. Thus, embryonic development proceeds from a single cell with dispersed chromatin to differentiated cells with nuclei that show compact chromatin domains being located in the periphery (Sect. 6.1). Accordingly, the physical relocation of a gene from the nuclear periphery to the center would unlock it to be expressed in a future developmental stage.

Another level of chromatin architecture in the interphase nucleus is the location of whole chromosomes in separate chromosome territories, which are separated by an inter-chromosomal compartment. Chromosomes fold in their territories in such a way that active and inactive TADs are found in distinct nuclear compartments. Active regions are preferentially located in the nuclear interior, whereas inactive TADs accumulate at the periphery. In addition, TADs that are heavily bound by tissue-specific transcription factors are in different neighborhoods than those interacting with repressive PRCs (Fig. 2.5, right). To some extent, chromosome territories intermingle, which could explain inter-chromosomal interactions. Nevertheless, interactions between loci on the same chromosome are much more frequent than contacts between different chromosomes. Since the volumes of chromosome territories depend on the linear density of active genes on each chromosome, **chromatin with higher transcriptional activity occupies larger volumes in the nucleus than silent chromatin**.

2.5 Epigenetic Methods

The rapid maturation of next-generation sequencing technologies led to the exponential development of methods in the field of epigenetics/epigenomics. These methods investigate various aspects of chromatin biology, such as DNA methylation, histone modification state, and 3D chromatin structure (Fig. 2.6). In general, next-generation sequencing methods have the advantage that they provide, in an unbiased and comprehensive fashion, information on the entire epigenome. Thus, conclusions drawn from a few isolated genomic regions might be extended to other parts of the genome. Global epigenomic profiling allows hypothesis-free exploration of new observations and correlations. Individual research teams as well as large consortia, such as the *ENCODE Project* and the *Roadmap Epigenomics Project* (Sect. 1.3), have already produced thousands of epigenome maps from hundreds of human tissues and cell types. The integration of these data, for example, transcription factor binding and characteristic histone modifications, allows the prediction of enhancer and promoter regions as well as monitoring their activity and many additional functional aspects of the epigenome.

The key epigenetic methods determine DNA methylation, transcription factor binding and histone modification, accessible chromatin, and 3D chromatin architecture. The biochemical cores of these methods are

- different chemical susceptibility of nucleotides, such as bisulfite treatment of genomic DNA, in order to distinguish between cytosine and 5-methylcytosine (5mC, Sect. 3.1)
- affinity of specific antibodies for chromatin-associated proteins, such as transcription factors, modified histones and chromatin modifiers
- endonuclease-susceptibility of genomic DNA within open chromatin compared to inert closed chromatin

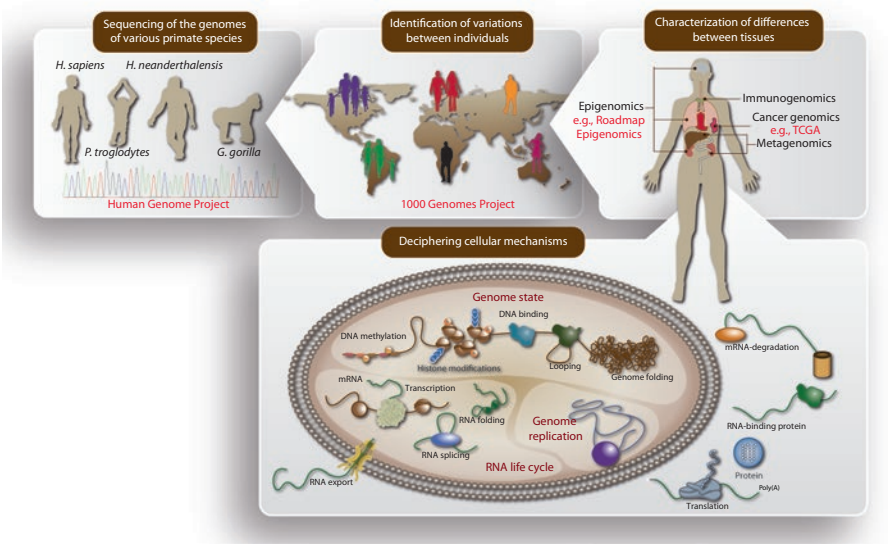


Fig. 2.6 **Roadmap of sequencing science**. The *Human Genome Project* (Box 1.1) created a reference genome. Nowadays, also the genomes of all other primate species are known including some extinct human species (**top left**). Whole genome sequencing of several thousand individuals is performed in large consortia, such as the *1000 Genomes Project* (**top center**). Moreover, the genetic and epigenetic differences between tissues and cell types of the same individual are collected in cancer genomics and epigenomics projects, such as *The Cancer Genome Atlas* (*TCGA*, Sect. 8.1) and the *Roadmap Epigenomics Project* (**top right**). The application of different next-generation sequencing methods, such as ChIP-seq, RNA-seq or ATAC-seq, allows integrating many different processes within the cell (**bottom**)

- physical separation of protein-associated genomic DNA (of fragmented closed chromatin) in an organic phase from free DNA (of accessible chromatin) in the aqueous phase
- proximity ligation of genomic DNA fragments that *via* looping got into close physical contact.

Chromatin immunoprecipitation (ChIP) followed by sequencing (ChIP-seq) maps the genome-wide binding pattern of chromatin-associated proteins, such as transcription factors and chromatin modifiers, including post-translationally modified histones. The core of this method is the immunoprecipitation of cross-linked protein-DNA complexes from sonicated chromatin with an antibody that is specific for the protein of interest (Fig. 2.7). Genomic DNA fragments within these complexes are purified from the enriched pool and sequenced by using massive parallel sequencing. Sequence tags representing the DNA fragments are aligned to the reference genome. Assemblies of the tags are called "peaks" and indicate genomic regions where the protein of interest was binding at the moment of cross-linking. ChIP-seq of histone modifications tends to produce broader peaks, *i.e.*, more diffuse

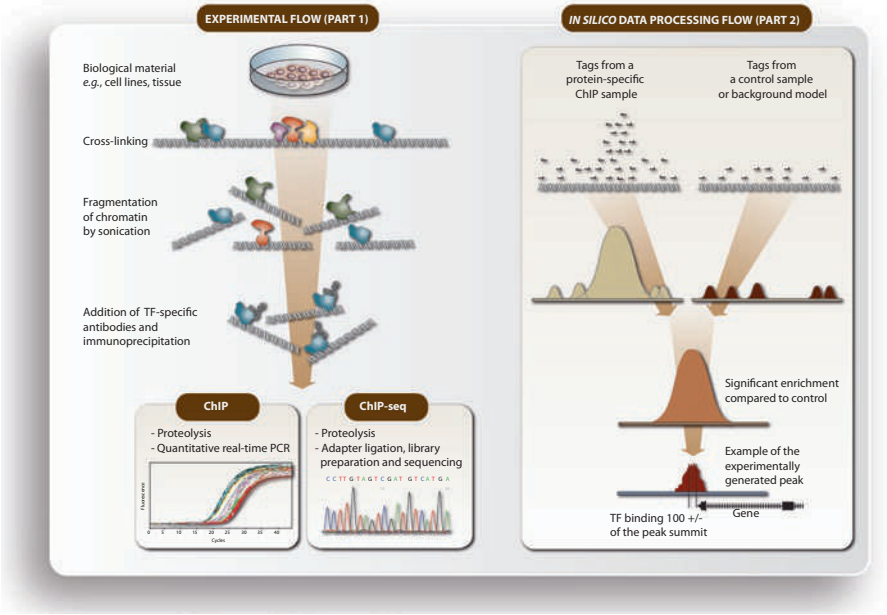


Fig. 2.7 **ChIP-seq and its analysis**. Short chromatin fragments are prepared from cells in which nuclear proteins are covalently attached to genomic DNA by short-term formaldehyde cross-linking. After chromatin fragmentation, immobilized antibodies against a protein of interest, such as a transcription factor (TF) or a histone mark, are used to immunoprecipitate the chromatin fragments associated with the respective protein (**left**). All genomic fragments are subjected to massive parallel sequencing, for example, by the use of an Illumina Genome Analyzer. Typically, sequencing runs provide tens of millions of sequencing tags (small arrows) that are uniquely aligned to the reference genome (**right**). Clusters of these tags form peaks that represent transcription factor binding loci when they show significantly higher binding than the control sample

regions of enrichment, than transcription factors that bind sequence-specifically and create sharper peak profiles.

Although ChIP-seq is a mature method, it is restricted by the need for large amounts of starting material (1–20 million cells), limited resolution, and the dependence on the quality of the applied antibodies. A new variation of ChIP-seq, ChIPmentation, takes advantage of a library preparation using the Tn5 transposase ("tagmentation") as in the method assay for transposase-accessible chromatin using sequencing (ATAC-seq). The sequencing library is prepared using fragmented and immunoprecipitated chromatin instead of the standard purified, *i.e.*, protein-free, immunoprecipitated genomic DNA. This tagmentation step reduces the number of cells needed in the experiments by a factor of 10–100. Finally, single-cell approaches (Box 2.2) allow even more powerful analyses of chromatin states and their associated gene regulatory networks.

In the past, the main assays for genome-wide mapping of open chromatin were DNase I hyper-sensitivity followed by sequencing (DNase-seq) and formaldehyde-assisted identification of regulatory elements followed by sequencing (FAIRE-seq).

Box 2.2 Single-Cell Analyses

In the past, epigenome methods were performed with larger numbers of cells. Since cell populations are known to be heterogenous, the respective results represent the average chromatin state for thousands or even millions of cells. Recent technological advances allowed executing genome-wide analyses on single cells. For example, single-cell RNA-seq showed substantial heterogeneity of cell types in various tissues and identified novel cell populations. The single-cell technology has been extended to the genome and DNA methylome. Bisulfite sequencing of single cells indicates substantial variations in DNA methylation patterns across otherwise homologous cells residing in the same tissues. Recently, single-cell ATAC-seq was developed, which allows single-cell analyses of chromatin accessibility. In general, single-cell epigenomics will provide insights into the combinatorial nature of chromatin, such as which combinations of epigenetic marks and structures are possible and what mechanisms control them.

The core for the DNase-seq method is a limited digestion of chromatin with the endonuclease DNase I that releases nucleosome-depleted fragments of genomic DNA. The FAIRE-seq assay takes advantage of the fact that genomic DNA within open chromatin regions is particularly sensitive to shearing by sonication. Chromatin is isolated from formaldehyde cross-linked cells, sonicated, and subjected to a phenol-chloroform extraction. Protein-free genomic DNA can be isolated from the aqueous phase, while protein-bound DNA remains in the organic phase. However, the present state-of-the-art method for mapping chromatin accessibility is ATAC-seq, which uses the Tn5 transposase in the sequencing library preparation. Like in ChIPmentation, the use of tagmentation largely reduces the number of cells needed per experiment.

For assessing the 3D organization of chromatin, chromosome conformation capture (3C)-based methods are used that combine protein cross-linking and proximity ligation of DNA in order to detect long-range chromatin interactions. These methods quantify the interaction frequency between genomic loci that are close to each other in 3D but may be separated by thousands of bases in the linear genome. This identifies loops of genomic DNA, for example, between promoter and enhancer regions. The genome-wide version of 3C, high-throughput chromosome capture (Hi-C) uses sequencing of all ligated fragments instead of PCR targeted to specific sites.

Epigenome profiling leads to maps of DNA methylation, histone marks, DNA accessibility, and DNA looping that can be visualized with an appropriate web browser, such as the UCSC Genome Browser (Box 2.3). Although visualization can be highly illustrative and may induce hypotheses, epigenome maps are primarily descriptive, *i.e.*, they are used for annotation. Enhancers, promoters, and other genomic features have characteristic epigenomic signatures, such as H3K4me1 marks for enhancers and H3K4me3 marks for promoters (Sect. 4.2), on the basis of which they can be identified within epigenome maps.

Box 2.3 Visualizing Epigenomic Data

A typical way of visualizing epigenomic data, such as those from the *ENCODE Project*, is to display a selected subset of them in a browser, like the UCSC Genome Browser (http://genome.ucsc.edu/ENCODE). Datasets can be inspected without downloading them by creating a dynamic UCSC Genome Browser track hub that can be visualized on a local mirror of the UCSC Browser. Other visualization tools supporting the track hub format, such as Ensembl (www.ensembl.org), can also be used. For every given genomic position, a graphical display provides an intuitively understandable description of chromatin features, such as histone acetylation and methylation, that can be read in combination with experimentally proven information about transcription factor binding, as obtained from ChIP-seq experiments.

There are a variety of approaches to integrate epigenomics data within or across omics layers, such as correlation or co-mapping. When the datasets, which are to be compared, have a common driver, or if one regulates the other, correlations or associations should be observed. This requires the application of appropriate statistical methods, many of which have been recently developed for the omics field. In most cases, more than two datasets, which often derive from different omics layers, are integrated (often referred to as modeled) in gene regulatory networks.

Further Reading

Bonev B, Cavalli G (2016) Organization and function of the 3D genome. Nat Rev. Genet 17:661–678

Buschbeck M, Hake SB (2017) Variants of core histones and their roles in cell fate decisions, development and cancer. Nat Rev. Mol Cell Biol 18:299–314

Carlberg C, Molnár F (2016) Mechanisms of gene regulation. Springer Textbook ISBN: 978-94-007-7904-4

Klemm SL, Shipony Z, Greenleaf WJ (2019) Chromatin accessibility and the regulatory epigenome. Nat Rev. Genet 20:207–220

Lappalainen T, Scott AJ, Brandt M, Hall IM (2019) Genomic analysis in the age of human genome sequencing. Cell 177:70–84

Pombo A, Dillon N (2015) Three-dimensional genome architecture: players and mechanisms. Nat Rev. Mol Cell Biol 16:245–257

Schmitt AD, Hu M, Ren B (2016) Genome-wide mapping and analysis of chromosome architecture. Nat Rev. Mol Cell Biol 17:743–755

Stricker SH, Köferle A, Beck S (2017) From profiles to function in epigenomics. Nat Rev. Genet 18:51–66

Talbert PB, Henikoff S (2017) Histone variants on the move: substrates for chromatin dynamics. Nat Rev. Mol Cell Biol 18:115–126

Zhou K, Gaullier G, Luger K (2019) Nucleosome structure and dynamics are coming of age. Nat Struct Mol Biol 26:3–13

DNA Methylation

3

Abstract

Cytosine methylation of genomic DNA is the best-understood epigenetic mark. In most cases, DNA methylation leads to the formation of heterochromatin and subsequent gene silencing. Coordinated DNA methylation and its recognition *via* methylation-sensitive DNA-binding proteins have a large impact on health, such as genetic imprinting, *i.e.*, the expression of a gene in a parent-of-origin-specific manner. In contrast, a disorganized DNA methylome contributes to tumorigenesis and thus to cancer.

Keywords

DNA methylation · CpG islands · DNA methyltransferase · TET proteins · 5mC modifications · Gene silencing · Insulator · CTCF · Genetic imprinting · X chromosome inactivation · *Xist* · Imprinting disorders · Hyper-methylation

3.1 Cytosines and Their Methylation

The identity of each of the 400 human tissues and cell types is based on their respective unique gene expression patterns, which in turn are determined by differences in their epigenomes. **For the proper function of our tissues, it is essential that cells memorize their respective epigenetic status** and pass it to daughter cells when they are proliferating. The main mechanism for this long-term epigenetic memory is the methylation of genomic DNA at the fifth position of cytosine (5mC) (Fig. 3.1, bottom left).

DNA methylation is often associated with transcriptional silencing of repetitive DNA (Box 1.1) and genes that are not needed in a specific cell type. CG dinucleotides (CpGs) are the only cytosine-containing dinucleotides that are symmetrically methylated on both the top and bottom DNA strand. Therefore, only the methylation pattern of CpGs persists during DNA replication and can be inherited to both

C. Carlberg, F. Molnár, *Human Epigenetics: How Science Works*,
https://doi.org/10.1007/978-3-030-22907-8_3

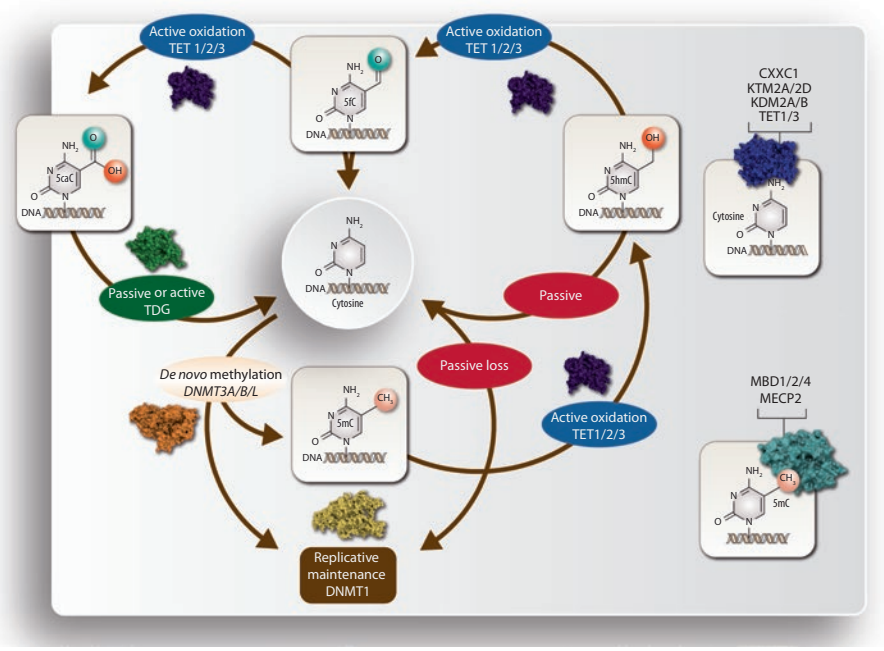


Fig. 3.1 **Writing, erasing, and reading cytosine methylations**. DNMT1, as well as DNMT3A and DNMT3B, catalyze the methylation of cytosines at position 5, *i.e.*, they act as "writer"-type chromatin modifiers (**left**). The dioxygenase enzymes, TET1, TET2, and TET3, oxidize 5mC to 5hmC and further to 5fC and 5caC, which leads *via* the action of the DNA glycosylase TDG to the loss of DNA methylation; *i.e.*, both types of enzymes function as "erasers" (**center**). Different sets of proteins either specifically recognize unmethylated cytosines or 5mC, *i.e.,* they are "readers" (**right**)

daughter cells. Nevertheless, also **the methylation of CpH dinucleotides (H = A, C, or T) contributes to the epigenetic memory of somatic cells**, but it does not get inherited (Sect. 9.2). Non-CpG methylation (mCH) occurs in all human tissues, but it is most common in long-lived cell types, such as stem cells and neurons. Proteins that specifically bind methylated DNA, such as MECP2 (methyl-CpG-binding protein 2), do not only interact with sites of methylated CpGs but also with methylated CpH loci.

CpG islands are defined as genomic regions of at least 200 bp in length showing a CG percentage higher than 55%. Due to these conditions only a minority (10%) of all CpGs belong to CpG islands. Our genome contains approximately 28,000 CpG islands, and many of our 20,000 protein-coding genes have such a region close to their TSS (CpG-rich promoters). In fact, genes are distinguished into those with and without CpG islands in the vicinity of their promoters. Interestingly, actively transcribed gene bodies carry both 5mC and 5-hydroxymethylcytosine (5hmC) marks, whereas active promoters are unmethylated (Sect. 3.2).

DNMTs are chromatin modifying enzymes that catalyze in a one-step reaction the transfer of a methyl group from S-adenosyl-L-methionine (SAM) to cytosines of genomic DNA (Fig. 3.1). Since **the DNA methylation pattern of somatic cells represents an epigenetic program of global repression of the genome** and specific settings of imprinted genes (Sect. 3.3), it is important to maintain the DNA methylome during replication. This is the main responsibility of DNMT1 in collaboration with its partner UHRF1 (ubiquitin-like plant homeodomain and RING finger domain 1) that preferentially recognizes hemi-methylated CpGs. In contrast, in the absence of functional DNMT1/UHRF1 complex, successive cycles of DNA replication lead to passive loss of 5mC, such as the global erasure of 5mC in the maternal genome during pre-implantation (Sect. 6.1). In particular, during the development of primordial germ cells (PGC), genomic DNA is widely demethylated. This creates pluripotent states in early embryos and erases most of the parental-origin-specific imprints in developing PGCs. With the exception of imprinted genomic regions (Sect. 3.3), DNMT3A and DNMT3B perform *de novo* DNA methylation during early embryogenesis, *i.e.*, together with DNMT1 they act as writers of DNA methylation (Fig. 3.1 left). Interestingly, the first approved epigenetic drug, decitabine (5-aza, 2′-deoxy-cytidine), is used for the therapy of leukemia and other forms of blood cancer in which hematopoietic progenitor cells do not maturate. Decitabine blocks DNA methylation *via* the inhibition of DNMTs (Sect. 8.5).

Active demethylation of genomic DNA is a multi-step process that involves the methylcytosine dioxygenase enzymes TET (ten-eleven translocation) 1, 2, and 3, which convert 5mC to 5hmC (Fig. 3.1). 5hmC is found in most cell-types but only in levels of 1–5% compared to 5mC rates. However, adult neurons are an exception since their 5hmC level is 15–40% of that of 5mC (Sect. 9.2). In two further oxidation steps TETs convert 5hmC into 5-formylcytosine (5fC) and to 5-carboxylcytosine (5caC). 5fC and 5caC are significantly less prevalent (0.06–0.6% and 0.01% of 5mC rates, respectively) than 5hmC, *i.e.*, TETs tend to preferentially halt at the 5hmC stage. Oxidized cytosines are deaminated to 5-hydroxyuracil (5hmU) so that they create a 5hmU:G mismatch, which is recognized and removed by the enzyme TDG (thymine-DNA glycosylase) (Fig. 3.1 top). The abasic site is then repaired by the base excision repair machinery, *i.e.*, by a regular DNA repair process, and results in the overall demethylation of the respective cytosine. **The oxidative modification of 5mC *via* the TET/TDG pathway allows a dynamic regulation of DNA methylation patterns**. DNA-binding proteins that specifically recognize either unmethylated or methylated genomic DNA (Fig. 3.1 right) then read the information stored in DNA methylation patterns and translate them into biological actions.

3.2 The DNA Methylome

The DNA methylome, *i.e.*, a genome-wide map of 5mC patterns and its oxidized modifications, is an essential component of the epigenome. Many datasets of different human tissues and cell types are already publicly available *via* the *IHEC* consortium (Sect. 1.3). Global DNA methylation methods measure cytosine

methylation at base resolution over the whole human genome. These are either affinity-based methods, such as methylated DNA immunoprecipitation sequencing (MeDIP-seq), or base-resolution mapping methods, such as bisulfite sequencing (Fig. 3.2). Both types of methods consistently identify genomic regions in which cytosines are frequently modified. Affinity-based assays use an antibody that enriches methylated fragments from sonicated genomic DNA. Like in ChIP-seq (Fig. 2.8) the resolution of these assays is highly dependent on the DNA fragment size and CpG density; *i.e.*, they are more qualitative than quantitative. In contrast, bisulfite sequencing directly determines the methylation state of each cytosine of the whole genome. In this chemical conversion method, sodium bisulfite treatment of genomic DNA chemically converts unmethylated cytosines to uracils. Through PCR amplification all unmethylated cytosines become thymidines, *i.e.*, the remaining cytosines correspond to 5mC. Whole genome sequencing provides then single base resolution of the methylation pattern. Interestingly, bisulfite sequencing is one of the first epigenomic methods being successfully applied on the single-cell level (Box 2.3). Since 5mC and 5hmC, but not 5fC or 5caC, are resistant to bisulfite conversion, they cannot be distinguished from each other; *i.e.*, antibody enrichment or more advance chemical conversion methods need to be applied.

The average CG base pair percentage of our genome is 42%; *i.e.*, the haploid genome sequence is composed of some 700 million cytosines. In principle, each of them can be methylated, but those of CpGs are functionally most important. Our genome contains some 28 million CpGs, the majority (70–80%) of which are methylated across all tissues and cell types. Most of the CpGs are located within regions of repetitive genomic DNA, such as SINEs, LINEs, and LTRs (Box 1.1). LINEs and LTRs carry strong promoters that must be constitutively silenced *via* placing them into constitutive heterochromatin in order to prevent their activity (Fig. 3.3a). Therefore, these genomic regions are generally hyper-methylated. The silencing of the repetitive DNA happens primarily during early embryogenesis (Sect. 6.1), while in adult tissues *de novo* silencing is initiated by the methyl-binding domain (MBD) proteins MECP2, MBD1, MBD3, and MBD4. These proteins bind symmetrically methylated CpGs, but they have no sequence specificity; *i.e.*, MBD proteins are not classical transcription factors but act as readers (Fig. 3.1 right) and adaptors for the recruitment of chromatin modifiers, such as HDACs and KMTs, to methylated genomic DNA.

Our DNA methylome is bimodal (*i.e.*, it occurs in two major modes): it has a low methylation level at CpG-rich promoters and binding sites for methylation-sensitive transcription factors, such as CTCF, while the remaining CpGs are by default methylated. Methylated genomic DNA is transcriptionally repressed (Fig. 1.2); *i.e.*, in most cases there is an inverse correlation between DNA methylation of regulatory genomic regions, such as promoters and enhancers, and the expression of the genes that they are controlling. However, at their gene bodies, highly expressed genes show high levels of DNA methylation; *i.e.*, some methylated CpGs downstream of TSS regions positively correlate with gene expression (Fig. 3.3b). Genes driven by CpG-rich promoters are silenced when methylated (Fig. 3.3c), while genes without

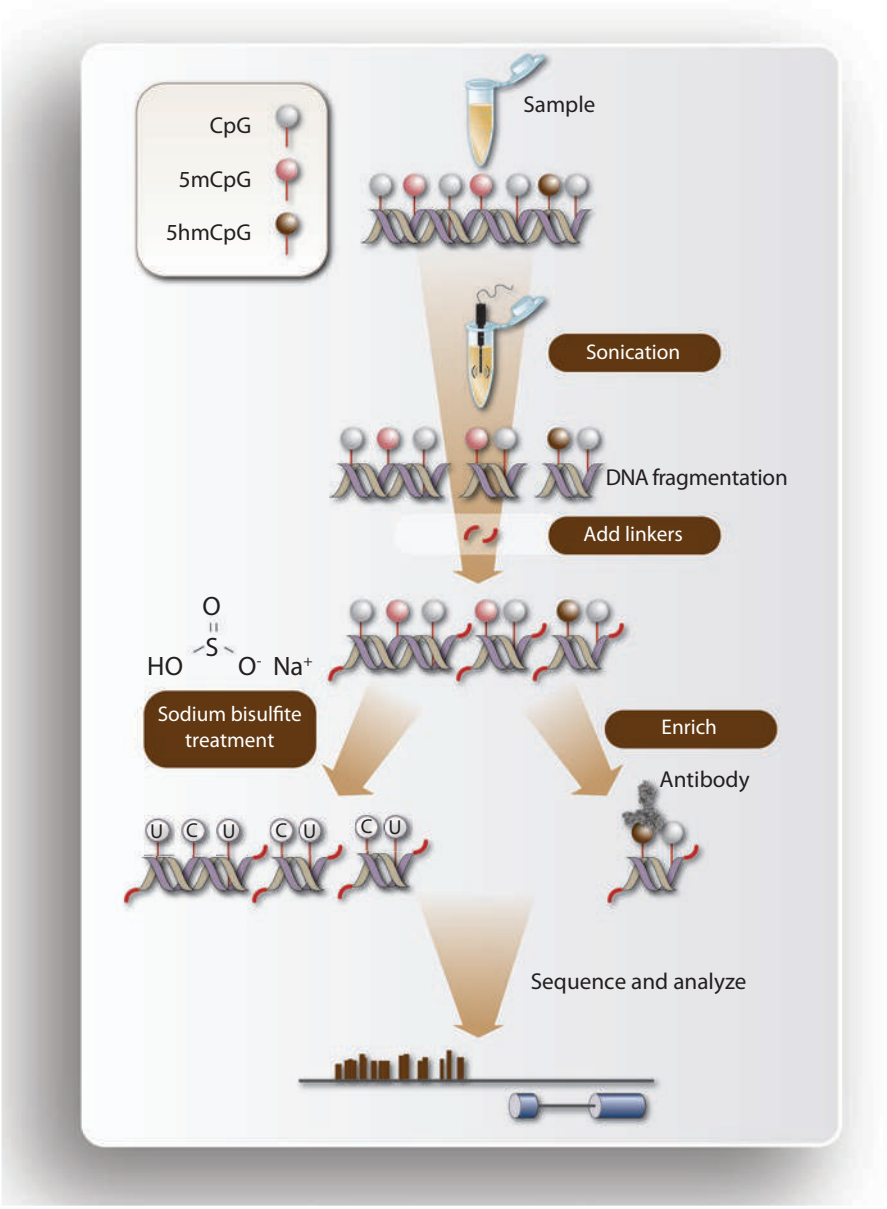


Fig. 3.2 **Analysis of DNA methylation**. Genome-wide 5mC and its oxidative derivative 5hmC are measured by conversion- and enrichment-based methods followed by sequencing. Bisulfite conversion allows quantification of 5mC but does not distinguish 5mC from 5hmC, while antibody enrichment enables qualitative measurement of 5mC and 5hmC. Bisulfite-converted or -enriched genomic DNA is purified, subjected to library construction, and clonally sequenced. Finally, the sequencing reads are aligned to the reference genome

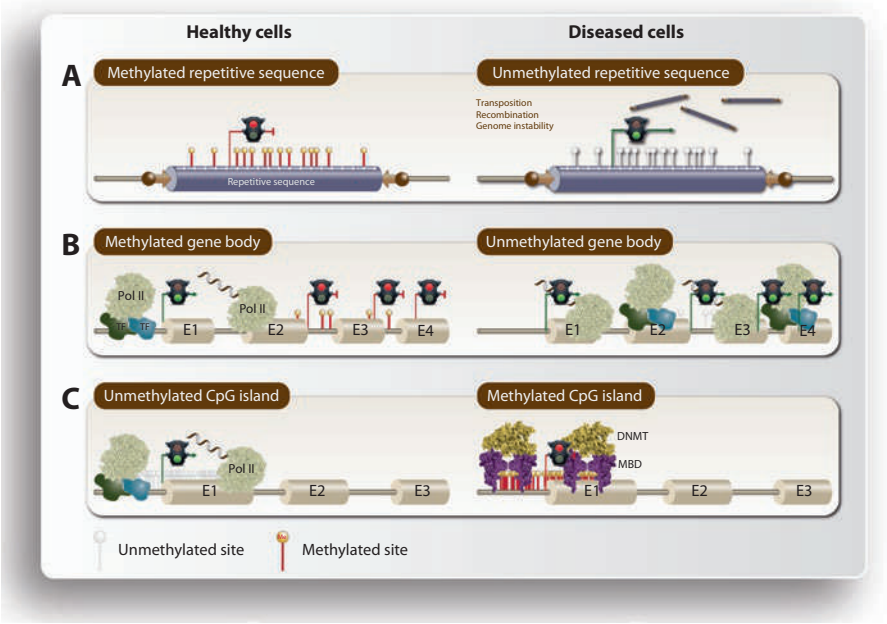


Fig. 3.3 **DNA methylation in different regions of the genome**. Scenarios of DNA methylation of healthy cells (**left**) or diseased cells (**right**) are displayed. (**a**) Repetitive sequences within our genome are normally hyper-methylated in order to prevent translocations, gene disruptions, and general chromosomal instability through the reactivation of retrotransposons. This pattern is altered in disease. (**b**) Methylation of the transcribed region of a gene facilitates transcription (traffic lights) by the prevention of spurious transcription initiations. Gene body tends to get demethylated in disease so that transcription may be initiated at several incorrect sites. (**c**) CpG islands close to TSS regions are normally unmethylated. This allows transcription, while hyper-methylation causes transcriptional inactivation

CpG islands close to their TSS regions are regulated by other mechanisms than DNA methylation, such as transcription factors binding to enhancers (Sect. 2.3).

In general, DNA methylation and histone modifications have different roles in gene silencing. **While most DNA methylation loci represent very stable silencing marks that are seldom reversed, histone modifications mostly lead to labile and reversible transcriptional repression** (Sect. 4.2). For example, genes for pluripotency transcription factors in embryogenesis (Sect. 6.2), such as *POU5F1* (also called *OCT4*) and *NANOG,* need to be permanently inactivated in later developmental stages in order to prevent possible tumorigenesis. This happens *via* H3K9 methylation at unmethylated CpGs on TSS regions of these genes, the attraction of heterochromatin protein 1 (HP1), *de novo* DNA methylation *via* DNMT3A and DNMT3B, and finally transcriptional silencing for the rest of the life of the individual. In contrast, when in differentiated cells these pluripotency genes are silenced only by histone modification, these cells can be rather easily converted to iPS cells.

Nevertheless, **the methylation status of some 20% of all CpGs within our genome is dynamically modified**. Differential DNA methylation is established by *de novo* methylation combined with active demethylation of CpG islands. During early embryogenesis, *i.e.*, in the pre-implantation phase, most CpGs are unmethylated (Sect. 6.1). After implantation DNMT3A and DNMT3B *de novo* methylate those CpGs that had not been packed with H3K4me3-marked nucleosomes. In contrast, H3K4me3-marked CpGs on TSS regions of CpG-rich promoters stay unmethylated.

Methylation and demethylation of CpGs modulate the DNA-binding affinity of transcription factors; *i.e.*, DNA methylation is a signal being differentially recognized by specific protein domains. Interestingly, a third of all of the some 1600 human transcription factors are positively affected by methylation of their DNA binding sites; half of all do not bind DNA when it is unmethylated, and only a fourth of all are negatively influenced by DNA methylation. A well-known example of the latter is CTCF in the context of genomic imprinting (Sect. 3.3). Taken together, **there are different forms of gene silencing ranging from flexible repressor-based mechanisms to a highly stable inactive state being maintained by DNA methylation**.

3.3 Genetic Imprinting

Insulators are genomic loci that separate genes located in one chromatin region from promiscuous regulation by transcription factors binding to enhancers of neighboring chromatin regions. Sect. 2.4 already presented the methylation-sensitive transcription factor CTCF as the main protein binding to insulator regions. In complex with other proteins, such as cohesin, CTCF mediates the formation of architectural loops, such as TADs, as well as of regulatory loops (Figs. 2.3, 2.4 and 2.5). In addition to the prevention of cross-border enhancer activity, insulators can act as boundary elements that inhibit spreading of heterochromatin from silenced genomic regions to transcriptionally active parts of the genome. This means that these boundary elements "insulate" closed from open chromatin, *i.e.*, inactive from active genes. Thus, **CTCF-bound insulators are epigenetic structures that are important for both specific gene regulation as well as chromatin architecture**.

CTCF is evolutionarily very conserved both in its protein structure as well as in its DNA-binding pattern. Genome-wide there are approximately 30,000 CTCF binding sites, some 15% of which are involved in the formation of TADs and only a few hundred control imprinting (see below). All CTCF binding sites are sensitive to methylation; *i.e.*, CTCF binding to methylated sites is drastically reduced. **Higher-order chromatin structures, such as DNA loops, that are stabilized by CTCF binding, represent another form of epigenetic memory**, which can be modulated by DNA methylation. Interestingly, only a small subset of unmethylated CTCF binding sites keep CTCF proteins bound throughout the cell cycle in order to protect these sites against *de novo* methylation. Thus, only those higher-order chromatin structures that are mediated by unmethylated CTCF sites can be inherited through

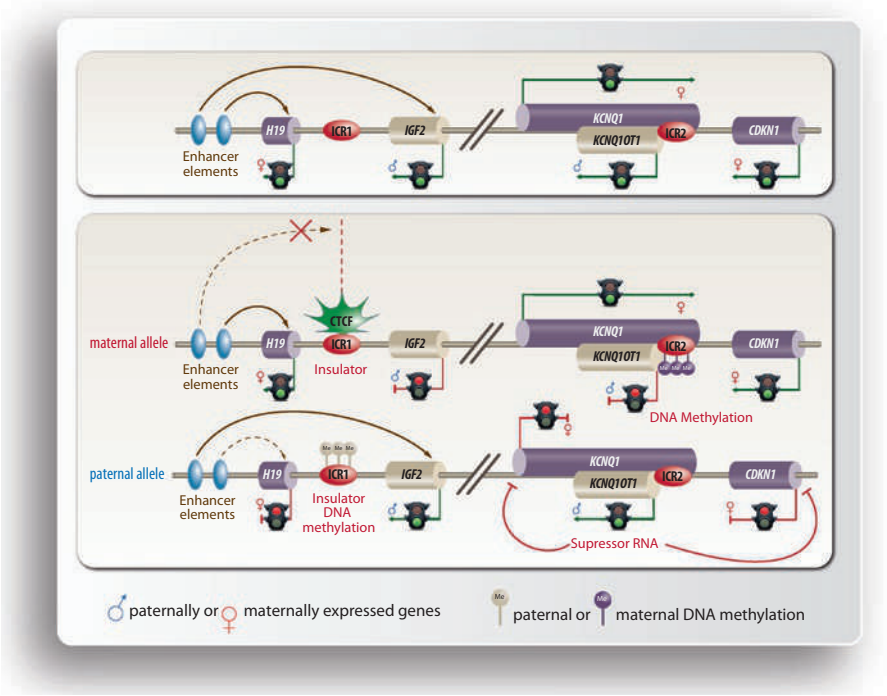


Fig. 3.4 **Control mechanisms of the 11p15 imprinted cluster**. General structure of the 11p15 cluster (**top**) and of scenarios of maternally (**center**) and paternally (**bottom**) controlled alleles. The Silver-Russell syndrome and the Beckwith-Wiedemann syndrome both are imprinting disorders that relate to this locus (Sect. 3.4). *IGF2* encodes for a growth factor, *H19* for a long ncRNA limiting body weight, *KCNQ1* for a potassium channel, *KCNQ1OT1* for an antisense transcript of *KCNQ1* that interacts with various chromatin components and finally *CDKN1C* for a cell cycle inhibitor

mitosis; *i.e.*, **CTCF-mediated chromatin structures represent a heritable component of phenotype-specific epigenetic programs**.

The DNA-methylation sensitive binding of CTCF to imprinting control regions (ICRs) provides a mechanistic explanation of the epigenetic process of genomic imprinting. ICRs represent a special subset of insulators that control the monoallelic expression of our more than 100 maternally and paternally controlled genes (www.geneimprint.com/site/genes-by-species.Homo+sapiens). Most imprinted genes occur in clusters, a master example of which is the chromosome 11p15 region that contains the protein-coding genes *IGF2* (insulin-like growth factor 2), *KCNQ1* (potassium voltage-gated channel subfamily Q member 1) and *CDKN1C* (cyclin-dependent kinase inhibitor 1C) as well as the ncRNA genes *H19* and *KCNQ1OT1* (Fig. 3.4, top). This imprinted genomic locus contains two ICRs and is regulated by enhancers downstream of the *H19* gene. In maternally controlled alleles, ICR1 is unmethylated and binds CTCF, while ICR2 is methylated and not bound (Fig. 3.4,

center). During post-implantation development CTCF binding is essential in order to maintain the hypo-methylated state of ICR1 and to protect it from *de novo* methylation in oocytes. CTCF blocks the long-range communication of the enhancers with the TSS region of the *IGF2* gene but allows the initiation of *H19* transcription. This results in the expression of *H19*, *KCNQ1,* and *CDKN1C* as well as in the repression of *IGF2* and *KCNQOT1* transcription. In contrast, in paternally controlled alleles, ICR1 is methylated and does not bind CTCF, while ICR2 is unmethylated (Fig. 3.4, bottom). This reverses the expression pattern, so that *IGF2* and *KCNQOT1* are produced but not *H19*, *KCNQ1* and *CDKN1C*. The physiological consequence of this imprinting is that **in maternally controlled cells growth and cell cycle are limited, while paternally controlled cells are primed for maximal growth**.

Another well studied example of imprinting is the inactivation of one X chromosome in female cells. The inactive X chromosome is observed as Barr body in female interphase cells. The epigenetic process behind X chromosome inactivation is the long ncRNA *Xist* (X inactive specific transcript) (Sect. 5.4), which is exclusively expressed from the X inactivation center of the inactive X chromosome. The action of *Xist* represents a special form of imprinting that affects a whole chromosome.

3.4 DNA Methylation and Disease

The approximately 100 imprinted genes in human have important roles during development, so that changes in their expression and function can lead to imprinting disorders. For example, the Silver-Russell syndrome, a disease leading to undergrowth and asymmetry, and the Beckwith-Wiedemann syndrome, a disease leading to overgrowth, are based on epigenetic errors in the 11p15 locus (Fig. 3.4). Individuals with the Beckwith-Wiedemann syndrome have a 1000-times increased chance of getting kidney tumors (mostly Wilms' tumors, however, in only 7% of those affected by the syndrome) and so-called embryonal tumors that arise from fetal cells and persist after birth; *i.e.*, **epigenetic changes precede and increase the risk of cancer rather than arise after tumor formation**. Most of the patients with Beckwith-Wiedemann syndrome lost the methylation at ICR2, resulting in the expression of the *KCNQ1OT1* ncRNA on both alleles (biallelic) and aberrant repression of *CDKN1C*, *i.e.*, in reduced cell cycle repression. Other Beckwith-Wiedemann syndrome patients show overexpression of *IGF2* caused by deletions in ICR1 on the maternal allele and disrupted CTCF binding leading to biallelic *IGF2* expression and loss of *H19* expression. Many individuals with Silver-Russell syndrome have an opposite epigenetic phenotype, where ICR1 is unmethylated, resulting in biallelic *H19* expression and loss of *IGF2* expression.

While imprinting disorders are very rare, perturbations in the DNA methylome are observed in the most cases of cancer (Sect. 8.1). For example, the CpG-rich promoters of tumor suppressor genes, such as *TP53*, are frequently hyper-methylated; *i.e.*, DNA methylation leads to the silencing of genes that are essential

for the inhibition of tumorigenesis progression. Moreover, in about 25% of adult cases of acute myeloid leukemia the *DNMT3A* gene is mutated, which disturbs the methylome and makes the regulatory landscape of pre-leukemic blood stem cells more vulnerable for additional mutations. Similarly, mutations in the genes of other chromatin modifiers can enhance the phenotype of cancer-driven mutations in transcription factor genes or their genomic binding sites.

Based on Waddington's model of an epigenetic landscape (Sect. 1.2), the epigenetic status of a cell, such as its methylation level, can be represented by a ball trapped in a valley. In case of normal differentiated cells, the borders of the valley are high and gene regulatory networks keep the cells in stable epigenetic homeostasis (Fig. 3.5, top left). This prevents the epigenetic state from moving too far from its equilibrium point in normal tissue. In contrast, a dys-regulation of the epigenome during tumorigenesis, such as overexpression of an epigenetic modulator or an inflammatory insult (Sect. 11.2), flattens the valley (Fig. 3.5, bottom left). Under these conditions of reduced regulation, the epigenetic status is more relaxed and

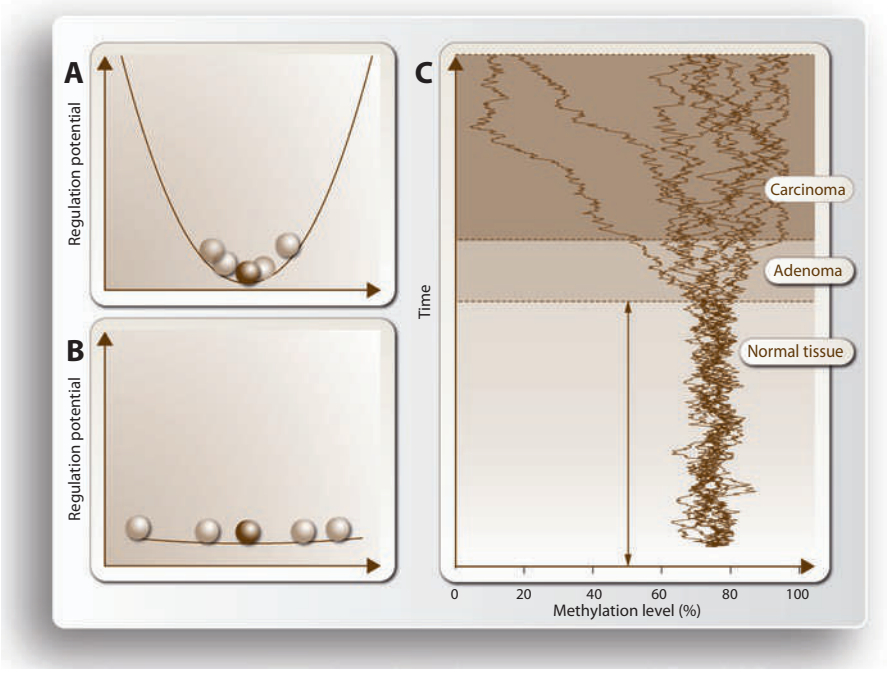


Fig. 3.5 **Exemplary model of epigenetic dys-regulation**. DNA methylation is used here as an example for epigenetic dys-regulation. The methylation level of a normal cell is illustrated as a ball at the bottom of a bin (**top left**), where regulatory forces, such as gene regulatory networks, allow only minor changes in the epigenetic status. In contrast, during tumorigenesis (**bottom left**) the landscape flattens, and the methylation levels can be far more variable. When the methylation level is modeled over time for 10 examples (**right**), the variations in the transition from normal tissue to adenoma and carcinoma become obvious as wider methylation ranges

influenced by stochastic variations. Thus, during tumorigenesis, DNA methylation levels diffuse away from the initial state in normal cells (Fig. 3.5, right). A plot of CpG methylation levels during the transformation of normal cells into adenoma and carcinoma cells shows a tight distribution in normal tissue, but a progression from adenoma to carcinoma. This explains the substantial level of epigenetic variation for a given cancer type across individuals or between metastatic cells originating from the same primary tumor. Accordingly, **there is no defined epigenetic signature for cancer**.

Aberrant DNA methylation is not only a well established marker of cancer and disturbed genomic imprinting, but it also can lead to general instabilities of the genome through reduced heterochromatin formation on repetitive sequences (Fig. 3.3a, right). One of the most frequent mutations found in human diseases is a C to T transition at methylated CpG islands; *i.e.*, the epigenetic mark cytosine methylation induces a genetic point mutation by reducing the efficiency of DNA repair at these sites. DNA methylation profiles, for example, of white blood cells that can be obtained from test persons with minimal invasion, may serve as biomarkers for evaluating the individual risk of cancer (Sect. 8.2) and a number of other diseases, such as type 2 diabetes (Sect. 10.3). Moreover, the DNA methylome indicates the progress of aging showing significant inter-individual differences (Sect. 7.3). Although biomarkers often do not explain the causality of a disease, they can monitor the disease state and may suggest appropriate therapy. Thus, **epigenomic profiles, such as DNA methylation patterns, in combination with genetic predisposition and environmental exposure, may be prognostic for personal risk of disease onset**.

Further Reading

Carlberg C, Molnár F (2016) Mechanisms of gene regulation. Springer Textbook ISBN: 978-94-007-7904-4.

Du J, Johnson LM, Jacobsen SE, Patel DJ (2015) DNA methylation pathways and their crosstalk with histone methylation. Nat Rev Mol Cell Biol 16:519–532

Krijger PH, de Laat W (2016) Regulation of disease-associated gene expression in the 3D genome. Nat Rev. Mol Cell Biol 17:771–782

Luo C, Hajkova P, Ecker JR (2018) Dynamic DNA methylation: in the right place at the right time. Science 361:1336–1340

Monk D, Mackay DJG, Eggermann T, Maher ER, Riccio A (2019) Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. Nat Rev. Genet 20:235–248

Schübeler D (2015) Function and information content of DNA methylation. Nature 517:321–326

Wu X, Zhang Y (2017) TET-mediated active DNA demethylation: mechanism, function and beyond. Nat Rev. Genet 18:517–534

Zhu H, Wang G, Qian J (2016) Transcription factors as readers and effectors of DNA methylation. Nat Rev. Genet 17:551–565

Histone Modifications

4

Abstract

Post-translational modifications are a general mechanism for instructing proteins about their function. Moreover, they allow proteins to "memorize" their encounters, such as contacts with other proteins. Histone proteins are the key examples of this protein communication system. In particular, lysine acetylations and methylations of histone tails have a large functional impact. Genome-wide profiling of the large set of post-transcriptional histone modifications provides the basis of the histone code. This code leads to the understanding, how the epigenome directs transcriptional regulation and stores information.

Keywords

Post-translational histone modification · Histone methylation · Histone acetylation · Genome-wide profiling · Histone code · Epigenetic memory

4.1 Histones and Their Modifications

Reversible post-translational modifications, such as phosphorylation, acetylation, and methylation, of key amino acid residues within proteins are the major mechanisms of communication and information storage in the control of signaling networks in cells. This means that **many proteins "remember" their functional tasks *via* their specific pattern of post-translational modifications**.

Master examples of such information-processing circuits via post-translational modifications are the nucleosome-forming histone proteins H2A, H2B, H3 and H4 as well as the linker histone H1 (Fig. 4.1). The tails and globular domains of these histone proteins provide over 130 amino acid residues for post-translational modifications, the information content of which is summarized as the histone code. The general feature of the core histone proteins is their small size of some 11–15 kD and their disproportional high content of the positively charged amino acids lysine

C. Carlberg, F. Molnár, *Human Epigenetics: How Science Works*,
https://doi.org/10.1007/978-3-030-22907-8_4

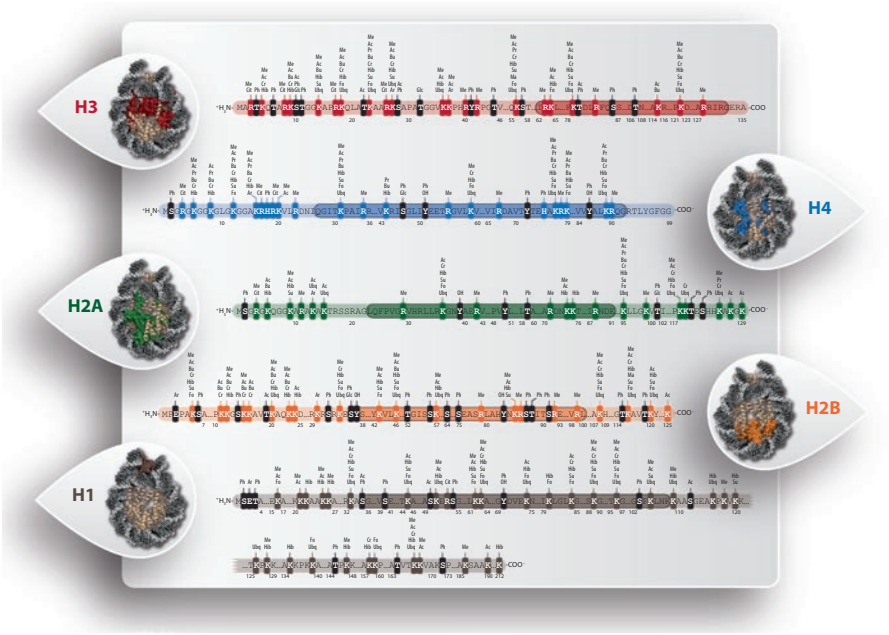


Fig. 4.1 **Post-translational modifications of histone proteins**. All presently known post-translational modifications of the nucleosome forming histones H2A, H2B, H3 and H4 and the linker histone H1 are indicated. Amino acids that can be modified (K, lysine; R, arginine; S, serine; T, threonine; Y, tyrosine; H, histidine; E, glutamate) are highlighted; most of them can carry different modifications, but not in parallel. **Me**, methylation (K, R); **Ac**, acetylation (K, S, T); **Pr**, propionylation (K); **Bu**, butyrylation (K); **Cr**, crotonylation (K); **Hib**, 2-hydroxyisobutyrylation (K); **Ma**, malonylation (K); **Su**, succinylation (K); **Fo**, formylation (K); **Ub**, ubiquitination (K); **Cit**, citrullination (R); **Ph**, phosphorylation (S, T, Y, H); **OH**, hydroxylation (Y); **Glc**, glycation (S, T); **Ar**, ADP-ribosylation (K, E)

and arginine, in particular at their amino-termini. Histones H2A, H2B and H3 exist in multiple variants (Box 4.1) and more than 100 human genes encode for histone proteins. This makes the histone gene family one of the largest within our genome. In contrast to transcription factors, **DNA binding of histones is not sequence-specific**.

Post-translational modifications of histones are frequent and important epigenetic signals that control many biological processes, such as cellular differentiation in the context of embryogenesis (Sect. 6.1). **Acetylations and methylations of lysines in histone tails are understood best and may be the most important epigenetic marks affecting histones**. However, there are also a number of other acylations, such as formylation, propionylation, malonylation, crotonylation, butyrylation, succinylation, glutarylation and myristoylation, the functional impact of which is far less understood (Fig. 4.1). In addition, there are phosphorylations at tyrosines, serines, histidines and threonines, ADP ribosylations at lysines and glutamates, citrullinations of arginines, hydroxylations of tyrosines, glycations of serines and threonines as well as sumoylations and ubiquitinations of lysines.

Box 4.1 Histone Variants

The canonical histones H2A, H2B, H3 and H4 represent the majority of histone proteins. In addition, there are eight variants of H2A (H2A.X, H2A.Z.1, H2A.Z.2.1, H2A.Z.2.2, H2A.B, macroH2A1.1, macroH2A1.2 and macroH2A2), two variants of H2B (H2BFWT and TSH2B) and six variants of H3 (H3.3, histone H3-like centromeric protein A (CENP-A), H3.1T, H3.5, H3.X and H3.Y), but humans have no variants of H4. Canonical histones are assembled into nucleosomes behind the replication fork to package newly synthesized genomic DNA. By contrast, the incorporation of histone variants into chromatin is independent of DNA synthesis and occurs throughout the cell cycle. Interestingly, canonical histones have no introns; *i.e.*, they have no splice variants while most of the genes for the histone variants do have introns and thus alternative splice variants. Histone variants are often subjected to the same modifications as canonical histones, but there are also variant-specific modifications on residues that differ from their canonical counterparts. Accordingly, histone variants also directly influence the structure of nucleosomes. For example, H2A.Bbd lacks acidic amino acids at its carboxy terminus, as a consequence of which only 118–130 bp (versus 147 bp) of genomic DNA are wrapped around the respective histone octamer. This leads to the formation of less compact and more accessible chromatin, which facilitates gene expression. The same nucleosome may contain multiple histone variants. There are homotypic nucleosomes, which carry two copies of the same histone, and heterotypic nucleosomes, which contain a canonical histone and a variant histone or two different histone variants. This allows for greater variability in nucleosome formation, stability and structure. For example, nucleosomes that contain H2A.Z and H3.3 are less stable than canonical nucleosomes and are often found at nucleosome-depleted regions of active promoters, enhancers and insulators. These labile H2A.Z/H3.3-containing nucleosomes serve as "place holders" and prevent the formation of stable nucleosomes around regulatory genomic regions. They can be easily displaced by transcription factors and other nuclear proteins that are not able to bind genomic DNA in the presence of a nucleosome composed of canonical histones. Thus, variable composition of nucleosomes can directly influence gene expression.

The enzymes responsible for histone modifications (writers) are often highly residue-specific (Sect. 5.1). Covalent modifications of histone proteins alter the physio-chemical properties of the nucleosome and are recognized by specific proteins (readers). Similarly, basically all covalent histone modifications are reversible *via* the action of specific enzymes (erasers). In general, chromatin acetylation is associated with transcriptional activation and controlled by two classes of antagonizing chromatin modifiers, HATs and HDACs. When a HAT adds an acetyl group to the amino group in the side chain of a lysine, the positive charge of this amino acid is neutralized (Fig. 4.2). In reverse, an HDAC can remove the acetyl group from

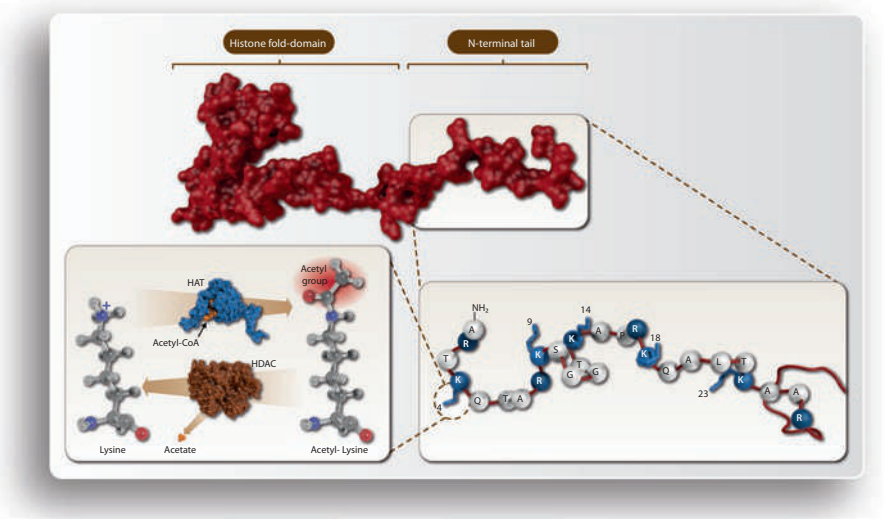


Fig. 4.2 **Histone acetylation**. Acetylation is shown as an example of a post-translational modification of histone proteins. A space-filling surface model with secondary structures of histone H3 (**top**) is displayed in combination with a zoom into its amino-terminal tail. The positively charged amino acids lysine (K) and arginine (R) are indicated in blue. The activity of HATs removes the positive charge, while HDACs can reverse this process (**bottom**)

H3K4ac and restore the positive charge of the lysine residue. Thus, **chromatin modifiers determine through the addition or removal of a rather small acetyl group the charge of the nucleosome core, which has major impact on the attraction between nucleosomes and the density of chromatin packaging**.

In analogy, for histone methylation there are two classes of enzymes with opposite functions, KMTs and KDMs (Sect. 5.1). Since histone methylation can be a repressive as well as an active marker, the exact position in the histone tail and its degree of methylation (mono-, di- or tri-methylation) is critical.

4.2 Epigenetic Marks of Transcriptional Regulation

Lysine is the most frequently modified amino acid in proteins, since it can accommodate a number of different modifications, such as several types of acylations and methylation, and reactions with ubiquitin and ubiquitin-like modifiers. These modifications occur in a mutually exclusive manner, so that specific lysine residues, such as H3K27, can serve as hubs for the integration of different signaling pathways (Fig. 4.3). Methylation is a special type of post-translational modification. Since the methyl group is small, it contributes only in a minor way to the steric properties of the amino acids. The methylation of lysines and arginines does not affect the charge of these residues; *i.e.*, also in their methylated form they are positively charged.

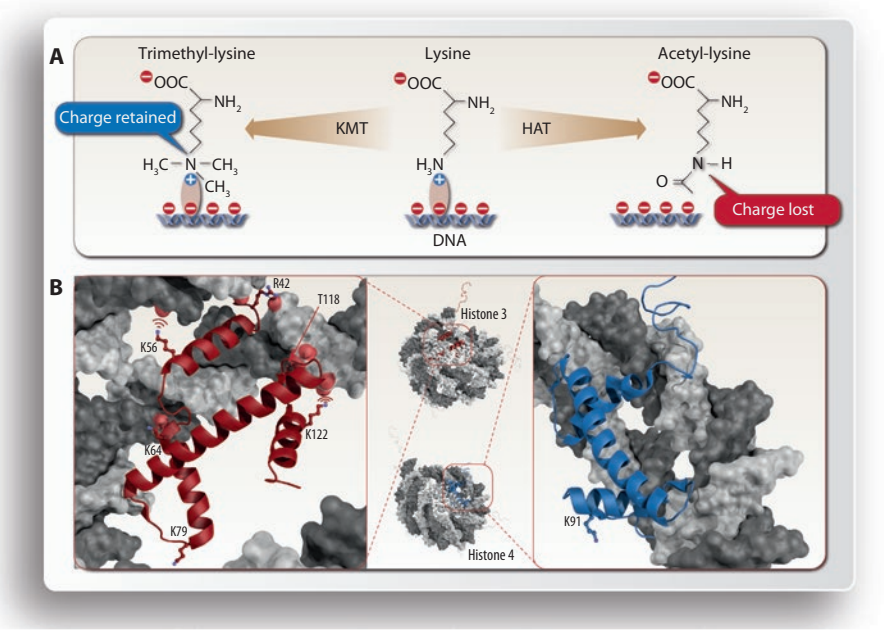


Fig. 4.3 **Nucleosome stability through histone modifications**. Unmodified lysine residues are positively charged and can form a salt bridge with negatively charged genomic DNA (both at physiological pH). The acetylation of lysines by HATs introduces a bulkier side chain and in parallel removes the positive charge (**top**). This decreases the affinity between DNA and the nucleosome and may destabilize the latter. The methylation of lysines by KMTs does not change the charge but - dependent on the number of added methyl groups - introduces various degrees of bulkiness. Crystal structures of the nucleosome are shown with highlighted key amino acids (**bottom**)

Lysines can be methylated up to three times and arginines up to two times, respectively. Histone methylations are more stable modifications than phosphorylations or acetylations; *i.e.,* their turnover is lower, and they mark more stable epigenetic states.

Post-translational modifications of histones either directly affect chromatin density and accessibility or serve as binding sites for effector proteins, such as chromatin modifying (Sect. 5.1) or remodeling complexes (Sect. 5.3). This ultimately modulates the initiation and elongation of transcription; *i.e.*, histone modifications have an impact on the transcriptome (Fig. 4.4). In addition, histone marks are able to store information (Sect. 4.3). Multiple histone modifications act in a combined way to generate a very specific chromatin structure that determines a specific expression level for each class of genes (Sect. 5.2). An individual histone modification is considered as a "letter" of the histone code, while multiple modifications combine to "words" of different specific meanings.

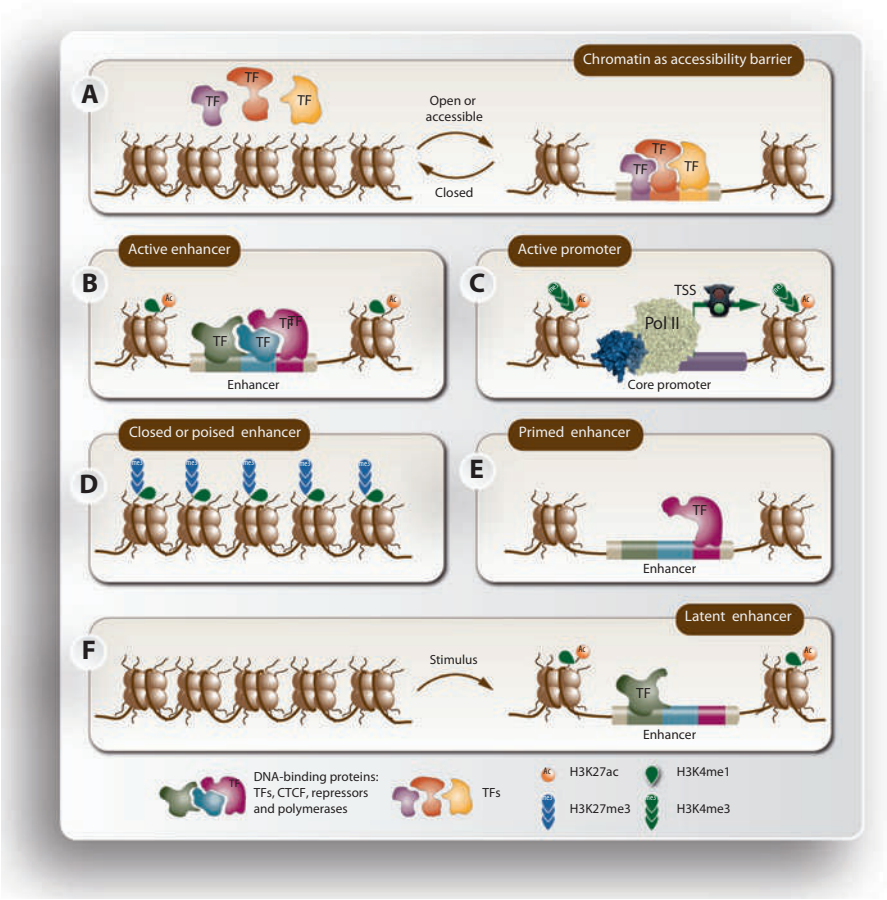


Fig. 4.4 **Chromatin accessibility and histone marks**. Chromatin restricts the access of transcription factors, Pol II and other nuclear proteins to the genome (**a**). Enhancers are often marked by H3K27ac and H3K4me1 modifications (**b**), while H3K27ac and H3K4me3 mark active promoters (**c**). Closed enhancers are termed "poised" when they carry both active histone marks (for example, H3K4me1) as well as repressive marks (for example, H3K27me3) (**d**). Enhancers are considered "primed" when they are pre-marked by H3K4me1 (**e**), while they are termed "latent" when they do not show any specific histone mark, but can become accessible through a stimulus (mostly an extra-cellular signal activating a signal transduction cascade) so that flanking nucleosomes acquire H3K4me1 and H3K27ac marks (**f**)

4.3 Genome-Wide Interpretation of the Histone Code

Maps of genome-wide histone modifications with patterns of chromatin accessibility, transcription factor binding as well as RNA expression from multiple tissues are provided by the *ENCODE Project* and the Roadmap Epigenomics Project and coordinated by *IHEC* (Sect. 1.3). The integrated data identified novel relationships

Table 4.1 Most conserved histone marks. Non-exclusive list of the relation of post-translational histone marks and their functional impact as confirmed by genome-wide analysis

Histone	Epi-mark	Most conserved co-marks	Status	Genomic region	Biological inference
H3	K4me1	K27me3, K4me2	poised	enhancer	
	K4me2	K4me3	active	regulatory regions	
	K4me3	K4me2	active or poised	promoter	poised enhancers regulate at many genes as bivalent promoters do
	K9me1		active	gene body and enhancer	
	K9me3		repressive	heterochromatin at promoters and enhancers	repetitive sequence (SINEs, LINEs, LTRs)
	K27ac	K4me1/2/3	active	enhancer (K4me1/me2)or promoter (K4me2/me3)	H3K27ac marks enhancers and promoters
	K27me1		active	enhancer	
	K27me3	K4me1/2/3	repressive	polycomb repressed region, promoters	X chromosome inactivation center
	K36me1		active	enhancer	
	K36me3	K27ac, K4me1	active	gene body	not correlated with H3K27me3, may be a neglected mark of active enhancers
H4	K20me1		active	gene body	
	K20me3		active or poised	regulatory regions	

between histone modifications and related chromatin structures. This provided a genome-wide basis of the histone code and its impact. With 15 chemical modifications that can occur at more than 130 sites on the five canonical histones (Fig. 4.1) and some 30 histone variants (Box 4.1), the theoretical number of possible combinations of signals forming the histone code is very large. In analogy to the alphabets of human languages, the histone code is very rich in "letters" that may be combined to a large number of "words" with different meanings (Table 4.1). Thus, **histone modifications represent a text rich in information about the local chromatin status**. This more fine-grained distinction suggests that the epigenome has far more differential functions than "on" and "off".

For some common histone marks, correlations between their presence and the activity of different genomic elements are well established, such as H3K9me3 and H3K27me3 at inactive or poised promoters (Fig. 4.4D), H3K27ac and H3K4me3 on active enhancers and promoters as well as H3K36me3 in transcribed gene bodies, respectively (Table 4.1). Poised genes, promoters and enhancers are found within facultative heterochromatin (Sect. 1.1); *i.e.*, the respective genomic regions are in a waiting position for activating signals.

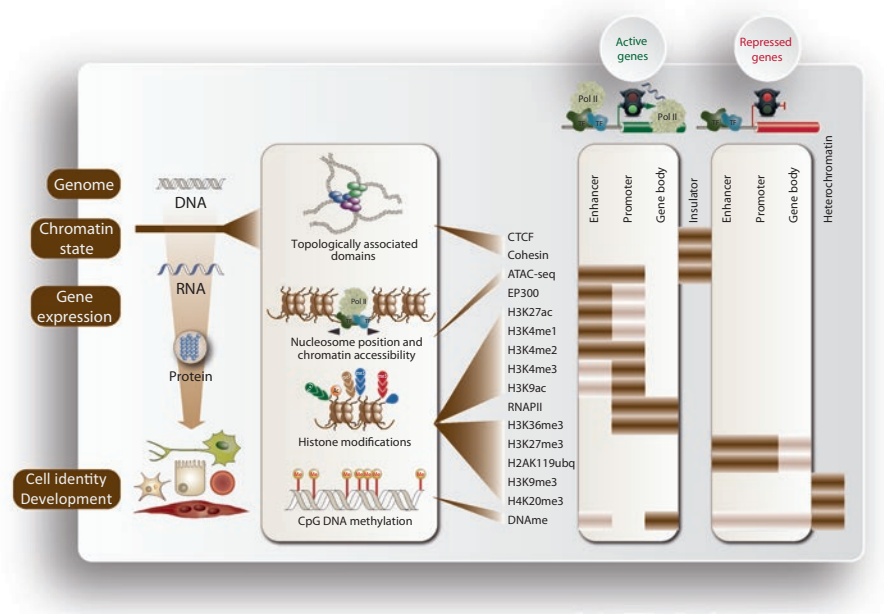


Fig. 4.5 **Impact of the epigenome for gene expression**. Chromatin acts as a filter for the genome concerning gene expression and in this way determines cell identity (**bottom left**). Epigenomic regulation happens at various scales of chromatin states (**center**), such as topological organization, chromatin accessibility, histone modifications and DNA methylation. Key histone modifications and binding nuclear proteins that are characteristic for these chromatin states are indicated and distinguished between active and repressed genes (**right**). CTCF and cohesin are involved in chromatin organization, the HAT EP300 (E1A binding protein p300, also called KAT3B) marks enhancers, and both Pol II and H3K36me3 indicate actively transcribed genes. For protein binding or histone modification lighter shades mark a lower or variable degree of modifications, while for DNA methylation it indicates that the genomic region can be regulated by methylation

Chromatin states, as marked by histone modifications, characterize genomic elements, such as enhancers, promoters, insulators and gene bodies (Fig. 4.5). Thus, **chromatin represents a cell type-specific filter of genomic sequence information that determines, based on the histone code, which genes are transcribed into RNA**. The following characteristics of the histone code are already well understood:

- Euchromatin is characterized by general acetylation of lysines in the tails of histones H3 and H4 as well as H3K27ac and H3K4me3 marks.
- In heterochromatin H3K9, H3K27 and H4K20 are either mono-, di- or tri-methylated.
- Acetylation and deacetylation of histone tails represent major regulatory mechanisms during gene activation and repression.
- Actively transcribed regions of the genome tend to be hyper-acetylated, whereas inactive regions are hypo-acetylated.

- The overall degree of acetylation rather than any specific residue is critical.
- In contrast to acetylation, there is a clear functional distinction between histone methylation marks, both concerning the exact histone residues as well as their degree of modification, such as mono-, di- or tri-methylation.
- H3K9me3 and H4K20me3 are enriched near boundaries of large heterochromatic domains, while H3K9me1 and H4K20me1 are found primarily in active genes.
- H3K4me3 is detected specifically at active promoters, while H3K27me3 is correlated with gene repression over larger genomic regions. Both modifications are usually located in different chromatin domains, but when they coexist on enhancers and/or promoters, the respective genomic regions are termed bivalent.
- Latent enhancers are initially not labeled by H3K4me3 or H3K27ac, but acquire these active marks and transcription factor binding upon stimulation of cellular signaling pathways (Fig. 4.4f).
- H3K36me3 levels correlate with levels of gene transcription, since KMTs deposit this mark when interacting with elongating Pol II; *i.e.*, expressed exons have a strong enrichment for this histone mark.

Histone modification profiles allow the identification of distal enhancer regions, as they show relative H3K4me1 enrichment and H3K4me3 depletion. Interestingly, chromatin patterns at enhancer regions seem to be far more variable, as they show enrichment not only for H3K27ac, but also for H2BK5me1, H3K4me2, H3K9me1, H3K27me1 and H3K36me1, suggesting the redundancy of these histone marks.

Further Reading

Biggar KK, Li SS (2015) Non-histone protein methylation as a regulator of cellular signalling and function. Nat Rev Mol Cell Biol 16:5–17

Carlberg C, Molnár F (2016) Mechanisms of Gene Regulation. Springer Textbook ISBN: 978-94-007-7904-4

Choudhary C, Weinert BT, Nishida Y, Verdin E, Mann M (2014) The growing landscape of lysine acetylation links metabolism and cell signalling. Nat Rev Mol Cell Biol 15:536–550

Sabari BR, Zhang D, Allis CD, Zhao Y (2017) Metabolic regulation of gene expression through histone acylations. Nat Rev Mol Cell Biol 18:90–101

Shlyueva D, Stampfel G, Stark A (2014) Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet 15:272–286

Tessarz P, Kouzarides T (2014) Histone core modifications regulating nucleosome structure and dynamics. Nat Rev Mol Cell Biol 15:703–708

Chromatin Modifying Proteins and RNAs

5

Abstract

Chromatin modifying enzymes either add or remove post-translational modifications to histone proteins and in this way change the functional profile of the epigenome. Antagonizing enzymes, such as HATs and HDACs, often co-localize at chromatin regions and fine-tune gene expression. Chromatin modeling complexes change the structure of chromatin by shifting, evicting or replacing nucleosomes. Finally, also long ncRNAs, such as *Xist*, can affect chromatin structure and function.

Keywords

Chromatin modifiers · Bromodomain · Chromodomain · PHD finger · HATs · HDACs · KMTs · HDMs · Writers · Erasers · Readers · ATP-dependent remodeling complex · Nucleosome dynamics · Long ncRNAs · *Xist*

5.1 Chromatin Modifiers

An average human cell has only some 100,000 open loci within its chromatin; *i.e.*, more than 90% of the genome is buried in heterochromatin and not accessible to transcription factors and Pol II. However, **many of these accessible chromatin regions are not static as they are dynamically controlled by chromatin modifying and remodeling proteins** (Fig. 5.1). These enzymes catalyze the methylation of genomic DNA (Sect. 3.1), the post-translational modification of histone proteins (Sect. 4.1) or the positioning of nucleosomes (Sect. 5.3). Our genome expresses in a tissue-specific fashion hundreds of these chromatin modifiers and remodelers that recognize (read), add (write), and remove (erase) chromatin marks. Writer-type enzymes, such as HATs, KMTs, and DNMTs, add acetyl- or methyl groups to histone proteins or cytosines of genomic DNA, respectively. In contrast, eraser-type enzymes, such as HDACs, KDMs, and TETs, reverse these

C. Carlberg, F. Molnár, *Human Epigenetics: How Science Works*,
https://doi.org/10.1007/978-3-030-22907-8_5

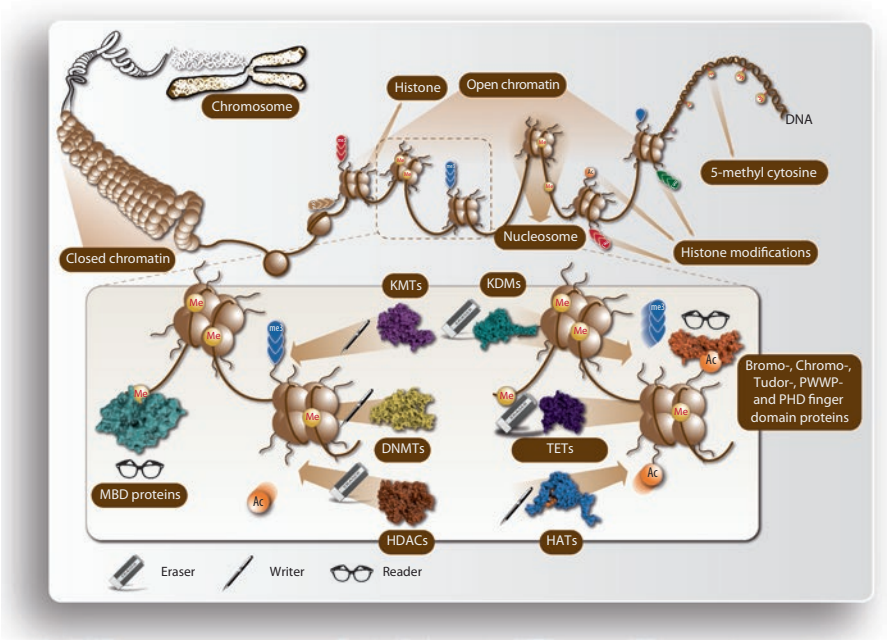


Fig. 5.1 **Central role of chromatin**. Covalent modifications of histones and genomic DNA, such as methylations, control the accessibility of chromatin to transcription factors and other regulatory proteins (**top**). These chromatin marks are introduced by writers, interpreted by readers and can be removed by erasers (**bottom**). The interplay between these nuclear proteins is essential for controlling gene expression

reactions and eliminate the respective marks. DNA methylation-specific reader-type proteins, such as MBD proteins and CTCF, were already discussed in Sect. 3.2. These proteins bind DNA depending on its methylation status. Moreover, also components of the chromatin remodeling complexes are able to read chromatin marks (Sect. 5.3).

Our genome encodes for 22 members of the HAT family and 18 members of the HDAC family taking care on the acetylation status of chromatin. The Zn^{2+}-dependent HDACs 1-11 act predominately in the nucleus and the cytoplasm, while nicotinamide adenine dinucleotide $(NAD)^{+}$-dependent sirtuins (SIRTs) 1-7 are found in addition also in mitochondria (Sect. 10.2). Furthermore, we have 66 KMTs and 20 KDMs that use both histone and non-histone proteins as substrates; *i.e.,* these enzymes control the methylation status of chromatin and other proteins. KDMs are either flavin adenine dinucleotide (FAD)-dependent monoamine oxidases or Fe(II) and α-ketoglutarate-dependent dioxygenases; *i.e.,* like SIRTs they sense via these metabolites the energy status of cells (Sect. 10.2).

Chromatin acetylation is generally associated with transcriptional activation, whereas the exact residue of the histone tails that is acetylated is not very critical (Sect. 4.3). In contrast, histone methylation, as mediated by KMTs, mainly mediates chromatin repression, but at certain residues, such as H3K4, it results in activation. Therefore, for histone methylation, in contrast to acetylation, the exact position of

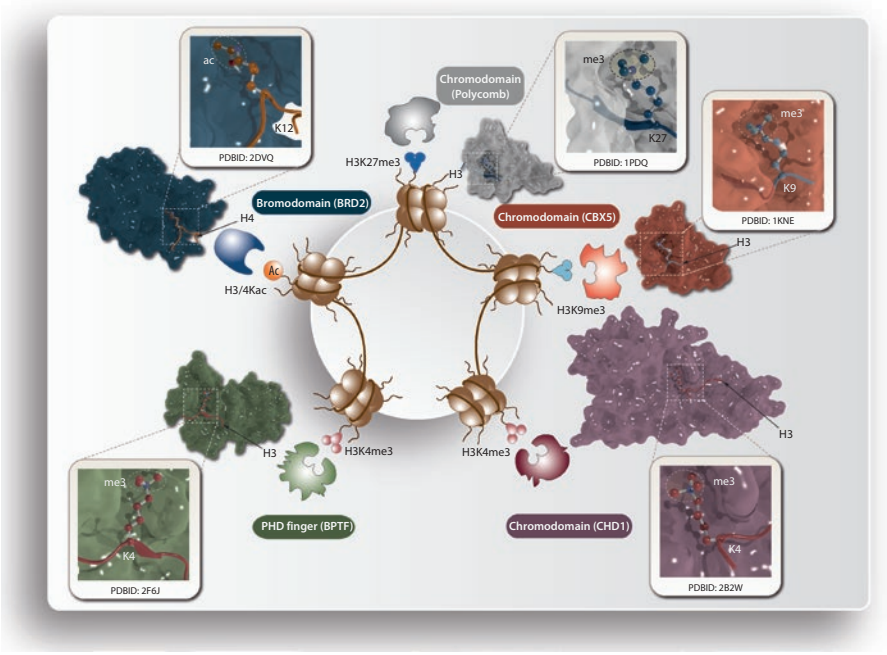


Fig. 5.2 **Histone reading proteins**. There are three major types of domains, such as chromodomains, PHD fingers, and bromodomains, by which histone modifications are recognized (read). While chromodomains and PHD fingers bind to specific histone methylations, bromodomains are rather unspecific and recognize all forms of acetylated histones. BPTF, bromodomain PHD finger transcription factor; BRD2, bromodomain containing 2; CBX5, chromobox 5, also called HP1; CHD1, chromodomain helicase DNA binding protein 1

the residue within the histone tail as well as its degree of methylation (mono-, di- or tri-methylation) is of critical importance.

The effects of chromatin modifiers, such as HATs and HDACs, are primarily local and may cover only a few nucleosomes up and downstream of the starting point of their action. The same applies to KMTs and chromatin remodeling enzymes, such as the SWI/SNF complex (Sect. 5.3). In case there is more HAT activity, chromatin is locally acetylated, the attraction between nucleosomes and genomic DNA decreases and the latter gets accessible for activating transcription factors, basal transcription factors and Pol II. In this euchromatin state chromatin remodeling enzymes, such as SWI/SNF, may have to fine-tune the position of the nucleosomes in order to obtain full accessibility of the respective binding sites. In the opposite case, when HDACs are more active, acetyl groups get removed and the packaging of chromatin locally increases. KMTs then methylate the same or adjacent amino acid residues in the histone tails that attract heterochromatin proteins, such as HP1, and further stabilize the local heterochromatin state (Fig. 1.2).

Specific histone marks are specifically recognized (read) by a large number of chromatin modifiers *via* a small set of common recognition domains (Fig. 5.2).

Repressive proteins of the Polycomb family use chromatin-organization modifier domains (chromodomains) in order to interact with methylated chromatin. Some 10 different HATs have a plant homeodomain (PHD) finger, which is a specific reading motif for H3K4me2 and H3K4me3 marks. There are even 46 human proteins (HATs, HAT-associated proteins, KMTs, helicases, ATP-dependent chromatin remodeling proteins, transcriptional co-activators and nuclear scaffolding proteins) that carry a bromodomain in order to recognize acetylated lysines. Chromodomains are far more specific for a given chromatin modification than bromodomain proteins; *i.e.*, chromodomain-containing nuclear proteins recognize their genomic targets with far higher accuracy. In addition, a smaller set of proteins use a YEATS domain for sensing acetylated and crotonylated lysines. Finally, proteins with a Tudor domain recognize methylated lysines and arginines.

5.2 Gene Regulation *via* Chromatin Modifiers

Cells are constantly exposed to a multitude of signals, such as the extra-cellular matrix, cytokines, peptide hormones and other active compounds, the majority of which are transmitted by receptors at the membrane. These extra-cellular signals induce *via* membrane proteins intra-cellular signal transduction cascades. These pathways often terminate at nuclear proteins, such as transcription factors, chromatin modifying and remodeling proteins; *i.e.*, they modulate the epigenome and transcriptome. Most of the signals vary over time and usually have an "on" or "off" character, but the resulting changes in the transcriptome rather resemble a waveform (Fig. 5.3, left). For example, a signal can either directly activate chromatin modifiers, which then write or erase histone marks, or act indirectly *via* the activation of chromatin remodelers that alter the nucleosome composition (Sect. 5.3). In this way, **chromatin-associated proteins act as signal converters and integrators**.

In general, the methylation of histones has a longer half-life than its acetylation or phosphorylation, respectively. Thus, signals can be stored within the epigenomic landscape for shorter and longer time periods. In particular, **the histone methylome is suited for a long-term epigenetic memory**. Histone modifications act in combination with DNA methylation and transcription factor activity (Fig. 5.3, right), which increases the diversity of their outputs. Some of the information stored in the histone modification pattern can even be maintained throughout DNA replication; *i.e.*, **a part of the histone marks can be inherited** (Sect. 7.1).

Genome-wide histone modification maps correspond to different genomic features, such as promoters, enhancers, and gene bodies, or activation states, such as actively transcribed, poised, or silenced, and often exist in combinations. The three main modes for the activity of genes are distinguished:

- **Active genes** are expressed genes that are associated with histone acetylation, H3K4me1, 2 & 3 marks and H2A.Z occurrence in their TSS regions as well as a number of different marks (H2BK5me1, H3K9me1, H3K27me1, H3K36me3,

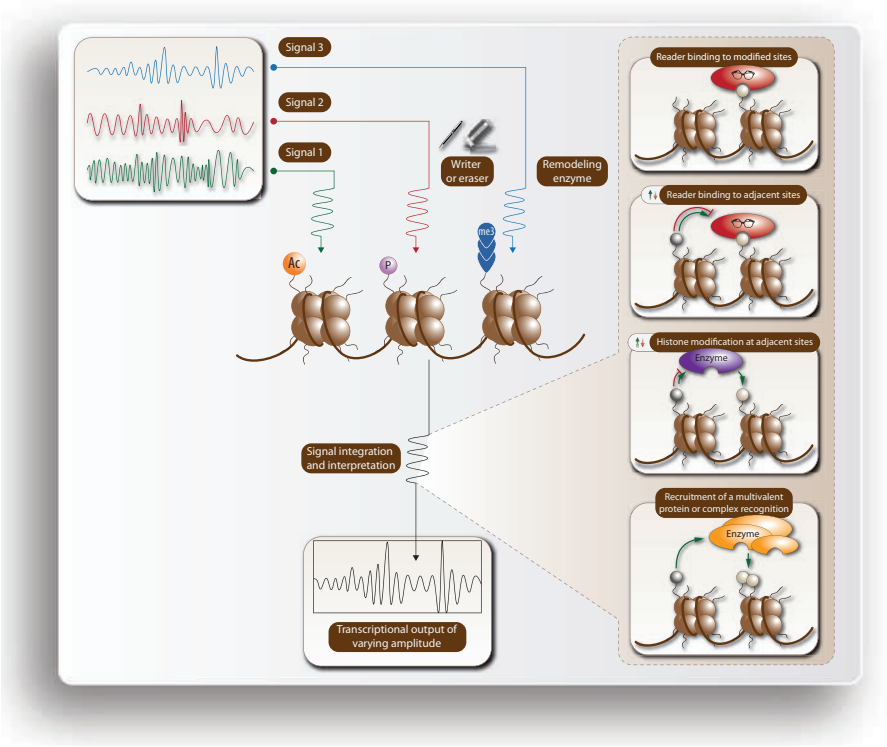


Fig. 5.3 **Signal storage and interpretation *via* chromatin modifiers and readers**. Signals deriving from various, mostly membrane-based signal transduction cascades are integrated on chromatin through modifications at histone tails. Multiple inputs occurring over time can be stored. These inputs can affect chromatin directly or are transmitted *via* chromatin modifiers, such as the writers HATs and KMTs, as well as the erasers HDACs and KDMs, and chromatin-remodeling proteins (**left**). There are a number of mechanisms how this dynamic epigenetic landscape is constantly interpreted by reader proteins, such as changing the ability of a reader protein to recognize an adjacent mark, recruiting enzymes that modify additional sites, and creating a combinatorial display for recognition in various binding events (**right**). The net result of the signal integration can be observed as transcriptional output, *i.e.*, as a change of the transcriptome (**bottom left**)

H3K79me1, 2 & 3 and H4K20me1) in their gene bodies (Fig. 5.4, top). Highest levels of both HATs and HDACs are found associated within these genes. Their presence correlates positively with mRNA expression and Pol II levels. The antagonizing activities of the chromatin modifiers are the basis for precise fine-tuning of gene expression *via* the homeostasis of active chromatin loci and apply both for TSS and enhancer regions (Fig. 4.4). Thus, active genes are found in euchromatin.

- **Poised genes** are not expressed and do not associate with significant histone acetylation, but they show H3K4 methylation marks and H2A.Z occurrence (Fig. 5.4, center). As long these genes wait for their activation, only low level of HATs or HDACs are associated with them; they are found in facultative

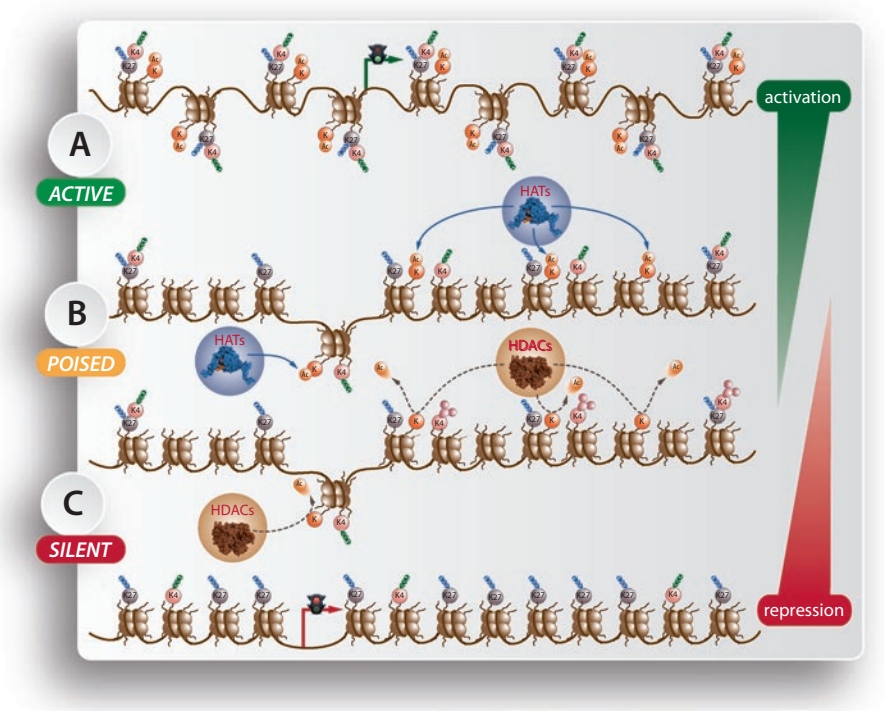


Fig. 5.4 **The epigenome of active (a), poised (b) and silent (c) genes**. Both HATs and HDACs are found at active and poised genes. HDACs remove acetyl groups that had been added by HATs after being recruited by elongating Pol II. Lower levels of HATs and HDACs are found at poised genes being primed by H3K4 methylation. HDACs prevent Pol II binding and thereby repress transcription *via* the removal of acetyl groups that had been added by transiently binding HATs. At silent genes that are devoid of H3K4 methylation no HAT or HDAC binding is detectable

heterochromatin. A cycle of transient acetylation and deacetylation keeps these genes primed but inactive and maintains their promoter regions in a poised state waiting for future activation *via* external signals.

- **Silent genes** either carry H3K27me3 marks together with Polycomb proteins or do not have at all any known chromatin marker (Fig. 5.4, bottom). At these genes neither HATs nor HDACs are found, and they are located within heterochromatin.

Taken together, **chromatin modifiers maintain the epigenome and in this way control gene expression**. Thus, these nuclear enzymes have central importance during embryogenesis as well as in cell fate decisions in health and disease.

5.3 Chromatin Remodelers

A cell's phenotype depends on its gene expression pattern, which basically is influenced how the genomic DNA is packed into chromatin. Nucleosomes often block the access of transcription factors to their genomic binding loci, since the packing of genomic DNA around histone octamers hides one side of the DNA. Within a stretch of 200 bp genomic DNA 147 bp contacting a histone octamer. Binding sites that are located close to the center of these 147 bp are generally inaccessible to transcription factors. Sites closer to the edge of the nucleosome-covered sequence are a bit better reachable, but only within the 50 bp between two neighboring nucleosomes genomic DNA is fully accessible. Thus, in some cases only minor shifts in the position of the nucleosomes are necessary in order to get transcription factor binding sites accessible, whereas in other cases a whole nucleosome needs to be depleted.

Since nucleosomes have a rather strong electrostatic attraction for genomic DNA, the catalysis of the sliding, removal or exchange of individual subunits, or even the eviction of whole nucleosomes has to dissolve all histone-DNA contacts and requires the investment of energy in the form of ATP. Thus, chromatin remodelers are multi-protein complexes that use the energy of ATP hydrolysis in order to affect nucleosomes in at least four ways (Fig. 5.5). These are

- the movement (**sliding**) of the histone octamer to a new position within the same chromatin region
- the complete displacement (**ejection**) of the histone octamer, for example, from TSS regions of heavily expressed genes
- the **removal** of H2A-H2B dimers from the histone octamer
- the **exchange** of regular histones by their variant forms, such as H2A by H2A.Z (Box 4.1).

Chromatin remodelers make TSS and enhancer regions either more or less accessible to the transcriptional apparatus, thereby allowing transcription factors to activate or repress, respectively, the transcription of their target genes. Thus, nucleosome occupancy and composition are tailored genome-wide.

Our genome encodes for four families of chromatin remodeling complexes that are distinguished based on the differences in their catalytic ATPases and associated subunits. The existence of different complexes implies that the respective chromatin remodelers have different mechanisms of action, but all of them contain an ATPase-DNA translocase, *i.e.,* an ATP consuming enzyme that dissolves DNA-protein contacts and allows moving of the nucleosome. Each remodeler family has multiple subtypes that provide cell type- or developmentally specific functions:

- The imitation switch (ISWI) complex assembles nucleosomes and spaces them in order to limit chromatin accessibility and gene expression (Fig. 5.6, left).

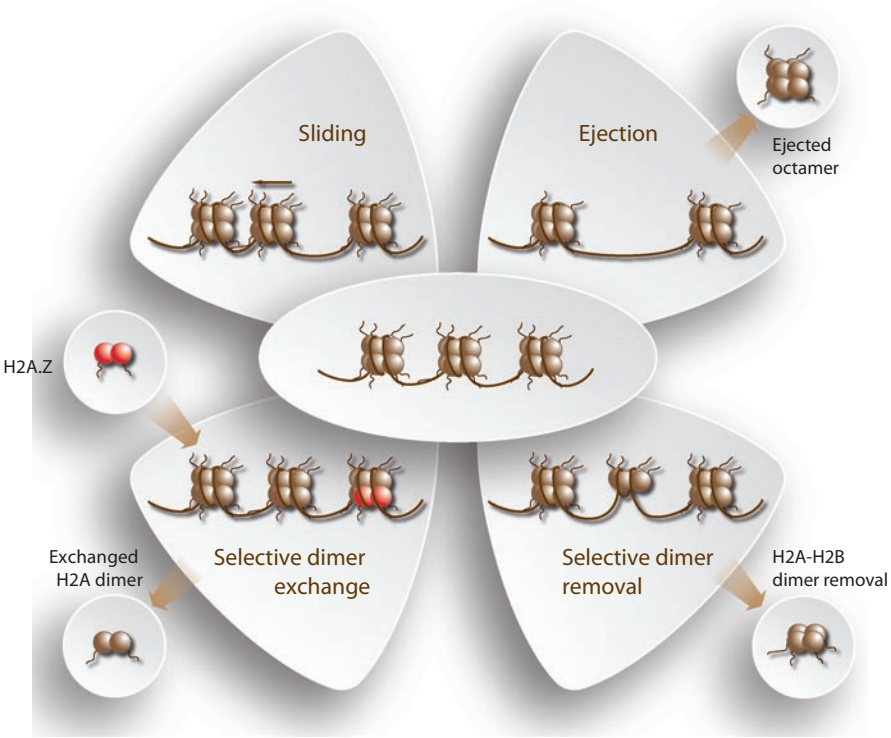


Fig. 5.5 **Mobility and stability of nucleosomes**. Chromatin remodelers enable access to genomic DNA through sliding, ejection, H2A-H2B dimer removal, or selective dimer exchange from nucleosomes. ATP-dependent remodeling complexes as well as thermal motion influence the mobility of nucleosomes. The stability of nucleosomes is affected by its detailed octamer composition and the pattern of histone modifications. For example, the incorporation of histone variants into nucleosomes alters the interactions with histone and non-histone proteins

- Chromodomain-helicase-DNA binding (CHD) remodelers conduct assembly (spacing nucleosomes), access (exposing promoters), and editing (incorporating histone H3.3) (Fig. 5.6, left).
- Switch/sucrose non-fermentable (SWI/SNF) complexes slide and eject nucleosomes; *i.e.,* they modulate chromatin access in order to activate of repress gene expression (Fig. 5.6, center).
- Inositol requiring (INO) 80 remodelers primarily have nucleosome editing functions (Fig. 5.6, right).

Remodeling complexes contain proteins with bromodomains, chromodomains and PHD domains that can read histone marks (Sect. 5.1). For example, ISWI remodelers use a PHD domain for targeting H3K4me3 marked histones, in CHD remodelers chromodomains bind to methylated histones, in SWI/SNF remodelers a

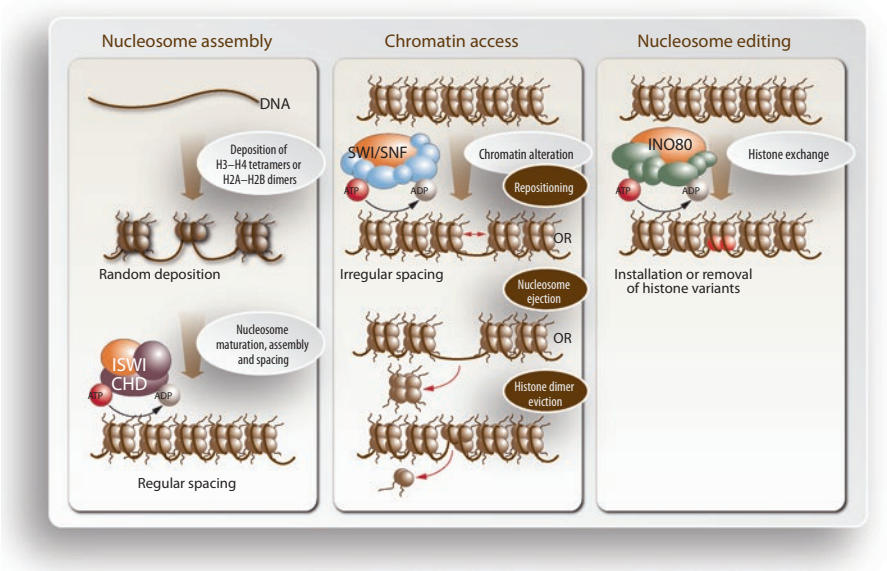


Fig. 5.6 **Function of human remodeling enzyme complexes**. ISWI and CHD remodelers are involved in the random deposition of histones, the maturation of nucleosomes and their spacing (**left**). SWI/SNF remodelers alter chromatin by repositioning nucleosomes, ejecting octamers or evicting histone dimers (**center**). INO80 remodelers change nucleosome composition by exchanging canonical and variant histones, such as installing H2A.Z variants (**right**)

bromodomain is used for recognizing acetylated histone H3, while INO80 remodelers employ a bromodomain for binding to acetylated H4 (Fig. 5.2).

At steady state, chromatin remodelers ensure dense nucleosome packaging in the vast majority of the genome, while at particular genomic loci they allow the rapid access of transcription factors and other nuclear proteins. For example, constitutively active genes typically have a nucleosome-depleted region upstream of their TSS, within which key transcription factor binding sites reside. Genome-wide studies indicated that often a 200 bp nucleosome-depleted region upstream of the TSS is flanked on either side by well positioned nucleosomes.

The action of SWI/SNF family members is mostly associated with transcriptional activation. Interestingly, the activity of many chromatin remodelers is affected by the presence of histone variants that they themselves introduce into the chromatin; *i.e.*, they control each other's action through the exchange of histones. The histone variants MacroH2A and H2A.Bbd reduce the efficiency of the SWI/SNF complex, whereas H2A.Z stimulates remodeling by ISWI complexes. The INO80 complex removes H2A.Z from inappropriate locations. In general, H2A.Z resides at open TSS regions and positively regulates gene transcription. The unique amino-terminal tail of this histone variant becomes acetylated when a gene is active.

5.4 Long ncRNAs as Chromatin Organizers

During the past 15 years tens of thousands of RNA transcripts were discovered in human tissues and cell types that resemble mRNAs but do not translate into proteins; *i.e.*, they are ncRNAs. When ncRNAs are longer than 200 nt, they are called long ncRNA. Long ncRNAs are heterogeneous in their biogenesis, abundance, and stability, and they differ in the mechanism of action. Some long ncRNA have a clear function, such as in regulation of gene expression, while others, such as enhancer RNAs (eRNAs), may be primarily side products non-precise of Pol II transcription.

Despite their rather recent discovery, ncRNAs are probably evolutionary older than proteins; *i.e.*, in early cells they mediated most of the regulatory actions, many of which were taken over later by proteins. Long ncRNAs carry out their cellular functions by interacting with proteins to form macromolecular complexes (Fig. 5.7). The complex formation is enabled *via* elements within ncRNAs, such as short

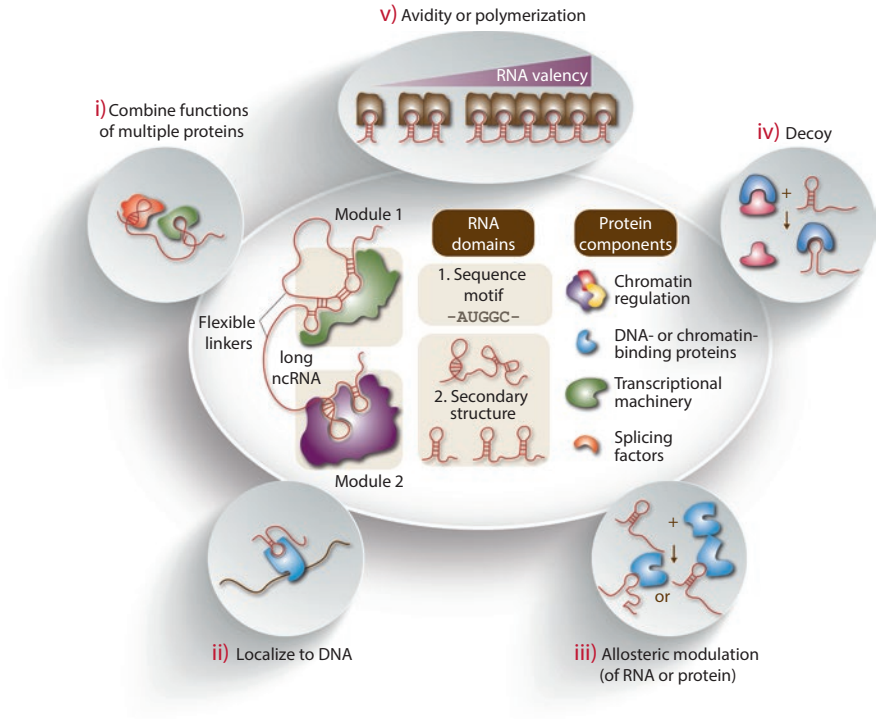


Fig. 5.7 **Principles of long ncRNA action**. Long ncRNA molecules have various regions for the molecular interaction with distinct protein complexes. These interactions have functions, such as combining the functions of multiple proteins, localizing long ncRNAs to genomic DNA, modifying the structure of long ncRNAs or proteins, inhibiting protein function as decoys, and providing a multi-functional platform, in order to increase the avidity of protein interactions or to promote RNA-protein complex polymerization (RNA valency)

sequence motifs or larger secondary or tertiary structures, that interact specifically with a large set of molecular structures in proteins, RNA, and DNA. This allows a large variety of functions, such as the ability to

- scaffold and recruit multiple regulatory proteins
- localize to specific targets on genomic DNA
- utilize and shape the 3D structure of the nucleus.

A classic example of an RNA scaffold is the telomerase RNA TERC that assembles the telomerase complex. A number of chromatin modifying and remodeling proteins, such as PRC components, KMT1C, KDM1A, DNMT1, and the SWI/SNF complex, interact with nuclear long ncRNAs. These RNA-protein interactions

- recruit chromatin regulatory complexes to specific genomic sites in order to regulate gene expression
- competitively or allosterically modulate the function of nuclear proteins
- combine and coordinate the functions of independent protein complexes (Fig. 5.7).

The already discussed long ncRNAs *Xist* (Sect. 3.3) is the key initiator of X chromosome inactivation in female cells and serves as the master example of how ncRNAs contribute to chromatin organization (Sect. 6.3). Before the expression of *Xist* both X chromosomes are transcriptionally active, not strongly associated with the nuclear lamina and structurally organized similar to autosomal chromosomes; *i.e.*, they are subdivided into more than hundred TADs. In early embryonic development, *Xist* is expressed by one allele, spreads across the X chromosome and interacts with lamin B receptor, which relocates the chromosome to the nuclear lamina. In this context, active genes are sequestered into the *Xist* compartment and silenced. Other long ncRNAs, such as *HOX* transcript antisense RNA (*HOTAIR*), direct some KDMs, such as KDM1A within the REST co-repressor (RCOR) complex, to their chromatin target sites. RCOR is a large protein complex that also contains HDACs and contributes to transcriptional repression.

Further Reading

Atlasi Y, Stunnenberg HG (2017) The interplay of epigenetic marks during stem cell differentiation and development. Nat Rev Genet 18:643–658

Carlberg C, Molnár F (2016) Mechanisms of Gene Regulation. Springer Textbook ISBN: 978-94-007-7904-4

Clapier CR, Iwasa J, Cairns BR, Peterson CL (2017) Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. Nat Rev Mol Cell Biol 18:407–422

Engreitz JM, Ollikainen N, Guttman M (2016) Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. Nat Rev Mol Cell Biol 17:756–770

Marazzi I, Greenbaum BD, Low DHP, Guccione E (2018) Chromatin dependencies in cancer and inflammation. Nat Rev Mol Cell Biol 19:245–261

Mozzetta C, Boyarchuk E, Pontis J, Ait-Si-Ali S (2015) Sound of silence: the properties and functions of repressive Lys methyltransferases. Nat Rev Mol Cell Biol 16:499–513

Rowley MJ, Corces VG (2018) Organizational principles of 3D genome architecture. Nat Rev Genet 19:789–800

Sheikh BN, Akhtar A (2019) The many lives of KATs – detectors, integrators and modulators of the cellular environment. Nat Rev Genet 20:7–23

Embryogenesis and Cellular Differentiation

6

Abstract

Early embryonic development is very vulnerable to environmental influences. Therefore, from all phases of our life, embryogenesis is the period, where epigenetics has the largest impact. Epigenetics directs the programing of PGCs, induced pluripotency as well as the function of adult stem cells in tissue homeostasis. These master examples demonstrate the impact of epigenetics on the organization of our body in health and disease.

Keywords

Embryogenesis · ES cells · Cell lineage commitment · PGCs · Cellular reprograming · Induced pluripotency · Master transcription factors · Gene regulatory networks · Adult stem cells · Tumorigenesis

6.1 Epigenetic Changes During Early Human Development

Our body is composed of some 30 trillion cells forming more than 400 different tissues and cell types. Embryonic development is a tightly regulated process that produces from an identical genome this high diversity of human cell types. In each cell, chromatin serves as a specific filter of genomic information and determines which genes are expressed and which not; *i.e.*, **cellular diversity is based rather on epigenomics than on genomics**. Thus, the differentiation program of embryogenesis is a perfect system for observing the coordination of cell lineage commitment and cell identity specification. Embryogenesis requires the coordination between an increase in cellular mass and the phenotypic diversification of the expanding cell populations. This is under the control of gene regulatory networks in the context of significant changes in the epigenetic landscape (Sect. 1.2).

In the process of fertilization, the two types of haploid gametes, oocyte and sperm, fuse and form the diploid zygote (Fig. 6.1). A series of cleavage divisions of

C. Carlberg, F. Molnár, *Human Epigenetics: How Science Works*,
https://doi.org/10.1007/978-3-030-22907-8_6

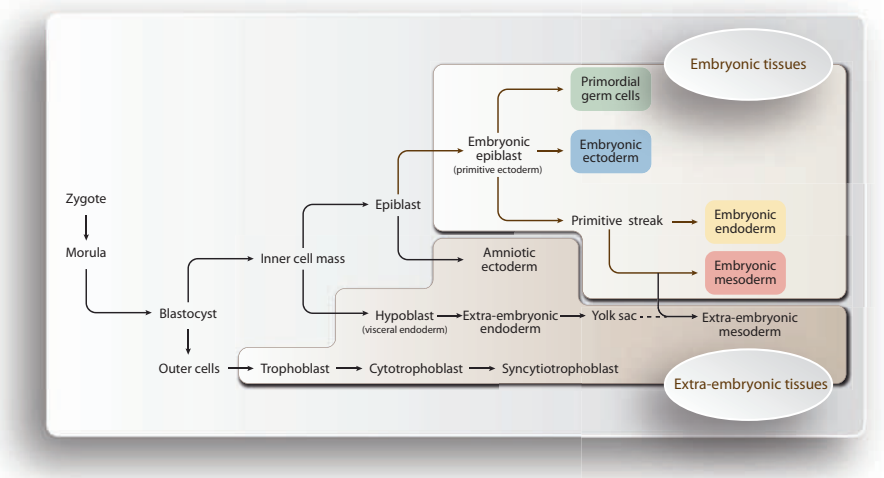


Fig. 6.1 **A road map of early human development**. Details are provided in the text. The dashed line indicates a possible dual origin of the extra-embryonic mesoderm

the zygote creates the totipotent 16-cell morula stage (Box 1.3). A few days after fertilization cells on the outer part of the morula bind tightly together forming a cavity inside. At the blastocyst stage (50–150 cells) a first differentiation occurs. The outer cells (trophoblasts) are the precursors to extra-embryonic cytotrophoblasts that form chorionic villi and syncytiotrophoblasts, which ingress into the uterus; *i.e.*, these cells form the placenta and other extra-embryonic tissues. Even before the blastocyst becomes implanted into the uterine wall, its inner cells, the so-called inner cells mass, begin to differentiate into two layers, the epiblast and the hypoblast. The epiblast gives rise to some extra-embryonic tissues as well as to all the cells of the later stage embryo and fetus. In contrast, the hypoblast is exclusively devoted to making extra-embryonic tissues including the placenta and the yolk sac.

Some of the embryonic epiblast cells form PGCs. These cells are the founders of the germ line and provide a link between different generations of an individual's family (Sect. 7.1). During the gastrulation phase the other cells of the embryonic epiblast turn into the three germ layers ectoderm, mesoderm and endoderm that are the precursors of all somatic tissues. The cells of these germ layers are only multipotent; *i.e.*, they cannot differentiate to every other tissue (Box 1.3). For example, in a series of sequential differentiation steps, ectoderm cells can form epidermis, neural tissue, and neural crest, but not kidney (mesoderm-derived) or liver cells (endoderm-derived).

Before fertilization the CpG methylation rate of the haploid genomes of sperm and oocyte is 90 and 40%, respectively (Fig. 6.2, left). The nucleus of sperm is 10-times more condensed than that of somatic cells, and most histones are replaced by protamines, which are small, arginine-rich, nuclear proteins that allow denser packaging of DNA. As a consequence, the genome of sperm is transcriptionally

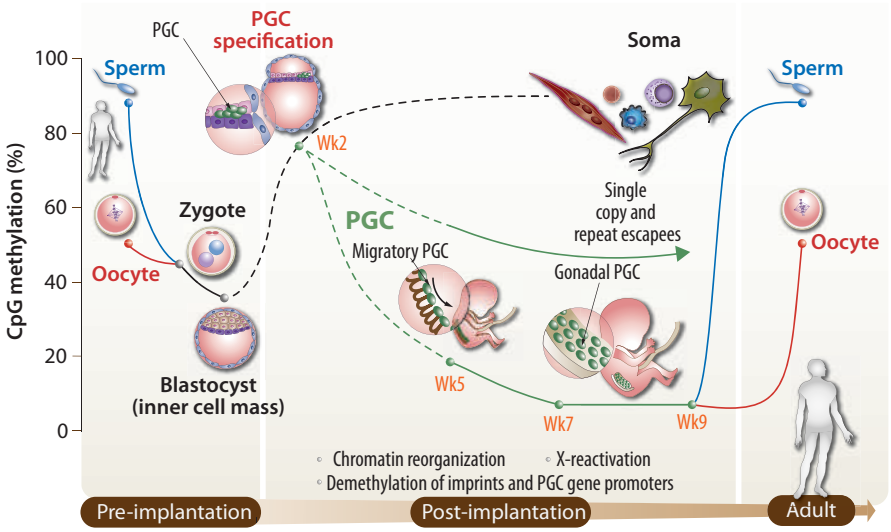


Fig. 6.2 **Epigenetic reprograming during embryogenesis**. DNA methylation is the most stable epigenetic modification and often mediates permanent gene silencing, during embryogenesis as well as in the adult. Accordingly, there is a hierarchy of events where DNA methylation marks are typically added or removed after changes in histone modifications; *i.e.*, they mostly occur at the end of the differentiation process. Therefore, in this graph the percentage of CpG methylation is indicated as representative marker of epigenome changes. Two waves of global demethylation occur during embryogenesis: the first in the pre-implantation phase (week 1) affecting all cells of the embryo, and the other applies only to PGCs during the phase of their specification reaching a minimum of some 5% CpG methylation at weeks 7–9. Some single copy genomic regions and a number of repeat loci remain methylated and are candidates for transgenerational epigenetic inheritance (Sect. 7.1). Dotted lines indicate the methylation dynamics. Genome-wide *de novo* DNA methylation takes place after implantation. In PGCs a second wave of *de novo* DNA methylation happens as early as after week 9 in males, however, as late as after birth in females. Importantly, **most CpG-rich promoters remain unmethylated at all stages of embryogenesis**; *i.e.*, they are not concerned by these waves of methylation and demethylation

silent, while in the oocyte many genes are active. After fertilization, at the zygote stage, the two haploid parental genomes initially remain separate in their rather different state of chromatin organization. Then both genomes get extensively demethylated, but the paternal genome much faster than the maternal genome (Fig. 6.2, left). In parallel, the protamines within the paternal chromatin get exchanged back to canonical histones. In the following cell divisions, preceding the blastocyst stage, there is passive demethylation in both parental genomes, until the methylation patterns are reestablished in a lineage-specific manner. In addition, in the paternal epigenome, there is rapid increase in 5hmC and 5fC/5caC, which is a sign of TET-mediated 5mC oxidation (Sect. 3.1). This process accelerates the demethylation of the paternal epigenome. Importantly, there is no complete demethylation of the epigenome and some 5% of the 5mC marks remain active, possibly *via* protection by methyl-binding proteins, such as MBD proteins (Fig. 6.2, right). **This process serves as the basis for transgenerational epigenetic inheritance** (Sect. 7.1).

Genome-wide epigenetic reprograming during pre-implantation development resets the zygotic epigenome for naïve pluripotency. This process is far more pronounced in PGCs than in other cells of the embryo in order to erase imprints and most other epigenetic memories (Fig. 6.2). Since DNA methylation is an important epigenetic silencer that modulates gene expression and maintains genome stability (Sect. 3.1), the transient loss of DNA methylation in PGCs bears the risk of causing activation of retrotransposons, proliferation defects, and even cell death. Therefore, genome-wide reorganization of repressive histone modifications *via* pluripotency transcription factors safeguards genome integrity in this phase of embryogenesis.

In early embryogenesis the maternal and paternal epigenomes also differ significantly concerning their histone modification patterns. The global pattern of histone marks within the maternal epigenome resembles that of somatic cells, while the paternal epigenome, due to the protamine-histone exchange process, is hyperacetylated, rapidly incorporates histone variant H3.3, and is devoid of H3K9me3 and H3K27me3 markers of constitutive heterochromatin. On the paternal epigenome the first mono-methylation occurs at H3K4, H3K9, H3K27, and H4K20. L, at these positions, different KMTs perform di- and tri-methylation. This takes place also on the maternal epigenome, directly after fertilization, where features of heterochromatin, such as H4K20me3 and H3K64me3 marks, are actively removed, while H3K9me3 marks are lost passively. This initial methylation asymmetry in the paternal and maternal epigenome becomes largely equalized in the course of subsequent development. However, certain genomic regions, such ICRs, stay asymmetric between both alleles, not only on the level of DNA methylation but also concerning histone modifications, such as H3K27me3. This is the **basis for genomic imprinting, *i.e.*, for paternal- and maternal-specific gene expression** (Sect. 3.3).

During pre-implantation development the absence of typical heterochromatin parallels with in general more open chromatin. This keeps the chromatin widely accessible and is necessary for epigenetic reprograming, when gamete-specific modifications are removed and new marks are reestablished. During the course of further development, such as the blastocyst level, cells of the inner cell mass (forming the embryo) show higher levels of DNA methylation and H3K27 methylation, as well as lower levels of histone H2A and/or H4 phosphorylation than cells of the trophectoderm (forming the placenta). This epigenomic asymmetry is a sign of differentiation of the respective cell types and regulates lineage allocation in the early embryo. Accordingly, precise and robust gene regulation *via* the control of access and activity of promoter and enhancer regions is essential (Sect. 6.2).

6.2 Stem Cells and Cellular Pluripotency

Stem cells have the property of both self-renewal, *i.e.*, the ability to run through numerous cell cycles while staying undifferentiated, and the capacity to differentiate into specialized cell types. Thus, stem cells need to be either totipotent or pluripotent (Box 1.3). ES cells are found only in the inner cell mass of blastocysts (Fig. 6.2),

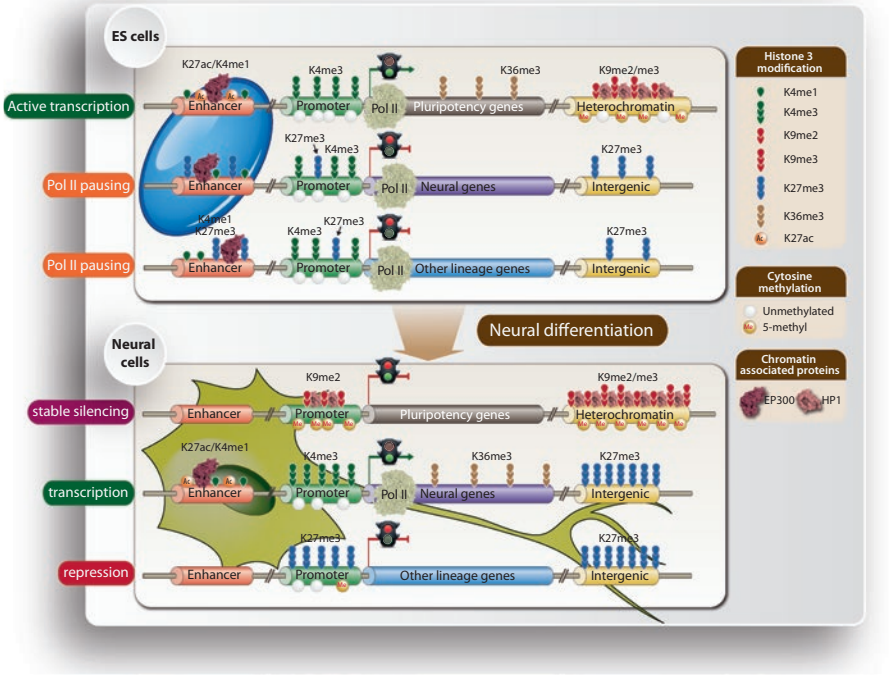


Fig. 6.3 **Chromatin states of ES cells in comparison to lineage-specific cells**. The chromatin stages at enhancers, promoters, gene bodies and intergenic heterochromatin regions of pluripotent genes, neuronal genes and other lineage genes are compared between ES cells (**top**) and, as an example neural cells (**bottom**). More details are found in the text

while many adult tissues contain multipotent stem cells that are able to self-renew and to differentiate into various tissue-specific cell types. Thus, adult stem cells are crucial for tissue homeostasis and regeneration. For example, hematopoietic stem cells (HSCs) (Sect. 11.1) differentiate into myeloid and lymphoid progenitors that give rise to all cell types of the blood and are responsible for the constant renewal of the tissue. The timing of the expression of master transcription factors plays a key role in this differentiation processes. Moreover, most developmental genes are regulated by multiple enhancers having both overlapping as well as distinct spatiotemporal activities. Key lineage-specific genes are associated with dense cluster(s) of highly active enhancers, often referred to as super-enhancers that show stepwise binding of lineage-determining master transcription factors. Thus, **in the process of development from early embryo toward terminally differentiated cells, the epigenome at promoter and enhancer regions changes significantly**.

Genome-wide analysis confirmed that histone marks distinguish ES cells from terminally differentiated cells and pluripotency genes from lineage-specific genes. In ES cells (Fig. 6.3, top), enhancer regions of both pluripotency genes are enriched with H3K4me1 and H3K27ac marks. These genes are actively transcribed, because

also their TSS regions are marked with H3K4me3 and their gene bodies show H3K36me3 modifications. In contrast, the enhancers of lineage-commitment genes carry H3K4me1 marks and repressive H3K27me3 instead of H3K27ac marks, which keeps the genes in a poised state, even if their TSS regions carry H3K4me3 marks. Thus, enhancers and promoters of poised genes (Sects. 4.3 and 5.3) comprise both activating and repressing histone marks; *i.e.*, they are examples of bivalent chromatin states from which they either get fully activated or repressed.

After differentiation toward a specific lineage, such as neurons (Fig. 6.3, bottom), only lineage-specific genes are marked by H3K27ac at both enhancer and promoter regions as well as by H3K4me1 at their enhancers. Then Pol II pausing is released and mRNA transcription continues. Genes of other lineages lose marks at their enhancers and obtain repressive H3K27me3 marks at their TSS regions. Moreover, pluripotency genes attain H3K9me3 marks and DNA methylation at their promoter regions in order to keep them stably silenced for the rest of the life. During the differentiation process, heterochromatin regions are marked by H3K9me2 and H3K9me3 modifications, HP1 binding and DNA methylation are expanded, and thus chromatin becomes more condensed. In repressed genes as well as in intergenic regions H3K27me3 marks also increase. In contrast, during cell lineage commitment, KDMs remove H3K27me3 marks from specific promoter-associated CpGs in order to make the respective genes transcriptionally permissive. This also includes depletion of nucleosomes from TSS regions via chromatin remodelers (Sect. 5.3).

Some 15 years ago, the technique of cellular reprograming of terminally differentiated cells into induced pluripotency, *i.e.*, the creation of so-called iPS cells, was invented. The method uses the ectopic expression (*i.e.*, the abnormally high expression) of the pluripotency transcription factors OCT4, SOX2, KLF4, and MYC in order to reprogram the somatic epigenome and to induce a stable pluripotent state, similar to that of an ES cell. OCT4, SOX2, and KLF4 cooperatively suppress lineage-specific genes and activate pluripotency genes (Fig. 6.3), while MYC overexpression stimulates cell proliferation, induces a metabolic switch from an oxidative to a glycolytic state, and mediates pause release and promoter reloading of Pol II. Nevertheless, the induction of pluripotency is very inefficient, and only 0.1–3% of a cell population get fully reprogramed; *i.e.*, there seems to be a number of epigenetic barriers that stabilize the identity of somatic cells and prevent their aberrant transdifferentiation. However, **cellular reprograming has the potential to regenerate diseased organs and also provides further insight into the principles of epigenetic control of cellular differentiation**.

During normal development from stem cells *via* progenitor cells to terminally differentiated cells, there is a gradual placement of repressive epigenetic marks, such as DNA methylation and histone methylation combined with more restricted accessibility of genomic DNA. Therefore, the overexpression of pluripotency transcription factors results in prominent changes of the epigenetic stage of somatic cells establishing that of pluripotent cells. In view of Waddington's model of an epigenetic landscape (Sect. 1.2), in which a ball rolling downhill represents development from a stem cell to a terminally differentiated cell (Fig. 1.4), cellular

reprograming means to roll back the ball/cell to the top of the hill *via* changes of the epigenome. Moreover, balls/cells may move only a part of the way up the hill and roll back down passing a different "valley", a discrete number of troughs or even travel from one valley to another without going back uphill. This process is termed transdifferentiation and ends up in a different type of terminally differentiated cell.

The first targets of reprograming transcription factors in somatic cells are accessible genomic regions with H3K4me2 and H3K4me3 marks. The following targets are genomic regions with H3K4me1 marks being in a poised state (Sect. 5.2). Reprograming transcription factors act as pioneer factors, *i.e.*, as transcription factors that bind genomic DNA even in the presence of nucleosomes. In contrast, regular transcription factors are not able to nucleosome-covered genomic DNA. Pioneer factors then recruit other transcription factors and chromatin modifiers. Also bivalent genes with active H3K4me3 marks and repressive H3K27me3 marks belong to this category, but they need to be changed from an active state in somatic cells to a poised state typical for pluripotent cells. The most difficult targets for pluripotency transcription factors are heterochromatin regions with repressive H3K9me3 marks. These regions need extensive multi-step chromatin remodeling in order to get transcriptionally activated. Surprisingly, DNA methylation does not play any essential role in cellular reprograming. In contrast, DNA demethylation of pluripotency genes, by either active or passive mechanisms, is crucial for effective reprograming.

6.3 Epigenetic Dynamics During Differentiation

The long ncRNA *Xist* is the key initiator of X chromosome inactivation in female cells (Sects. 3.3 and 5.4) and recruits a series of regulatory complexes at different stages of this X chromosome-wide transcriptional silencing process (Fig. 6.4). In female ES cells, both X chromosomes are actively transcribed and carry markers of active chromatin, such as H3K4me1, H3ac, and H4ac. However, in early embryonic development during the blastula stage of approximately 100 cells, X chromosome inactivation is initiated in one of the two X chromosomes by inducing *Xist* expression, which gradually spreads over the whole chromosome. Through the interaction with heterogeneous nuclear ribonucleoprotein U (HNRNPU) *Xist* recruits the RNA-binding protein SMRT/HDAC1-associated repressor protein (SHARP). The co-repressor protein NCOR (nuclear receptor co-repressor) 2 recruits HDAC3, which finally leads to demethylation of H3K4 and ejection of Pol II. In addition, *Xist* recruits the complexes PRC1 and PRC2 that deposit H2AK119ub and H3K27me3 marks, respectively. Moreover, KMTs add repressive H3K9me2 and H3K9me3 marks. In differentiated cells, X chromosome inactivation is maintained by DNA methylation *via* DNMTs and the incorporation of the histone variant macroH2A.

Differentiating cells share accessible chromatin regions with the ES cell they are derived from, but the similarity in the epigenetic landscape (Sect. 1.2) decreases when cells mature. After commitment to a specific lineage, the cellular repertoire

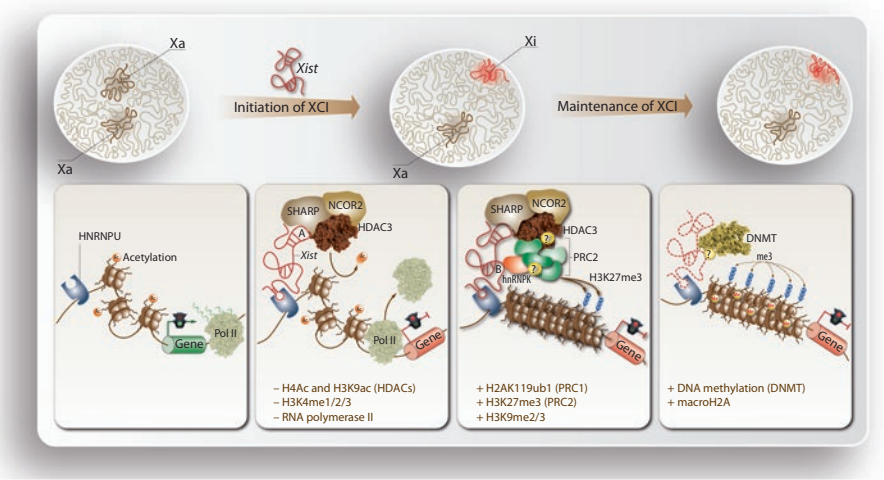


Fig. 6.4 **Mechanisms of *Xist*-induced gene silencing**. In ES cells both X chromosomes are actively transcribed (Xa) and then marked by H3K4me1, H3ac, and H4ac. X chromosome inactivation starts early in embryonic development, when *Xist* expression is randomly initiated on one of the two X chromosomes, and gradually spreads across the whole inactive X chromosome (Xi). More details are provided in the text. XCI, X chromosome inactivation

expands for accessible regulatory regions that contain motifs for transcription factors being specific to that lineage, whereas it clearly decreases for transcription factor binding site of other lineages. Thus, the epigenetic landscape of terminally differentiated cells is constrained by the walls of valleys, the height of which are determined by a gene regulatory network. This network is formed by appropriate levels of DNA methylation and histone modifications as well as by a proper 3D architecture. In this way, cells are prevented from switching states (Fig. 6.5, left). However, **in response to relevant intra- and extra-cellular signals, the epigenome also allows cell state transitions**. When chromatin homeostasis is disturbed, for example, by epimutations (Sect. 8.1), cells do not respond appropriately to these signals. Overly restrictive chromatin networks create epigenetic barriers that prevent all types of cell state transitions (Fig. 6.5, center). In contrast, excessively permissive chromatin networks have very low barriers and allow multiple types of cell state transitions (Fig. 6.5, right). For example, deviations from the norm contribute to tumorigenesis (Sects. 6.4 and 8.4).

Changes in cell identity are reflected by alterations in the usage of the enhancer and promoter regions. Many of the regulatory regions that are active in early embryogenesis lose their activity during differentiation. This is compensated through the activity of TSS regions and poised enhancers, some of which turn into super-enhancers. Changes in enhancer usage require a chromatin topology that allows a new set of enhancers to interact with their target promoters. In parallel, heterochromatin foci become more condensed and more abundant in differentiated cells than in undifferentiated cells. While in ES cells H3K27me3 marks show only

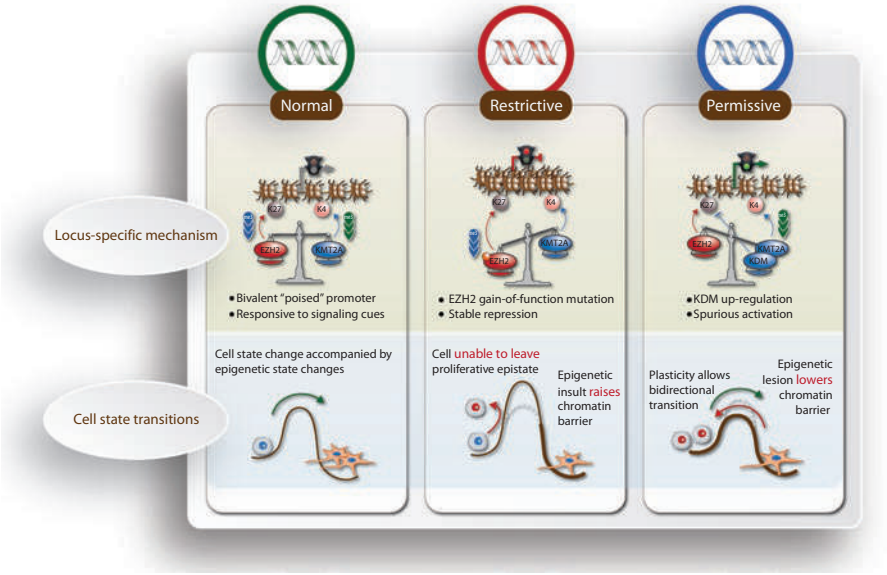


Fig. 6.5 **Chromatin structure, cellular identity and cell state transitions**. In normal cells (**left**) networks of chromatin proteins stabilize the states of cells but also mediate the response to intra- and extra-cellular stimuli and occasionally allow cell state transitions. However, cells in which the chromatin network is perturbed do not respond appropriately. In restrictive chromatin (**center**) epigenetic barriers prevent cell state transitions, while in overly permissive chromatin (**right**) these barriers are lowered and allow easy transition to other cell states. The scenarios are illustrated *via* an example of the underlying molecular mechanisms (**top**) or as cell state transitions (**bottom**). Blue nuclei represent normal cells, while red nuclei indicate cancer cells

focal distributions, in differentiated cells they largely expanded over silent genes and intergenic regions. This results in silencing of pluripotency genes, activating lineage-specific genes and repressing of lineage-inappropriate genes.

On the mechanistic level (Fig. 6.5, top) the scenarios of normal, restrictive and permissive chromatin can be explained, for example, by the actions of a KMT for repressive H3K27me3 marks, such as EZH2 (enhancer of zeste homolog 2, also called KMT6A), and a KMT for activating H3K4me3 marks, such as KMT2A. EZH2 is the catalytic core of the repressive PRC2 complex and KMT2A belongs to the Trithorax group family; both play antagonizing roles in hematopoiesis (Sect. 11.1). In normal cells both KMTs and their histone marks are in balance resulting in bivalent, poised constitutive heterochromatin at TSS regions (Sect. 5.2). This means that their respective target genes are transcribed only in response to appropriate stimuli. In restricted cells, EZH2 may have a gain-of-function epimutation, such as often observed in several forms of lymphoma, resulting in far higher levels of repressive H3K27me3 marks, stable heterochromatin and no gene transcription. In this state, cells may be blocked in differentiation and continue to grow with a high proliferation rate. In contrast, in permissive cells a demethylase, such as KDM6A,

inhibits the action of EZH2 and removes H3K27me3 marks. KDMs are often upregulated under stress conditions. In net effect, this leads to the dominance of H3K4me3 marks and to the activation of gene expression, such as of oncogenes, even in the absence of specific stimuli. In the cell state transition diagram (Fig. 6.5, bottom) the barrier between the cell states is either of medium height in normal cells, very high in restricted cells or rather low in permissive cells.

6.4 Development and Disease

In the homeostasis of adult tissues, dividing and differentiating resident stem cells replace damaged or dying cells when necessary. For example, human epidermis heavily relies on correct stem cell function, both in homeostasis and after wounding. In the epidermis stem cell pools reside exclusively in the basal layer and proliferation occurs only there. The basal layer continuously replenishes the whole tissue with fresh cells that migrate and differentiate through the different epidermal layers. During this epidermal differentiation process, the state of chromatin changes dynamically. Genome-wide levels of H3K27me3 marks decrease during differentiation, in particular at TSS regions of epidermal differentiation genes. This parallels with decreased binding of PRC2 and increased binding of the H3K27me3 demethylase KDM6B. Moreover, in epidermal differentiation the genome-wide histone acetylation level is decreased and terminally differentiated skin cells show higher expression of HDAC1 and HDAC2. Finally, also genome-wide DNA methylation levels decrease during epidermal differentiation.

Changes in the self-renewal of adult stem cells can either result in premature aging (Sect. 7.3), if it is impaired, or in predisposition to malignant transformation, if it is enhanced (Sect. 8.2). Both epigenetic and cellular (re)programing as well as tumorigenesis show similar genome-wide changes in chromatin structure and DNA methylation. Compared with terminally differentiated cells, both iPS and cancer cells have reduced levels of H3K9 methylation and aberrant hyper-methylation or hypo-methylation. For example, reduced methylation levels, such as a result of low *DNMT1* expression, can cause T cell lymphomas but also promote iPS cell formation. Similarly, *DNMT3A* mutations are found in leukemia and knockdown of the gene facilitates human iPS cell formation. Thus, **cancer cells need to overcome some of the same epigenetic barriers as iPS cells in order to alter their cellular states** (Fig. 6.5, bottom).

The interpretation of cellular reprograming and transformation as biochemical reactions illustrates that both processes have to overcome a comparable epigenetic barrier that stabilized the starting cells. Both “reactions” are based on multi-step processes that involve proliferation, change of cell identity and finally lead to the formation of immortal cells with tumorigenic potential. Moreover, compared to terminally differentiated cells, adult stem cells or progenitor cells far more likely form either iPS cells or tumor cells. This suggests that **the epigenetic state of stem and progenitor cells is more sensitive to both cellular reprograming and tumorigenesis**. In addition, pluripotency transcription factors induce a metabolic

switch to create ATP rather from glycolysis than from oxidative phosphorylation, which for cancer cells is known as the "Warburg effect". However, iPS cells keep the normal intact diploid genome of the starting cells, while cancer cells accumulate mutations and often get aneuploidy; *i.e.*, they have an abnormal number of chromosomes.

All four canonical pluripotency transcription factors, OCT4, SOX2, KLF4, and MYC, were found to be amplified or mutated in human cancer; *i.e.*, they belong to the some 500 human cancer genes (Sect. 8.2). Similarly, chromatin modifiers, such as KDM2B, that support cellular reprograming are also associated with tumorigenesis in leukemia and pancreas cancer. In contrast, the expression of histone variants, such as macroH2A, creates a barrier for both iPS cell formation and malignant progression of melanoma cells.

Taken together, the molecular hallmarks of epigenomics, such as histone modifications and DNA methylation, are important for the identity of normal terminally differentiated cells. However, most of these epigenetic hallmarks are reversible, such as by the chemical inhibition of chromatin modifiers enzymes (Sect. 12.3). Therefore, they can be used for the process of cellular reprograming. Thus, this epigenetic process can **serve as basis for the therapy of cancer, neurological, metabolic and immunological diseases**.

Further Reading

Avgustinova A, Benitah SA (2016) Epigenetic control of adult stem cell function. Nat Rev Mol Cell Biol 17:643–658

Li M, Belmonte JC (2017) Ground rules of the pluripotency gene regulatory network. Nat Rev Genet 18:180–191

Perino M, Veenstra GJ (2016) Chromatin control of developmental dynamics and plasticity. Dev Cell 38:610–620

Smith ZD, Sindhu C, Meissner A (2016) Molecular features of cellular reprograming and development. Nat Rev Mol Cell Biol 17:139–154

Tang WW, Kobayashi T, Irie N, Dietmann S, Surani MA (2016) Specification and epigenetic programing of the human germ line. Nat Rev Genet 17:585–600

Yadav T, Quivy JP, Almouzni G (2018) Chromatin plasticity: a versatile landscape that underlies cell fate and identity. Science 361:1332–1336

Population Epigenetics and Aging

7

Abstract

The epigenome has a memory function in both somatic and germ cells. The latter is the basis for transgenerational inheritance. Different types of human cohorts allow studying the impact of epigenetics on a population level. The progressive decline in the function of cells, tissues, and organs associated with aging is affected by both genetic and epigenetic factors; *i.e.*, there are characteristic epigenome-wide changes during aging acting as **epigenetic clocks** over years and decades. In contrast, **epigenetic circadian clocks** in the brain as well as in peripheral tissues coordinate a large set of physiological functions over a day.

Keywords

Epigenetic memory · Epigenetic drift · Transgenerational epigenetic inheritance · Human cohorts · Aging · Epigenetic clock · Circadian clock · Healthspan

7.1 Transgenerational Epigenetic Inheritance

The epigenome is able to preserve the results of cellular perturbations by environmental factors in form of changes in DNA methylation, histone modifications and 3D organization of chromatin. Thus, the epigenome has memory functions. Changes in epigenomic patterns, such as DNA methylation maps, are called epigenetic drifts. They primarily describe the lifelong information recording ("experience") of somatic cell types and tissues, but may also be inherited to daughter cells, when the cells are proliferating. In case of germ cells, epigenetic drifts may be, at least in part, even transferred to the next generation. This leads to the concept of transgenerational epigenetic inheritance, which suggests that the lifestyle of the parent and grandparent generation, such as daily habits in food intake or physical activity, may affect their offspring (Sect. 12.1).

C. Carlberg, F. Molnár, *Human Epigenetics: How Science Works*,
https://doi.org/10.1007/978-3-030-22907-8_7

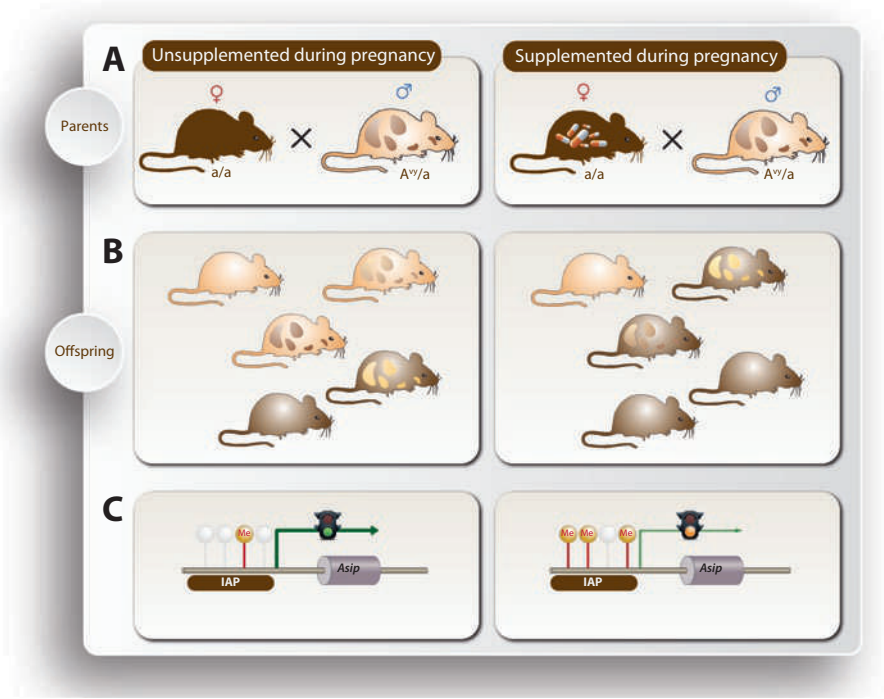


Fig. 7.1 Maternal dietary supplementation affects the phenotype and epigenome of A^{vy}/a offspring. The diets of female wild-type a/a mice are either not supplemented (**left**) or supplemented with methyl-donating compounds (**right**), such as folate, choline, vitamin B12, and betaine, for 2 weeks before mating with male A^{vy}/a mice, and ongoing during pregnancy and lactation (**a**). The coat color of offspring that are born to non-supplemented mothers is predominantly yellow, whereas it is mainly brown in the offspring from mothers that were supplemented with methyl-donating substances (**b**). About half of the offspring does not contain an A^{vy} allele and is therefore black (a/a, not shown here). Molecular explanation of DNA methylation and *Asip* gene expression: maternal hyper-methylation after dietary supplementation shifts the average coat-color distribution of the offspring to brown by causing the IAP retrotransposon upstream of the *Asip* gene to be more methylated on average than in offspring that are born to mothers fed a non-supplemented diet (**c**). White circles indicate unmethylated CpGs and yellow circles are methylated CpGs

The master example of the transgenerational epigenetic inheritance concept is the agouti mouse model. This transgenic mouse carries the retrotransposon IAP (intracisternal A particle) within the regulatory region of the gene *Asip* (agouti signaling protein) (Fig. 7.1). This creates a dominant allele of the *Asip* gene (termed A^{vy}), the expression of which is depending on the methylation level of IAP, *i.e.*, on its epigenetic status. Since the methylation of the IAP retrotransposon happens stochastically, A^{vy} behaves as a metastable epiallele. The *Asip* gene encodes for a paracrine signaling molecule that stimulates hair follicle melanocytes to synthesize yellow pheomelanin pigments instead of black or brown eumelanin pigments. Moreover, the ASIP protein is involved in the neuronal coordination of appetite

resulting in that yellow coat A^{vy} mice become obese and hyper-insulinemic. Heterozygous A^{vy}/a mice vary in their coat colors from yellow via mottled to wild-type dark coat color. When the IAP retrotransposon is methylated, the synthesis of the yellow pigment is downregulated and a dark coat color appears. In contrast, unmethylated IAP allows ubiquitous *Asip* gene expression leading to both yellow coat color and obesity. When A^{vy}/a mice inherit the A^{vy} allele maternally, *Asip* gene expression and coat color correlate with the maternal phenotype. Interestingly, the mottled phenotype indicates that IAP methylation is mosaic; *i.e.*, the *Asip* gene is not expressed in all cells. This suggests that the methylation pattern of the IAP retrotransposon is established early in development and the coat color provides an easy phenotypic readout of the epigenetic status of IAP throughout life. This makes the A^{vy} *Asip* mouse an ideal *in vivo* model for the investigation of a mechanistic link between environmental stimuli, such as nutrition, and epigenetic states of the genome.

The agouti mouse model was used for the following experiment: two weeks before mating with male A^{vy}/a mice, female wild-type a/a mice were either supplemented or not with methyl donors, such as folate, vitamin B12 and betaine (Fig. 7.1, Box 10.1). The supplementation was continued during pregnancy and lactation. While the F1 generation of non-supplemented mothers displayed the expected number of yellow color phenotypes, the offspring of supplemented mothers shifted toward a brown coat color phenotype. This suggests that maternal methyl donor supplementation leads to increased A^{vy} methylation in the offspring. Furthermore, this means that an environmentally induced epigenetic drift in the mothers was inherited to their children. The inheritance of an epigenetic programing to the next generation indicates that at least metastable epialleles, such as IAP, are able to resist the global demethylation of the genome before pre-implantation (Sect. 6.1).

Interestingly, when A^{vy} mice were fed with a soy polyphenol diet causing changes in their DNA methylation patterns, their offspring was protected against obesity and diabetes across multiple generations. In another mouse model, maternal undernutrition leads to low birth weight and glucose intolerance in male and female F1 offspring. Exposure to suboptimal nutrition during fetal development *in utero* leads to changes in the germ cell DNA methylome of male offspring, even when these males were nourished normally after weaning. These phenotypic differences are transmitted through the paternal line to the F2 offspring. More than 100 regions in the F1 sperm genome from maternally undernourished male offspring were found to be hypo-methylated compared to controls. This indicates that PGCs from nutritionally restricted fetuses did not completely remethylate their DNA. Taken together, the different rodent models indicate that **epigenetic memory can be passed from one generation to another by inheriting the same indexing of chromatin marks**. From the different types of chromatin marks, DNA methylation seems to be designed in particular for a long-term cell memory, while short-term "day-to-day" responses of the epigenome are primarily mediated by non-inherited changes in the histone acetylation level. Histone methylation levels are in between both extremes.

The mouse models impose the question whether the concept of an epigenetic memory and inheritance is also valid for humans. There are no comparable natural

human mutants, and for ethical reasons human embryonal feeding experiments are not possible. However, there are natural "experiments", such as the *Dutch Hunger Winter*, where individuals were exposed *in utero* to an extreme undernutrition occurring in the Netherlands during the winter of 1944/45. Fetal malnutrition led to impaired fetal growth. Low birth weight favors a thrifty phenotype that is epigenetically programed to use nutritional energy efficiently, *i.e.*, to be prepared for a future environment with low resources during adult life. Even many decades after birth the *in utero* undernourished individuals showed subtle (<10%) changes in DNA methylation at several loci in adulthood, for example, at the regulatory region of the imprinted gene *IGF2* (Sect. 3.3). This epigenetic pattern is associated with an increased risk of obesity, dys-lipidemia and insulin resistance, when the respective individuals are exposed to an obesogenic environment. Accordingly, **for humans there seems to be the same link between pre-natal nutrition and epigenetic changes as described for rodents**. Similarly, the famine of 1959–61 in China largely contributed to the overproportionally high raise of type 2 diabetes in the country. These examples led to the Developmental Origins of Health and Disease (DOHaD) concept indicating that early developmental events, such as perturbations of the nutritional state *in utero*, have significant effects on disease risk as adult. Thus, **environmental exposures of humans, in particular during early life, can be stored as epigenetic memory**.

Interestingly, the amount of brown adipose tissue that we carry in our body and using for thermogenesis, correlates with the month when we were conceived. Individuals conceived in cold months have significant differences in brown fat characteristics and metabolic phenotypes than those, who were conceived in warm months or environment. An analogous mouse experiment confirmed the observation in rodents and demonstrated that only the cold exposure of the fathers before conception affected the amount of brown adipose tissue in the offspring.

7.2 Population Epigenetics

The field of epigenetic epidemiology, *i.e.*, the study of epigenetics in populations, combines epigenome-wide methods (Sect. 2.5) with population-based epidemiological approaches. Epigenetic changes can occur at any time during life, although increased sensitivity exist during early embryogenesis (Sect. 6.1, Fig. 7.2). **Our epigenome primarily changes due to environmental exposures but also based on stochastic epigenetic drifts associated with aging** (Sect. 7.3). Epigenetic epidemiology studies different types of human cohorts with the goal to identify both the causes as well as the phenotypic consequences of epigenomic variations. These are cohorts of natural experiments, such the *Dutch Hunger Winter*, longitudinal birth cohorts, longitudinal studies on mono-zygotic twin cohorts, pre-natal cohorts and *in vitro* fertilization conception cohorts.

Studies of families are well suited to investigate epigenetic changes in the offspring that may be based on environmental exposures of parents during gametogenesis. Birth cohorts and *in vitro* fertilization cohorts track life from as early as

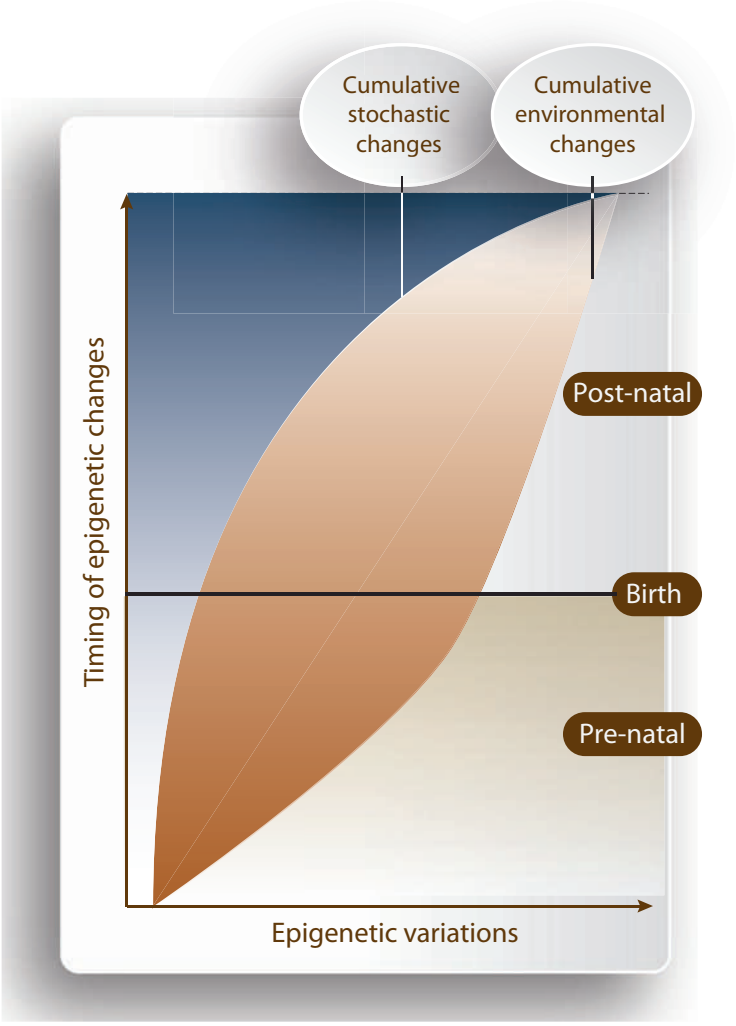


Fig. 7.2 **Epigenetic variation in populations**. Epigenetic changes can occur at any time during life, but there is significantly increased sensitivity during early pre-natal development

peri-conception (*i.e.*, around the time of conception) and allow the study of epigenetic changes based on the pre-natal environment and their association with disease phenotypes early in life. Cohorts based on natural experiments, *i.e.*, when the exposure to severe environmental conditions was not under experimental control, enable the investigation of the link of environmental exposures in early life with the onset of disease phenotypes decades later. Prospective cohorts, in particular those involving mono-zygotic twins having identical genomes, parents, birth date and gender, study in a longitudinally way the contribution of age-related epigenetic modifications in common diseases, such as type 2 diabetes, autoimmune diseases, cancer,

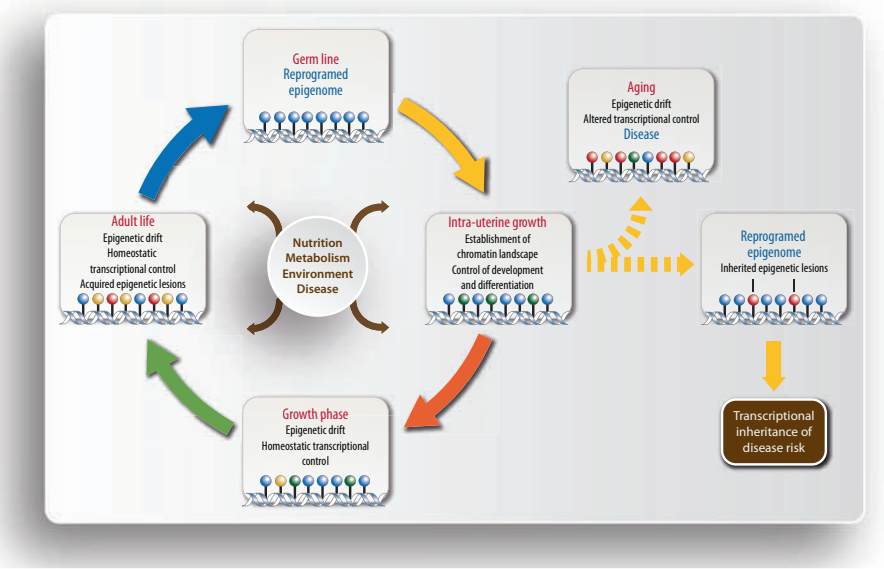


Fig. 7.3 **Epigenetic drift and transgenerational inheritance**. During embryogenesis epigenetic marks, such as DNA methylation and histone modifications, are established in order to maintain cell lineage commitment (Sect. 6.1). After birth, this epigenetic landscape stays dynamic throughout lifespan and responds to nutritional, metabolic, environmental, and noxious signals. Epigenetic drifts are part of homeostatic adaptations and should keep the individual in good health. However, when an adverse epigenetic drift compromises the capacity of metabolic organs to adequately respond to challenges as provided by nutrition and chronic inflammation, the susceptibility to diseases, such as type 2 diabetes or cancer, increases (Sect. 12.1). Some of these acquired epigenetic marks can be inherited to subsequent generations when they escape epigenetic reprograming during gametogenesis

Alzheimer's and autistic spectrum disorders. Interestingly, twin studies demonstrated that epigenetic variations significantly increase across lifespan. Short-term interventions, such as dietary studies, can identify specific environmental exposures that lead to tissue-specific epigenetic changes. Finally, the follow-up of long-lived families helps to identify the impact of epigenetics for healthy aging (Sect. 7.3).

Epigenetic drifts, such as hyper-methylation of CpG islands close to the regulatory regions of tumor suppressor genes, contribute to the risk for cancer (Sect. 8.1) and other diseases (Fig. 7.3). In particular the risk for diseases that are related to the exposure with environmental factors, such as microbes causing inflammation (Sect. 11.2) or overeating leading to obesity and type 2 diabetes (Sect. 10.4), have a large epigenetic contribution. Of special interest are diseases that have their onset a long time before the phenotype emerges, *i.e.*, where accumulations of epigenetic changes stepwise increase disease susceptibility.

Epigenetic information can also be transmitted *via* the germ line to the next generation. Although some 95% of DNA methylation marks are erased throughout the two rounds of demethylation during PGC generation (Sect. 6.1), some single-copy genomic regions escape both demethylation waves and remain methylated in

gametes. In case the DNA methylation pattern at these escape regions is susceptible to environmental influences, such as dietary molecules, **lifestyle choices of an individual may be transmitted to subsequent generations and may lead to phenotypic consequences**.

7.3 Epigenomics of Aging

Human lifespan comprises a period of growth and differentiation that ends up in sexual maturity, *i.e.*, in a period of maximal fitness and fertility, and a period of aging that comes with loss of function at the various levels of cells, tissues, and the organism as a whole. Due to a rather long time until sexual maturity (12–15 years) and some 15–20 years of childcare, evolution selected us to be for some 45 years mainly free of non-communicable diseases, such as cancer. Thus, we have a kind of "warranty" given by nature to get some 45 years old in order to guarantee the survival of our offspring. The remaining 75 years up to our maximal lifespan of 120 years can be considered as security measure. However, this latter phase is not under evolutionary control and therefore associated with a wide range of diseases, including metabolic disorders, cardiovascular and neurodegenerative diseases, and many cancers. Accordingly, healthspan is defined as the period of disease free health, represented by high cognition and mobility. Thus, understanding the changes that occur during aging, *i.e.*, the hallmarks of aging, and **identifying regulators of lifespan and healthspan is a key question for each of us**.

Individuals have a personal rate of aging that depends on gender (women tend to live longer than men), lifestyle choices, such as smoking or physical inactivity, and many environmental factors. There is a genetic basis for longevity, but the non-genetic contribution to aging is estimated to be more than 70% (Sect. 12.1). For example, some molecular markers of age, such as telomere length or the expression of genes in metabolic and DNA repair pathways, are sensitive to environmental stress. Moreover, various animal aging models suggested that calorie restriction, low basal metabolic rate, increased stress response and reduced fertility play a major role in determining the lifespan of individuals (Sect. 10.2). The molecular basis of all non-genetic factors of aging is cellular perturbations that result in the modulation of signal transduction pathways affecting the epigenome; *i.e.*, **epigenetic changes are a major contributor to the aging process**. Thus, not only diseases but also aging result in epigenetic drifts. Changes in the epigenome also can occur spontaneously (*i.e.*, stochastically) without the contribution of a cellular perturbation. Nevertheless, the likelihood for stochastic changes of the epigenome is increased by chemicals disrupting DNA methylation marks or by errors in copying the methylation status during DNA replication. Such chemicals are

- metals like cadmium, arsenic, and mercury
- peroxisome proliferators
- air pollutants, such as black carbon and benzene
- endocrine disruptors, such as diethylstilbestrol, bisphenol A, and dioxin.

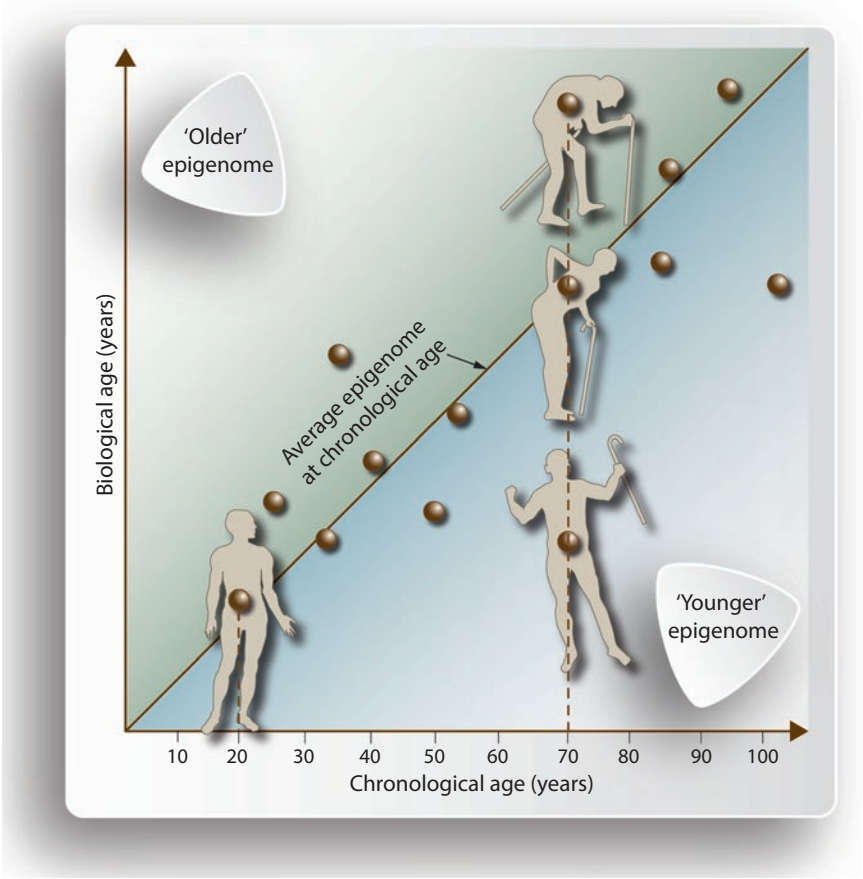


Fig. 7.4 **Epigenetic biomarkers of age**. Epigenomic patterns not only monitor cellular identities but also cellular health and age. For example, changes in the DNA methylation status of CpG islands, as measured in PBMCs (also referred to as liquid biopsies) taken from individuals of different age (symbols), can serve as a sensor for chronological age. However, there is significant deviation from the postulated linear fit (diagonal line), which suggests that methylation patterns also represent biological age

Changes of the epigenome, in particular the DNA methylome, are associated with the chronological age of an individual as well as with age-related diseases. The DNA methylation status at a few hundred key CpG islands, which can be measured from easily accessible tissues and cell types, such as skin or peripheral blood mononuclear cells (PBMCs), can serve as biomarkers. The respective chromatin landscapes are then correlated with chronological age and biological age, *i.e.*, the age at which the population average is most similar to the individual. DNA methylome profiling of a large cohort of individuals spanning over a wide age range provides a good correlation between chronological and biological age, but there are also significant inter-individual variations. At a given chronological age the investigated tissue of some individuals has a far "younger" epigenome, while that of others is

already "older" (Fig. 7.4). Accordingly, it can be expected that the latter individuals may have an earlier onset of age-related diseases from the respective tissue and eventually may die in younger years than the former individuals. This has been observed with individuals suffering from progeroid syndromes. In contrast, the blood of the offspring of super-centenarians, *i.e.*, individuals who reached an age of at least 105 years, has a lower epigenetic age than that of age-matched controls. Thus, **epigenomic signatures can serve as biomarkers of aging** and may be druggable targets in order to delay or reverse age-related disease.

The biological age of a specific tissue, for example, that of the liver of an obese person, may be significantly higher than that other tissues, such as PBMCs or muscle, from the same person. The DNA methylation-based age of a reference tissue, such as PBMCs, can be considered as an epigenetic clock. The epigenetic clock is a more accurate predictor for mortality by all causes in later life than other biomarkers of aging, such as changes in telomere length. For comparison, ES and iPS cells show in these assays as "ageless". Moreover, sperm cells might be classified "younger" than somatic cells from the same individual. Epigenetic clocks of mice (average lifespan some 2 years) tick faster than those of humans (average lifespan some 80 years). Moreover, a quantitative model of the aging methylome is able to distinguish relevant factors in aging, including gender and genetic variants.

Genomic instability is a hallmark of aging as well as of tumorigenesis (Sect. 8.1). The accumulation of DNA mutations or even aneuploidy, as during aging, has an effect on both the transcriptome and the epigenome. While cells of young individuals show a robust transcriptome and normal chromatin states, with increasing age the transcriptome gets instable and aberrant chromatin states are accumulating. For example, DNA damage stimulates the recruitment of chromatin modifiers that may induce abnormal chromatin states. In turn, epigenomic changes during aging can increase the susceptibility of the genome to mutations and in parallel reduce the precision of transcription. Moreover, errors in DNA repair and failure to correctly replicate the genome and epigenome not only increase the number of DNA mutations but also of epimutations. Thus, since genome surveillance and epigenetic remodeling influence each other, **environment-induced epigenomic instability throughout life is an important driver of the aging process**.

In addition to changes in DNA methylation, other key epigenetic hallmarks of aging are

- a general loss of histones due to local and global chromatin remodeling
- an imbalance of activating and repressive histone modifications
- site-specific loss and gain of heterochromatin
- significant nuclear reorganization
- changes in the transcriptome.

The general loss of histones in aging cells is tightly linked to cell division. For example, nuclear blebs of senescent human cells contain a large number of histones. Senescence is a cell cycle arrest in response to stress, which is linked to age-related tissue decline. Cells then develop senescence-associated heterochromatin foci,

which are regions of highly condensed chromatin showing heterochromatic histone modifications, heterochromatic proteins and histone variant macroH2A (Box 4.1). The epigenome of senescent cells shows an increase of H3K4me3 and H3K27me3 marks within LADs and a loss of H3K27me3 outside of LADs. The most significant molecular consequence of the loss of repressive histone marks and the gain of activating marks during aging is a change in gene expression, such as the upregulation of genes related to cell proliferation, cell adhesion and ribosomal proteins, while genes related to cell cycle, DNA base excision repair and DNA replication are downregulated.

Constitutive heterochromatin at telomeres, centromeres and pericentromeres is established during embryogenesis (Sect. 6.1) and is maintained throughout lifespan. However, senescent cells loose some of these regions of constitutive heterochromatin resulting in growth of euchromatic regions. Moreover, a loss of the nuclear lamina stimulates the breakdown of heterochromatin organization and the relocalization of heterochromatic proteins to regions in the genome where they contribute to the formation of region-specific foci. Laminopathies display deleterious changes to nuclear organization, such as the expression of progerin, a truncated dominant-negative lamin A protein, in the Hutchinson-Gilford progeria syndrome (HGPS). Cells from HGPS patients show abnormalities in nuclear morphology, defect DNA damage repair, changes in chromosome organization, increased rates of cellular senescence and many alterations in heterochromatin proteins, such as low levels of HP1, H3K9me3, and H3K27me3 and increased levels of H4K20me3. Moreover, the deficiency in the gene *WRN* (Werner syndrome RecQ like helicase) causes the progeroid Werner syndrome and leads to global loss of chromatin compaction, decreased H3K9me3 and H3K27me3 levels and increased phosphorylation of the histone variant H2A.X at centromeres.

7.4 Epigenetics of the Circadian Clock

Light-sensitive organisms, like humans, synchronize their daily behavioral and physiological rhythms with the rotation of Earth on its axis; *i.e.*, they display circadian (meaning "approximately one day") activity cycles. These rhythms are generated by the suprachiasmatic nucleus (SCN) of the hypothalamus (Fig. 7.5a). The core of this molecular 24 h clock is a series of transcription-translation feedback loops of the transcription factors ARNTL (aryl hydrocarbon receptor nuclear translocator-like, also called BMAL1) and CLOCK (clock circadian regulator) and their co-repressor proteins. Interestingly, CLOCK also has HAT activity with specificity for H3K9 and H3K14; *i.e.*, this central circadian regulator is a chromatin modifier and a direct **link of the molecular clock to epigenetics**. The ARNTL-CLOCK complex activates in a circadian fashion the expression of hundreds of genes both in the brain and in peripheral metabolic tissues, including also the genes *PER1* (period circadian clock 1) and *CRY1* (cryptochrome circadian clock 1). PER1 and CRY1 form a heterodimeric complex co-repressor complex that inactivates ARNTL-CLOCK creating a negative feedback loop (Fig. 7.5b). The genes *NR1D1*

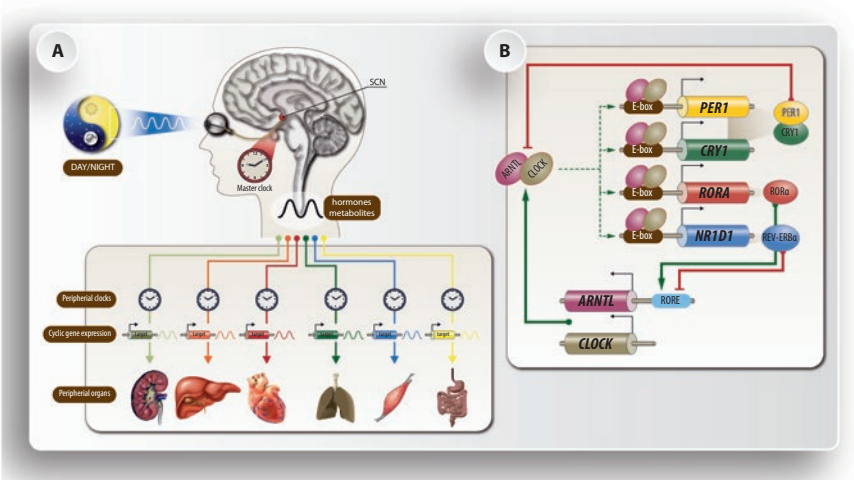


Fig. 7.5 **The circadian clock in humans**. Electrical and humoral signals from the SCN within the brain synchronize phases of circadian clocks in peripheral organs, which then generate time-dependent rhythms in gene expression, metabolism and other physiological activities (**a**). In the feedback loop of the molecular circadian oscillator, positive elements, such as the transcription factors ARNTL, CLOCK, and RORα, are shown in green, and negative elements, such as PER1, CRY1, and REV-ERBα, in red (**b**). The combined actions of hundreds of ARNTL-CLOCK target genes provide a circadian output in physiology

and *RORA*, encoding for the nuclear receptors REV-ERBα and RORα, are further targets of ARNTL-CLOCK. REV-ERBα negatively and RORα positively regulates the expression of the *ARNTL* gene; *i.e.*, both transcription factors form additional feedback loops in the control of the molecular clock.

The molecular clock is a self-sustained activity, but circadian gene transcription can also be modulated by metabolites, in particular by those representing energetic flux. For example, the AMP sensor AMP-activated protein kinase (AMPK, Sect. 10.3) connects the internal clock function to the nutrient state *via* phosphorylation and subsequent proteasomal degradation of CRY1, which reactivates ARNTL-CLOCK. Moreover, the cyclical activity of ARNTL-CLOCK is modulated by the demethylase KDM5A, which in turn is linked via its co-factors iron and α-ketoglutarate to cellular redox and mitochondrial energetics. A bidirectional interaction between circadian and metabolic signaling is the inhibition of ARNTL-CLOCK by the NAD^+-dependent deacetylase SIRT1. NAD^+ is an electron shuttle in oxidoreductase reactions and also acts as a co-factor in ADP-ribosylation modifications (Sect. 10.2). This represents another feedback control of the molecular clock, since the gene encoding for the critical enzyme for NAD^+ synthesis, nicotinamide mononucleotide phosphoribosyltransferase (NAMPT), is a direct ARNTL-CLOCK target. Thus, **epigenetics of the molecular clock links the sensing of the energy status of cells with circadian changes in gene expression.**

Epigenetics of the circadian clock coordinate daily behavioral cycles of sleep-wake and fasting-feeding with anabolic and catabolic processes in the periphery.

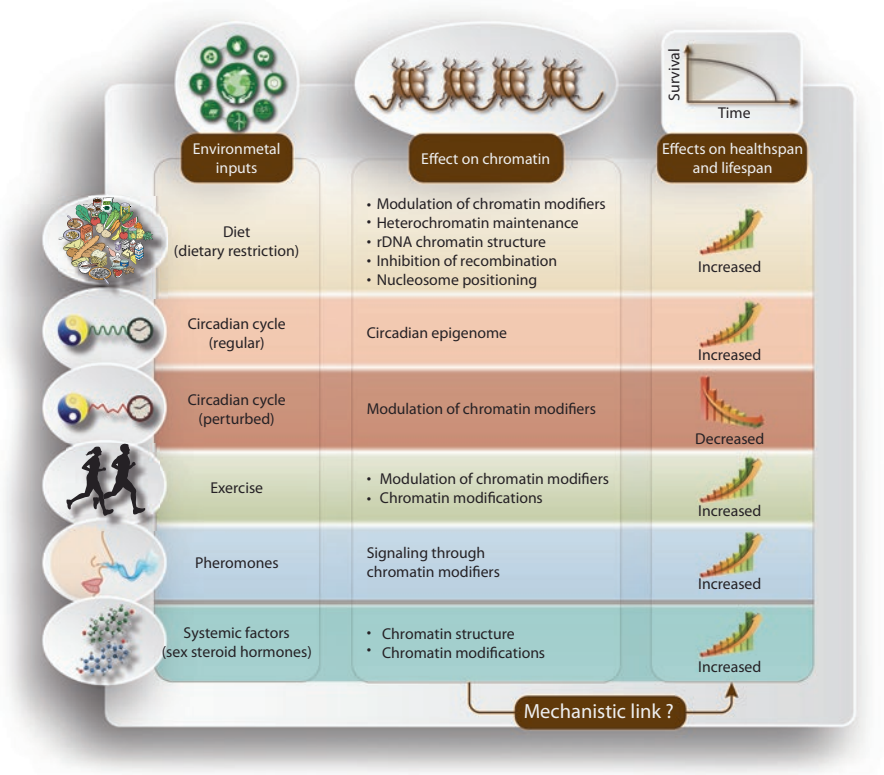


Fig. 7.6 **Effects of environmental inputs on longevity and chromatin**. Many environmental signals that modulate lifespan also affect chromatin. These are dietary restriction, the circadian cycle, physical activity, and sex steroid hormones. More details are provided in the text

NAD^+ oscillation, redox flux, ATP availability, and mitochondrial function are used in order to influence acetylation and methylation reactions of chromatin modifiers (Sect. 10.2). Dietary restriction is known to extend the lifespan in many model organisms, such as yeast, worms, flies, and even primates. Gene expression changes related to dietary restriction promote global preservation of genome integrity and chromatin structure, such as maintenance of heterochromatin (Fig. 7.6). Nutrition signaling pathways *via* insulin, the amino acid sensor TOR (target of rapamycin) and NAD^+-sensing SIRTs, all those integrate metabolic signals into chromatin responses. They inform the epigenome on nutrient availability and thus have a key role in determining lifespan. Since artificial light, shift work, travel and temporal disorganization disrupt the alignment between the external light-dark cycle and their internal clock, they cause a disadvantage for metabolic health. Longitudinal population studies and clinical investigations both indicated an association between shift work and diseases, such as type 2 diabetes, gastrointestinal disorders and cancer, that can be modulated by changes in the circadian rhythm. Furthermore, the

habit of altering bedtime on weekends, the so-called "social jet lag", has been associated with increased body weight.

Physical activity promotes healthy aging, as it prevents cognitive decline and is associated with a 30% reduction in all-cause mortality. Moreover, it induces changes in the chromatin of skeletal muscles, such as increased H3K36ac levels and the cellular localization of HDAC4 and HDAC5. Thus, **physical activity has direct effects on the epigenome**. Interestingly, the signaling of pheromones, which are compounds triggering a response in other members of a given species, works via chromatin modifiers; *i.e.*, pheromones have an effect on the epigenome. The systemic levels of sex steroid hormones (Fig. 7.6), such as estrogens in females and androgens in males, decline with age. For example, estrogens reduce the risk for age-related diseases, such as osteoporosis, sarcopenia (muscle weakness), cardiovascular diseases, reduced immune function, and neurodegeneration, while lower levels of steroid hormones increase the prevalence for these diseases. Estrogens and androgens act *via* their specific nuclear receptors, the ligand-responsive transcription factors estrogen receptor α and β and androgen receptor, respectively. These transcription factors are well characterized for their interaction with chromatin modifiers and remodelers; *i.e.*, their activation as well as the lack of their activity has a direct impact on the chromatin at the local regions of their genomic binding sites.

Further Reading

Benayoun BA, Pollina EA, Brunet A (2015) Epigenetic regulation of ageing: linking environmental inputs to genomic stability. Nat Rev Mol Cell Biol 16:593–610

Horvath S, Raj K (2018) DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat Rev Genet 19:371–384

Jirtle RL, Skinner MK (2007) Environmental epigenomics and disease susceptibility. Nat Rev Genet 8:253–262

Heard E, Martienssen RA (2014) Transgenerational epigenetic inheritance: myths and mechanisms. Cell 157:95–109

Radford EJ (2018) Exploring the extent and scope of epigenetic inheritance. Nat Rev Endocrinol 14:345–355

Sen P, Shah PP, Nativio R, Berger SL (2016) Epigenetic mechanisms of longevity and aging. Cell 166:822–839

Sun W, Dong H, Becker AS, Dapito DH, Modica S, Grandl G, Opitz L, Efthymiou V, Straub LG, Sarker G et al (2018) Cold-induced epigenetic programming of the sperm enhances brown adipose tissue activity in the offspring. Nat Med 24:1372–1383

Taudt A, Colome-Tatche M, Johannes F (2016) Genetic sources of population epigenomic variation. Nat Rev. Genet 17:319–332

Zimmet P, Shi Z, El-Osta A, Ji L (2018) Epidemic T2DM, early development and epigenetics: implications of the Chinese famine. Nat Rev Endocrinol 14:738–746

Cancer Epigenetics

8

Abstract

Compared to normal cells, cancer cells show **epigenetic drifts**, which are genome-wide changes in DNA methylation, histone modifications, and 3D chromatin structure representing **epimutations**. Moreover, many tumors reactivate programs of fetal development; *i.e.*, tumorigenesis is associated with **epigenetic reprograming**. The mechanistic bases of cancer epigenetics are specific genetic, environmental, and metabolic stimuli that disrupt the homeostatic balance of chromatin, which then either becomes very restrictive or permissive. Recent drug discovery efforts targeted the epigenome and inhibitors of chromatin modifiers are tested in clinical trials and some were already approved for therapy.

Keywords

Epimutation · Epigenetic reprograming · Tumorigenesis · Hallmarks of cancer · Epigenetic modifiers · Epigenetic mediators · Epigenetic modulators · Epigenetic therapy

8.1 Epimutations in Cancer

One in two of us will face during our lifespan the diagnosis of cancer, *i.e.*, the detection of a malignant tumor. Cancer is not a uniform disease but rather a collection of hundreds of different types of hyperplasia. Moreover, behind every newly diagnosed tumor there is an individual history of some 10–20 years of tumorigenesis. Cancer is typically considered as a disease of the genome caused by the accumulation of DNA point mutations as well as translocations and amplifications of larger genomic regions. However, tumorigenesis also comes along with abnormalities in cellular identity, different responsiveness to internal and external stimuli and major changes in the transcriptome, all of which are based on changes of the epigenome. In fact, most types of cancer carry mutations both in their genome and epigenome. For

C. Carlberg, F. Molnár, *Human Epigenetics: How Science Works*,
https://doi.org/10.1007/978-3-030-22907-8_8

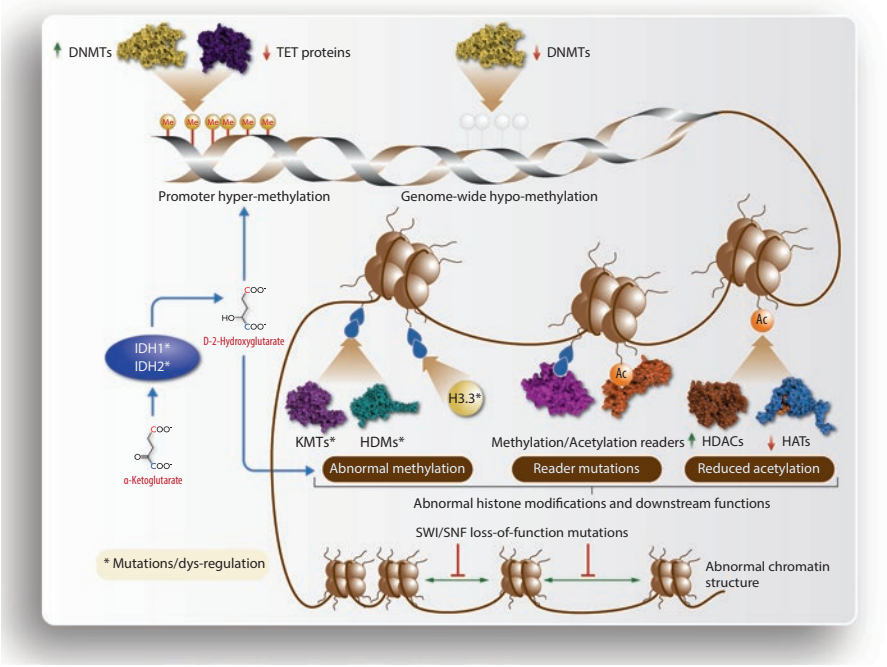


Fig. 8.1 **Epimutations in cancer**. There are four major types of epimutations affecting cancer: DNA hyper-methylation at promoters (**top left**), genome-wide DNA hypo-methylation (**top right**), abnormal modification of histones and/or their recognition (**center**), and abnormal chromatin structures caused by mal-functional chromatin remodelers (**bottom**). More details are provided in the text

example, large-scale cancer genomic projects, such as *TCGA* (Sect. 2.5), revealed that more than 50% of human cancers carry mutations in key chromatin-associated proteins. It is important to realize that **the epigenetic signature of a cell allows more variation than its primary genetic status**, since the error rate in inheritance of DNA methylation is some 4% for a given CpG per cell division, while the mutation rate of the genome during DNA replication is far lower. Thus, epigenetic variability leads in much shorter time to phenotypic selection, such as possible onset of cancer, than genetic mutations in a traditional view.

Tumorigenesis is a multistep process, in which a multitude of molecular alterations contribute to the initiation and progression of a tumor. As already briefly addressed in Sects. 3.4 and 6.4, dys-regulated epigenetic processes can act as drivers of early disruption of cellular homeostasis in pre-cancerous and cancerous cells. An important epigenetic change in cancer is the deregulation of CpG methylation patterns, *i.e.*, of the DNA methylome (Fig. 3.5). Like in aging (Sect. 7.3), tumorigenesis correlates with genome-wide DNA hypo-methylation (Fig. 8.1, top right), which leads to genome instability *via* the reactivation of pluripotency transcription factors (acting as oncogenes) and retrotransposons within repetitive DNA. In contrast, CpG islands at promoter regions of tumor suppressor genes get

hyper-methylated (Fig. 8.1, top left), which causes the inactivation of the respective genes and their tumor-protective function. This is an example of an epigenetic drift (Sect. 7.1). For example, silencing of the tumor suppressor gene *MLH1* (MutL homolog 1) via DNA hyper-methylation harms the DNA mismatch repair process and increases the risk of accumulating DNA mutations throughout the whole genome. Thus, **an epimutation can initiate a multitude of genetic changes**.

Epigenetic mutations in genes encoding for chromatin modifiers are either gain-of-function or loss-of-function. Abnormal histone methylation may be caused by mutations in genes encoding for KMTs and KDMs as well as for the histone variant H3.3, which reduces the genome-wide methylation of H3K27 and H3K36 (Fig. 8.1, center). Examples are gain-of-function and overexpression of EZH2 (a H3K27-specific KMT) and loss-of-function of the H3K36-specific KMT SETD2 (SET domain containing 2). Moreover, there are translocations of KMT2A (H3K4-specific) as well as translocations and overexpression of the H3K36-specific KMT NSD1 (nuclear receptor binding SET domain protein 1) and the H3K27-specific KMT NSD2. In addition, the amplification or overexpression of genes encoding for H3K4-, H3K9- and H3K36-specific KDMs have been described in the context of different types of cancer. Furthermore, also histone acetylation is reduced in cancer through the loss of the HATs EP300 (KAT3B) and CREBBP (CREB binding protein, also called KAT3A) and the overexpression of HDACs. Finally, not only the writer and eraser function of chromatin modifiers can be affected by mutations, but also their reader function. Examples are the overexpression or gain-of-function translocations of BRD4, which binds acetylated histones, or the overexpression of the PHD family member TRIM24 (tripartite-motif-containing protein 24), which recognizes H3K23ac.

Loss-of-function mutations in genes encoding for DNA demethylases (*TET1*, *TET2* and *TET3*) or increased expression of genes encoding for DNMTs (*DNMT1*, *DNMT3A* and *DNMT3B*) can cause promoter hyper-methylation in some cancers (Fig. 8.1, top). In contrast, genome-wide hypo-methylation is often based on loss-of-function mutations in the *DNMT3A* gene. Mutations in the genes encoding for the metabolic enzymes IDH (isocitrate dehydrogenase) 1 and 2 cause that the citric acid cycle intermediate α-ketoglutarate becomes transformed into the oncometabolite 2-hydroxyglutarate, which inhibits TETs and KDMs (Fig. 8.1, center left). This leads to the increased methylation of both DNA and histones. Taken together, **epimutations affect a large variety of changes in cellular homeostasis that result in the acceleration of tumorigenesis**.

8.2 Hallmark of Cancer: Epigenomic Disruption

The concept of the **hallmarks of cancer** highlighted the processes “sustained proliferative signaling”, “evading growth suppressors”, “resisting cell death”, “enabling replicative immortality”, “inducing angiogenesis”, and “activating invasion and metastasis” as common during tumorigenesis of basically all types of cancer (Fig. 8.2). Later on, the concept was extended to 10-12 hallmarks, including

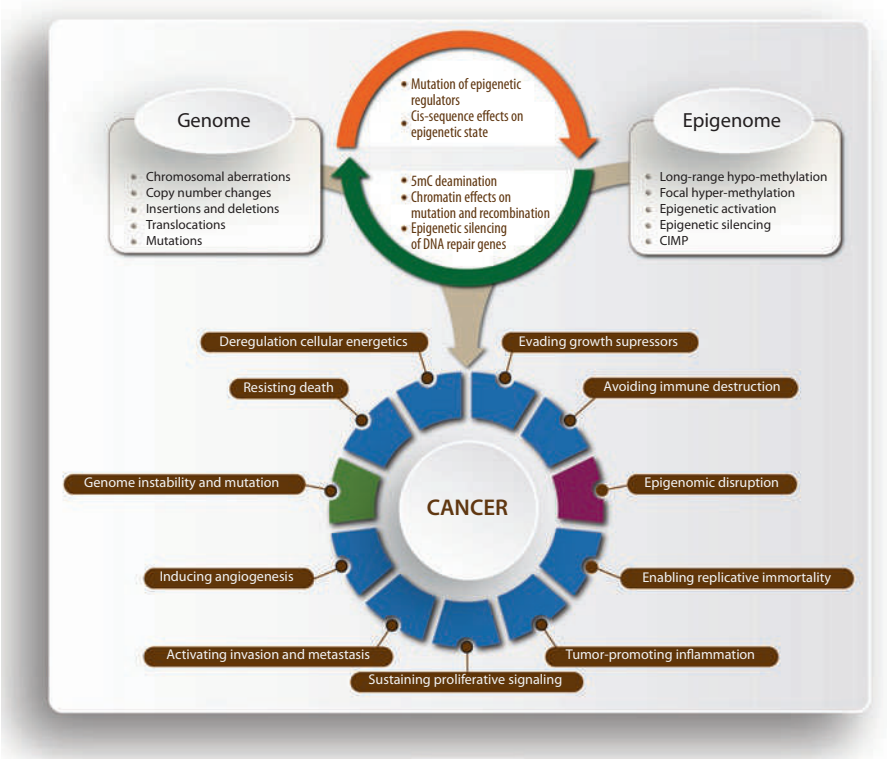


Fig. 8.2 **Interplay between genome and epigenome in cancer**. Changes in the genome can influence the epigenome and *vice versa*. This forms a network that produces genetically or epigenetically encoded variations in the phenotype that are subject to Darwinian selection for growth advantage and thus eventually achieving the hallmarks of cancer. CIMP, CpG island methylator phenotype

"genomic instability and mutation" and "**epigenomic disruption**". This acknowledges the results of large-scale cancer genomics projects, such as *TCGA*, on both genetic and epigenetic drivers of different types of cancer. For example, the cancer projects indicated high frequency of mutations in genes encoding for epigenetic mediators, some of which form hot spots.

The cancer genome and epigenome influence each other in a multitude of ways and can work mutually (Fig. 8.2). Both genetics and epigenetics offer complementary mechanisms to achieve similar results, such as the inactivation of tumor suppressor genes by either deletion or epigenetic silencing. Moreover, a gain-of-function activation of the oncogene *PDGFRA* (platelet-derived growth factor receptor α) that is important for achieving the hallmark "sustained proliferative signaling" may be based either on a genetic mutation within the coding region of the gene or by an epimutation that disrupts insulators at the borders of the TAD carrying the gene. Changes in DNA methylation patterns are key epigenetic dys-regulations occurring during tumorigenesis. Compared with normal cells of the same individual, the

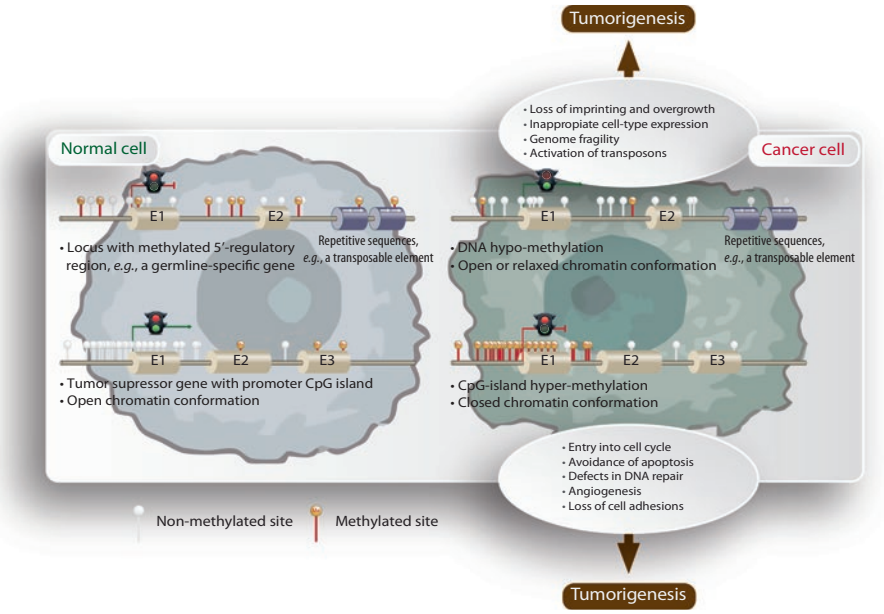


Fig. 8.3 **Changes of DNA methylation patterns during tumorigenesis**. Compared to normal cells (**left**) cancer cells are hypo-methylated on the genome-wide scale (**top right**), in particular at repetitive sequences, such as transposons. In addition, imprinted and tissue-specific genes often get demethylated. Hypo-methylation causes changes in the epigenetic landscape, such as the loss of imprinting, and increases the genomic instability that characterizes cancer cells. Another common alteration in cancer cells is the hyper-methylation of CpGs within regulatory regions of tumor suppressor genes (**bottom right**). These genes are then transcriptionally silenced, so that cancer cells lack functions, such as inhibition of the cell cycle

epigenome of tumor cells shows a massive overall loss of DNA methylation, while for certain genes also hyper-methylation at CpGs is observed (Fig. 8.3).

Global DNA hypo-methylation during tumorigenesis generates chromosomal instability, reactivates transposons and causes loss of imprinting. The resulting low DNA methylation favors mitotic recombination leading to deletions and promotes chromosomal rearrangements, such as translocations. The disruption of genomic imprinting, such as the loss of imprinting of the *IGF2* gene (Sect. 3.3), is a risk factor for different types of cancer, such as colon cancer or Wilms' tumor.

Hyper-methylated promoter regions of tumor suppressor genes, such as *TP53*, *RB1* (RB transcriptional corepressor 1) and *MGMT* (O-6-methylguanine-DNA methyltransferase), can serve as biomarkers that provide significant diagnostic potential in the clinic, in particular in early-detection screenings of individuals with a high familial risk of developing cancer. Many CpGs can become methylated already early in tumorigenesis, in particular in CIMP (CpG island methylator phenotype) of colon cancer, glioma and neuroblastoma. The promoter hyper-methylation rather affects the expression of genes involved in carcinogen metabolism, cell-to-cell interactions and angiogenesis than classical tumor

suppressor genes that control the cell cycle, DNA repair and apoptosis. The profiles of CpG hyper-methylation vary with tumor types. Each type of cancer can be characterized by its specific DNA hyper-methylome; *i.e.*, these epigenetic marks are comparable to traditional genetic and cytogenetic markers. From about 200 genes that are regularly mutated in various forms of human breast and colon cancers, on average 11 carry a mutation in a single tumor type. For comparison, 100-400 CpGs close to TSS regions are found to be hyper-methylated in a given tumor; *i.e.*, epigenetics is able to provide 10-times more information than genetics.

8.3 Epigenetic Reprograming in Cancer

Cancer genes are classified into dominant oncogenes, which can be activated by gain-of-function mutations, amplifications or translocations, and recessive tumor suppressor genes, the expression of which is often lost, for example, by loss-of-function mutations or methylation of their promoter regions. An alternative classification divides them into "drivers" (*i.e.*, one of approximately 500 cancer genes), the mutation of which directly affects tumorigenesis, and "passengers", which are mutated as a side product, but do not have a functional contribution on oncogenesis. The epigenetic perspective adds a further classification of cancer genes as encoding for epigenetic modifiers, mediators and modulators. Epigenetic modifier genes encode for proteins that directly modify the epigenome through DNA methylation, histone modification or structural changes of chromatin, *i.e.*, primarily chromatin modifiers and remodelers. For example, childhood tumors are often based only on a small number of genetic mutations, but these often occur in genes that encode for chromatin modifiers (Fig. 8.1, center). Moreover, the biallelic loss of the gene *SMARCB1* (SWI/SNF-related matrix-associated actin-dependent regulators of chromatin B1, Fig. 8.1, bottom), in pediatric rhabdoid tumors as well as in lung cancer and Burkitt's lymphoma, was a first indication that the **disruption of epigenetic control can serve as a driver for cancer**.

Epigenetic mediators are targets of epigenetic modification; *i.e.,* they are downstream of epigenetic modifiers. For example, due to the overactivity of an epigenetic modifier, such as a chromatin modifier of the HAT family, genes encoding for pluripotency transcription factors, such as NANOG, SOX2, or OCT4 (Sect. 6.4, Fig. 8.4), get activated in a somatic cell. Thus, in this situation the pluripotency transcription factors are epigenetic mediators and may transform a terminally differentiated cell back to a pluripotency stage. In this way, the cell forms a cancer stem cell and initiates the tumorigenesis process. Epigenetic modulators are gene products that are located upstream of epigenetic modifiers and mediators in signal transduction pathways. Epigenetic modulators influence the activity or localization of epigenetic modifiers in order to destabilize differentiation-specific epigenetic states. They represent a bridge between environment and epigenome. Inflammatory responses mediated by the transcription factor nuclear factor κB (NF-κB) are an example of an epigenetic modulator. They trigger an epigenetic switch to a positive feedback loop with the cytokine interleukin (IL) 6 and the transcription factor

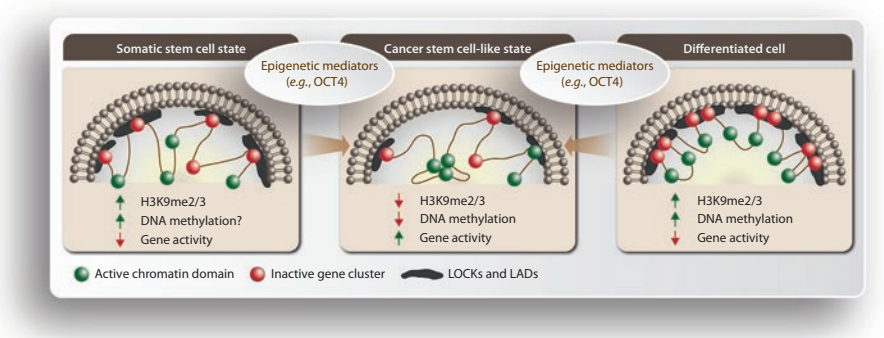


Fig. 8.4 **Reprograming of the nuclear architecture in cancer cells**. Epigenetic mediators, such as OCT4, can reprogram the epigenome of somatic stem cells (**left**) or differentiated cells (**right**) into cancer stem cells (**center**). Normal cells are characterized by high levels of H3K9 di- and tri-methylation as well as DNA methylation in LOCKs being a part of LADs. The latter are located close to the nuclear membrane and contain only a low number of active genes. In contrast, in cancer stem cells LOCKs and LADs are largely absent, and a larger variety of genes are active. This leads to phenotypic heterogeneity

STAT3 (signal transducer and activator of transcription 3) in the transformation of mammary epithelia. Thus, **the actions of epigenetic modulators are often the first steps in tumorigenesis resulting in changing epigenome patterns**.

Changes in DNA methylation during tumorigenesis are always combined with other epigenetic dys-regulations, such as aberrant patterns of histone modifications and overall changes in the nuclear architecture; *i.e.*, the overall epigenetic landscape of cancer cells is significantly distorted compared to somatic stem cells or differentiated cells (Fig. 8.4). These alterations in the 3D organization of chromatin exemplify epigenome changes during tumorigenesis. In differentiated cells developmentally repressed genes, *i.e.*, genes that are not needed in a given cell type, are often found within LADs that constitutively localize close to the nuclear periphery (Sect. 2.4). A significant fraction of these LADs represent so-called large organized chromatin K9-modifications (LOCKs), which are genomic regions that are enriched in repressive H3K9me2 and H3K9me3 histone marks (Fig. 8.4, right). This is further promoted by the recruitment of KDMs and HDACs to the repressive environment of the nuclear envelope and DNA hyper-methylation at these regions. Thus, 3D chromatin compaction mediates gene repression during lineage specification and represents a form of **epigenetic memory**. This results in reduced transcriptional noise and provides barriers for dedifferentiation. Although adult stem cells are in a less differentiated state than terminally differentiated cells, also in them specific LAD/LOCK structures are found (Fig. 8.4, left).

Proteins that regulate the interaction of chromatin with the lamina and recruit chromatin modifiers to the nuclear periphery act as epigenetic mediators. For example, the reactivated pluripotency transcription factor OCT4 can reprogram the epigenome of both differentiated cells and adult stem cells into cancer stem cells (Fig. 8.4, center). The activation of the epigenetic mediator dissolves most of the

LADs/LOCK structures, as a consequence of which a number of genes are reactivated. This provides cancer cells with phenotypic heterogeneity, such as increased variability in gene expression, in order to switch between different cellular states within the tumor. The loss of LOCKs also affects enhancer-promoter region communication within and between TADs, so that oncogenic super-enhancers are able to cluster. A similar process happens during epithelial-to-mesenchymal transition (EMT), which is a key process in normal wound healing, but also the first step toward metastasis. In EMT the activation of an H3K9-KDM, such as KDM1A, often is the initiating epigenetic event. When cancer cells have destabilized the epigenetic memory of the cells they originate from and form EMT-related chromatin structures, they gain phenotypic plasticity. Thus, the overall result of **the change in chromatin architecture might be an oncogenic transformation of the cell**.

The reprograming of a somatic cell to an iPS cell or its transformation to a cancer cell are related events (Sect. 6.4). In both cases an epigenetic barrier has to be overcome in a multi-step process that primarily involves epigenetic mediators, such as the transcription factors SOX2, KLF4, NANOG, OCT4, and MYC. Interestingly, all five transcription factors are encoded by oncogenes. Moreover, also the KMTs SUV39H1, EHMT2, SETDB1, KMT2A, KMT2D, KMT2C, DOT1L, EZH2, the KDMs LSD1, KDM2B and KDM6A, the SWI/SNF complex member ARID1A as well as DNMTA and DNMT3B have comparable roles both in cellular reprograming and in tumorigenesis. Both processes acquire *de novo* developmental programs and create cells with an unlimited self-renewal potential.

8.4 Epigenetic Mechanisms of Cancer

The idea that **cancer is fundamentally an epigenetic disease** is also reflected in the relationship between cancer and the epigenetic landscape. Since epigenetic modifiers, such as genes encoding for chromatin modifiers, are highly mutated in cancer, these mutations largely affect the stability of the epigenetic landscape. This is illustrated by permissive chromatin, which has a high rate of plasticity (Sect. 6.3, Fig. 6.5). Permissive chromatin allows cancer cells to acquire easily a number of different transcriptional states, for example, shifting to alternative developmental programs, some of which can be pro-oncogenic. When such an adaptive chromatin state propagates through mitosis, a new cell clone is created that may overgrow other cells due to increased fitness. This plasticity model can be considered as the epigenetic counterpart to the genetic model of genome instability being induced by carcinogen exposure or DNA repair defects. While permissive chromatin states allow oncogene activation or non-physiologic cell fate transitions, restrictive states prevent the induction of tumor suppressors or block differentiation.

For example, the cancer hallmark “evasion of growth suppressors” can be based on either a loss-of-function mutation of the tumor suppressor gene *CDKN2A* or by hyper-methylation of its promoter. The relative contribution of genetic and epigenetic mechanisms to the hallmarks of cancer differs between cancer types. Interestingly, the example of the adult brain tumor glioblastoma in comparison to

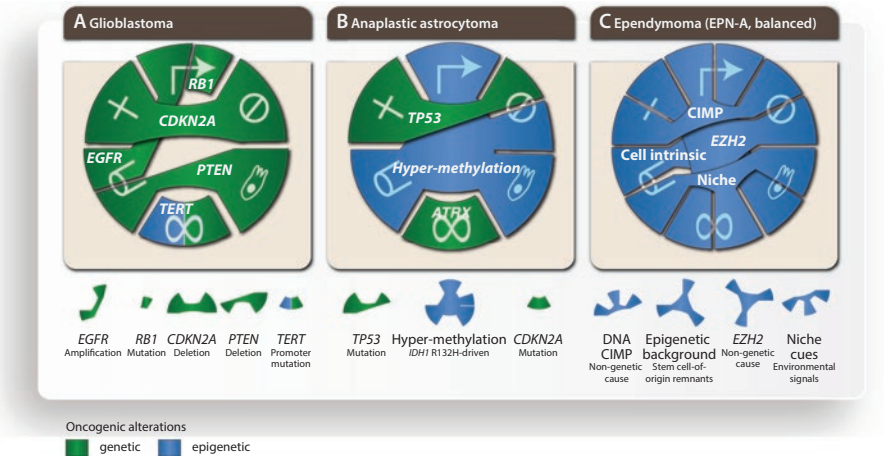


Fig. 8.5 **Genetic and epigenetic mechanisms underlying the hallmarks of cancer**. Both genetic (green) and epigenetic (blue) mechanisms are important factors in tumorigenesis, but their relative contribution to the hallmarks of cancer depends on the type of cancer. In glioblastoma (**left**), an adult brain tumor, most hallmarks relate to genetic drivers, while in ependymoma (**right**), a childhood tumor, primarily epigenetic effects dominate. Anaplastic astrocytoma (**center**) represents an example where both genetic and epigenetic factors contribute to the hallmarks

the childhood brain tumor ependymoma suggests that long-term tumorigenesis in adults may rather be based on genetic events, while short-term tumorigenesis has majorly an epigenetic origin (Fig. 8.5).

8.5 Epigenetic Cancer Therapy

The increasing appreciation of the contribution of altered epigenetic states to the phenotype of cancer cells suggests that **epigenetic cancer therapies** may have a clinical impact. This requires profound understanding of how epigenetic lesions drive cancers; *i.e.*, there is a need for conceptual and mechanistic models of cancer epigenetics in context with genetic models. Moreover, the application of new methods, such as epigenome-wide single-cell assays (Box 2.2), combined with the selection of most appropriate human cohorts will be essential. The main arguments for epigenetic cancer therapy are that genes encoding for epigenetic modifiers, such as KMTs and KDMs (Fig. 8.6), are frequent drivers in a larger range of cancer types and that epigenetic modifications, in contrast to genetic mutations, are largely reversible. So far, basically all molecules designed for epigenetic cancer therapy are inhibitors of enzymes affected by gain-of-function mutations, while targeting loss-of-function mutations remains difficult. Since histone methylation marks have a far more selective function than histone acetylation marks, KMT and KDM inhibitors promise to be more specific and may be less toxic than HDAC inhibitors or even DNMT inhibitors.

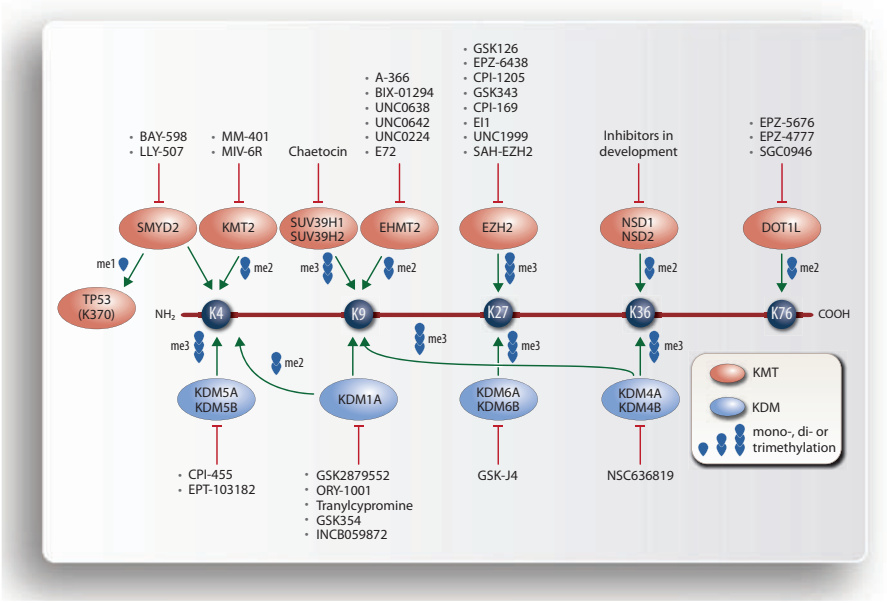


Fig. 8.6 **Targeting KMT and KDM mutations**. Inhibitors of KMTs (red) and KDMs (blue) are indicated that affect histone 3 lysines K4, K9, K27, K36, and K79 (blue). KMT inhibitors bind either within the SAM pocket, within the substrate pocket or at allosteric sites of the KMT protein. Respective inhibitors of DOT1L and EZH2 are already in clinical trials. Chaetocin is a non-selective inhibitor of KMTs, such as SUV39H1 and SUV39H2. KDM1A is a FAD-dependent KDM, which can be inhibited by molecules blocking its co-factor binding site. The KDM1A inhibitors tranylcypromine, GSK2879552, INCB059872, and ORY-1001 are in clinical trials. Most KDMs carry a catalytic domain, which can be inhibited by iron-chelating molecules

Several KMT inhibitors have been developed (Fig. 8.6), and those of the H3K27-KMT EZH2 and the H3K79-specific KMT DOT1L are already in clinical trials. Since EZH2 is the catalytic core of the PRC2 complex that also recruits DNMTs, EZH2 inhibitors may link both epigenetic repression mechanisms. The inhibition of EZH2 results in reduced levels of H3K27me3 marks, upregulation of silenced genes and inhibition of the growth of cancer cells with EZH2 gain-of-function mutations or overexpression. **KDMs use FAD, α-ketoglutarate or Fe(II) as co-factors and offer in this way a number of options for their inhibition** (Sect. 5.2). However, the catalytic domain of most KDMs is structurally highly conserved, which is a challenge for the design of specific KDM inhibitors (Fig. 8.6). Therefore, so far only inhibitors of the FAD-dependent H3K9-specific KDM LSD1, such as GSK2879552, tranylcypromine, INCB059872 and ORY-1001, are in clinical trials.

HDAC inhibitors reactivate the transcription of tumor suppressor genes, such as *CDKN1A* (encoding for the cyclin-dependent kinase inhibitor p21), by increasing histone acetylation, but they also mediate the acetylation of non-histone proteins, such as p53, and stabilize their activity. In this way, they have a wide impact on cancer cells and can induce apoptosis, cell cycle arrest and many other anti-cancer

actions. Three HDAC inhibitors, vorinostat (SAHA), belinostat and romidepsin, are approved for treatment of different types of leukemia and more than 10 others are in clinical trials for both blood and solid tumors. However, the selectivity and detailed mechanism of action of HDAC inhibitors is still not fully understood. Interestingly, HDAC inhibitors are also considered for the therapy of neuronal diseases (Sect. 9.4).

DNMT inhibitors are either nucleoside analogs that after incorporation into the DNA covalently trap DNMTs or non-nucleoside analogs that directly bind to the catalytic region of DNMTs. These molecules prevent DNA methylation leading to reduced promoter hyper-methylation and re-expression of silenced tumor suppressor genes. The two nucleoside analogs azacitidine (5-azacytidine) and decitabine (Sect. 3.1) have already been approved. Many other DNMT inhibitors are in development. Non-nucleoside analogs are less toxic, since they do not get incorporated into DNA, but they lack potency and specificity.

Finally, modulators of chromatin modifiers are not only effective in the therapy of different forms of cancer, but also in the prevention of tumor formation. Interestingly, a number of natural, food-derived compounds, such as epicatechin from green tea, resveratrol from grapes or curcumin from curcuma, are known to modulate the activity of chromatin modifiers (Sect. 10.1, Fig. 10.3). Thus, **ingredients of healthy diet have the potential to prevent cancer by keeping the activity of chromatin modifiers under control**.

Further Reading

Berdasco M, Esteller M (2019) Clinical epigenetics: seizing opportunities for translation. Nat Rev Genet 20:109–127

Feinberg AP, Koldobskiy MA, Gondor A (2016) Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. Nat Rev Genet 17:284–299

Flavahan WA, Gaskell E, Bernstein BE (2017) Epigenetic plasticity and the hallmarks of cancer. Science 357:eaal2380

Hitchins MP (2015) Constitutional epimutation as a mechanism for cancer causality and heritability? Nat Rev Cancer 15:625–634

Pfister SX, Ashworth A (2017) Marked for death: targeting epigenetic changes in cancer. Nat Rev Drug Discov 16:241–263

Shen H, Laird PW (2013) Interplay between the cancer genome and epigenome. Cell 153:38–55

Suva ML, Riggi N, Bernstein BE (2013) Epigenetic reprograming in cancer. Science 339:1567–1570

Neuroepigenetics

9

Abstract

Epigenetic regulation is critical for the normal development and functioning of our brain. **Dynamic DNA and histone methylation as well as their demethylation at specific gene loci play a fundamental role in learning, memory formation and behavioral plasticity**. The epigenome of neurons allows a molecular explanation for long-term memories. MECP2 is the best-characterized methyl-binding transcription factor and is involved both in gene activation and repression in the central nervous system. MECP2 is highly expressed in the brain and an important component of neuronal chromatin. **Mutations in the *MECP2* gene are the basis of the autism spectrum disorder Rett syndrome**. In addition, also histone acetylation levels in neurons contribute to the cell's proper function. Accordingly, HDAC inhibitors offer an effective therapy of some neurodegenerative diseases.

Keywords

Neuronal development · DNA methylation · mCH methylation · MECP2 · Rett syndrome · Histone acetylation · HDAC inhibitors · REST · Neurodegenerative diseases

9.1 Neuronal Development: The Role of Epigenetics

The frontal cortex region of our brain plays a key role in behavior and cognition and requires a coordinated interaction of neuronal and non-neuronal cells, such as supporting glial cells. The highly controlled process of neuronal development and maturation creates the physical structure of our brain. It starts during embryogenesis and continues until the third decade of our life. In parallel, after birth there is first a burst of synaptogenesis and then a pruning of unused synapses during adolescence. This is the cellular basis for experience-dependent plasticity and learning in children

C. Carlberg, F. Molnár, *Human Epigenetics: How Science Works*,
https://doi.org/10.1007/978-3-030-22907-8_9

and young adults. Its disruption can lead to behavioral alterations and neuropsychiatric disorders. Like other developmental programs of our body (Sect. 6.4), also neuronal development is controlled by precise epigenetic patterns, such as DNA methylation and histone modifications. Thus, in principle **neuroepigenetics is based on the same mechanisms as all other of our tissues and cell types**.

In general, epigenome-wide studies focus on 5mC marks at CpGs (mCG, red in Fig. 9.1a). However, like ES cells, neurons have the special property to carry significant levels of mCH marks (H = A, C or T, Sect. 3.1), some 70% of which are mCA marks. While mCH marks are hardly detectable in fetal brain, they increase during early post-natal development to a maximum of 1.5% of all CH dinucleotides at the end of adolescence (blue in Fig. 9.1a). mCH levels rise most rapidly during the primary phase of synaptogenesis, *i.e.*, within the first 2 years after birth, and correlate with the increase in synapse density (green in Fig. 9.1a). The widespread methylome reconfiguration, such as mCH, occurs in neurons, but not in glial cells, during fetal to young adult development and becomes the dominant form of methylation in our neuronal epigenome. Thus, **genome-wide DNA methylation patterns significantly change during brain development and maturation; they are the basis for neuronal plasticity**.

The frontal cortex develops post-natally in response to various inputs from the environment. Dependent on sensory input this leads to neuronal-specific DNA methylation and causes changes in gene expression and synaptic development. The chromatin modifier DNMT3A sets these mCH marks in particular at CA dinucleotides. Interestingly, although the average methylation percentage at CH is

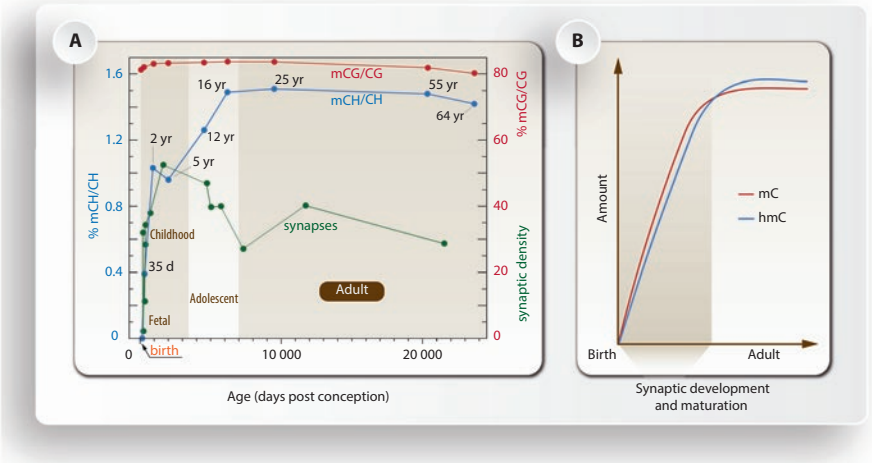


Fig. 9.1 Changes of the neuronal methylome during brain development and maturation. The levels of mCH (**a**) and 5hmC (**b**) accumulate in neurons of the human frontal cortex after birth, which coincides with active synapse development (indicated as synaptic density per 100 mm^3) and maturation (**a**). Please note that mCH marks cause gene silencing, while 5hmC marks are found at active genes. (Data are based on Lister et al. (*Science* 2013, 341:1237905))

clearly lower than at CG, but CpGs are more rare, in the adult brain mCH marks are by number even more abundant than mCG marks. Thus, **in the maturing brain mCA is a major epigenetic mark repressing gene expression.**

In general, mCA marks serve as a docking platform for methyl-binding proteins, for example, at the genomic region of the gene *BDNF* (brain-derived neurotrophic factor). The methyl-CpG-binding transcription factor MECP2 acts as a reader of DNA methylation marks (Fig. 3.1) and also recognizes with high affinity mCA marks. Therefore, genes that acquire mCA enrichment during neurodevelopment recruit MECP2 and are repressed in their transcriptional initiation and elongation. This regulatory process mediates distinct functions in the brains of adult versus that of newborns. Accordingly, the number of mCA marks at gene bodies is more predictive for the level of gene silencing than that of mCG marks at promoters or any measure of chromatin accessibility. Moreover, the impact of mCH marks is further highlighted by the observation that **between individuals mCH marks are more conserved than mCG marks**.

Interestingly, different regions of the brain, such as frontal cortex, hippocampus, and cerebellum, show significant age-dependent increase of 5hmC, *i.e.*, of the oxidized forms of 5mC (Fig. 9.1b). This increase is specific to neurons, which in adults have far higher 5hmC levels than any other human tissue of cell type. TET enzymes mediate the 5mC oxidation and active DNA demethylation in neurons. Based on mouse knockout studies, TET1 is most important for neurogenesis and synaptic plasticity. Thus, **5mC oxidation and possibly DNA demethylation are central epigenetic mechanisms of brain development and maturation**.

Like mCA also 5hmC marks recruit MECP2 and other methyl-binding proteins, but in this case the abundance of these transcription factors mostly positively correlates with gene expression. This fits with the observation that the *MECP2* mutation R133C, which occurs in some forms of the autistic spectrum disorder Rett syndrome (Sect. 9.2), specifically disrupts the ability of the protein to bind to 5hmC and results in target gene silencing. Thus, enrichment of 5hmC and depletion of 5mC both increase transcriptional activity and chromatin accessibility (Fig. 9.2). This also suggests that 5mC oxidation may be the key epigenetic mechanism in controlling gene expression in the context of synaptic plasticity, which is important for learning and memory. The latter involves effects on long-term potentiation, excitability and activity-dependent synaptic scaling of neurons. Thus, *via* controlling the activity of genes encoding for ion channels, receptors and trafficking mechanisms, the epigenome has the capacity to both sense extra-cellular signals reaching neurons and to control their output.

Taken together, during brain development neurons acquire epigenetic marks, such as 5mC (at CG and CH) and 5hmC, that stabilize the neuronal phenotype over our lifespan. **Although neurodevelopmental disorders are generally thought to be irreversible, the plastic nature of the dynamic part of the neural epigenome *via* 5hmC and demethylation has major implications for a possible epigenetic reprograming in the context of these diseases**.

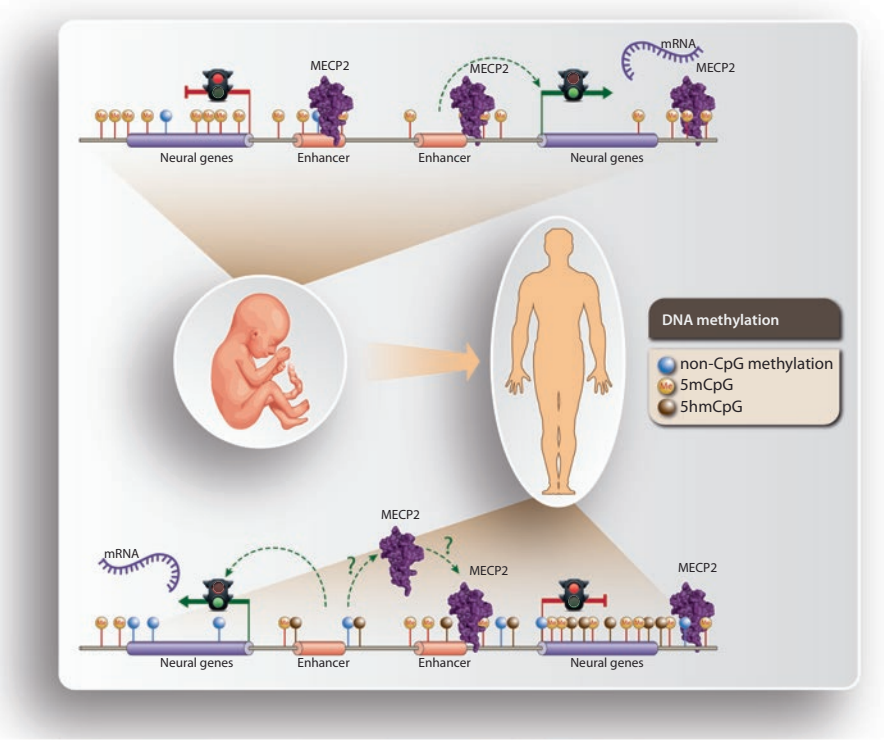


Fig. 9.2 **The neuronal methylome during brain development.** 5hmC marks label enhancers in the fetal brain (**top**) that will become demethylated and active in the adult brain (**bottom**). The DNA-binding pattern of methyl-DNA-binding proteins, which have different affinity for 5hmC and 5mC, can change. This affects the neuronal transcriptome

9.2 Epigenetic Basis of Memory

The acquisition of new information is defined as learning, while the ability to retain information for later reconstruction is memory. Memories provide the basis for our behavior, since they allow us to reliably navigate throughout our life and to interact with our (social) environment. For the consolidation of new information into memory, distinct gene expression profiles are activated in our neurons, which are based on changes of our epigenome. In contrast, impairments in learning and memory can have devastating consequences for an individual's ability to function independently in society. In addition, the persistence of undesirable memories, such as those of trauma (for example, war veterans) or violence (for example, child abuse), can contribute to mental illness leading to social inhibition. Thus, **(social) intelligence is based on our neuronal epigenome**.

Memories of our childhood start from an age of approximately 3 years; *i.e.*, they last for many decades. DNA is the only molecule in our body that has a comparable half-life; *i.e.*, only our genome can be the molecular basis for a long-term memory. Since memories do not change the sequences of our genome, our neuronal epigenome remains as the exclusive device for information storage. In previous chapters we used multiple times the term "epigenetic memory" (Sects. 1.3, 2.1, 3.1, 5.3, 7.1, and 8.3) and often meant the long-term stability of cell identities *via* preserved gene expression states and robust gene regulatory networks. For example, the actions of chromatin modifiers and remodelers within Polycomb and Trithorax complexes (Sect. 6.3) result in hundreds of genomic regions in long-term, mitotically heritable memory of silent and active gene expression states. Key proteins of these complexes are KMTs and KDMs, such as EZH2, KMT2A, KDM5C and KDM6B, while HATs and HDACs are missing. This reiterates the principle that **short-term "day-to-day" responses of the epigenome are primarily mediated by non-inherited changes in the histone acetylation level, while long-term decisions, for example, concerning cellular differentiation, are stored in form of histone methylation marks**.

DNA methylation not only silences gene expression (Sect. 1.1) but in addition also serves as an essential mediator of memory acquisition and storage. Neurons do not divide and get as old as we get; *i.e.*, their epigenome is the most likely location for our long-term memory. The knockout of the genes *Dnmt1* and/or *Dnmt3a* in mice demonstrated a loss of long-term potentiation and consecutively deficits in learning and memory. This indicates that **without active DNA methylation, information storage processes do not work**. Moreover, rather than being a first responder to extra-cellular signals, DNA methylation acts as a consolidator of previously established memory acquisition via histone acetylation and methylation.

Taken together, each of the approximately 100 billion (10^{11}) neurons within our brain has a tremendous data storage capability, since each of it contains some 700 million cytosine that may be methylated and some 30 million nucleosomes comprised of 240 million histone proteins that can be post-translationally modified in more than 100 individual ways.

9.3 MECP2 and the Rett Syndrome

MECP2 belongs to the family of intrinsically disordered proteins, which are characterized by a low level of secondary structure. This makes MECP2 well suited to interact with a large variety of different types of macromolecules, such as other proteins, DNA and RNA. Central to the primary structure the MECP2 protein is its DNA-binding domain that includes a MBD, but also contains several other structures, such as AT hook motifs. Accordingly, a specific target of MECP2 is methylated DNA, in particular CA dinucleotides, but MECP2 also interacts with a wide range of other proteins, such as heterochromatin protein HP1, the co-repressor NCOR1 and the HAT CREBBP.

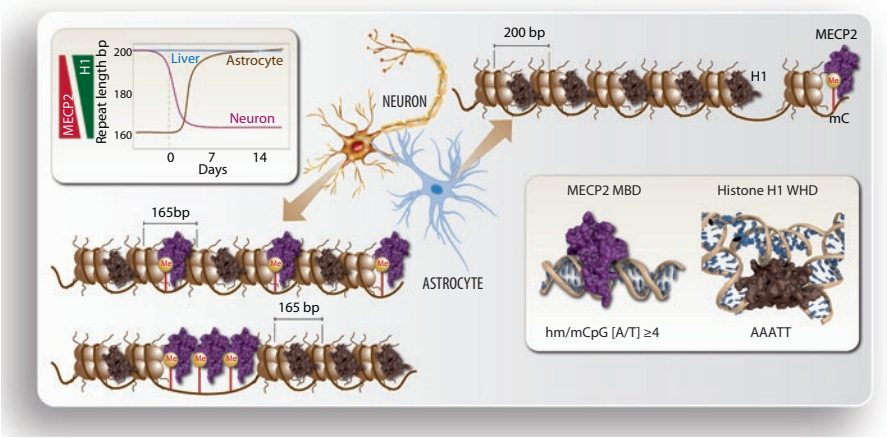


Fig. 9.3 MECP2 binding to chromatin of neurons and astrocytes. Astrocytes have lower MECP2 levels than neurons resulting in a regular chromatin repeat length of 200 bp compared to 165 bp for neurons. In neurons, MECP2 is evenly distributed throughout the chromatin, binds to sites of methylated DNA and replaces the linker histone H1, which decreases the repeat length. Changes of chromatin repeat length during development (in days, observed in a mouse model) before and after birth are indicated for astrocytes, neurons and as a reference liver tissue. A higher proportion of MECP2 in relation to histone H1 results in a shorter repeat length (**insert top left**)

MECP2 gets rapidly ubiquitinated, which limits its half-life to only 4 h. The protein is ubiquitously expressed in basically all human tissues and cell types, but it shows highest expression in the brain, in particular in neurons. Accordingly, in neuronal chromatin MECP2 is a very abundant protein appearing in average every second nucleosome (Fig. 9.3). MECP2 and the linker histone H1 (Sect. 2.1) compete for binding to the nucleosome. When MECP2 levels rise during the neuronal development, the protein replaces histone H1. This causes the reduction of the chromatin repeat length, *i.e.*, the distance from the center of a nucleosome to the center of its neighbor, from 200 bp to 165 bp (Fig. 9.3). Thus, **MECP2 increases the density of chromatin packaging**. In this configuration, MECP2 is part of tightly folded heterochromatin structures suggesting that the protein would function mainly as a transcriptional repressor. Nevertheless, there are a number of MECP2 interaction partners that stimulate gene expression. This indicates that MECP2 rather is a transcriptional regulator that, depending on its interaction partners, binds to genomic regions of activated or repressed genes. Furthermore, MECP2 has opposite roles when binding to promoter regions or gene bodies. DNA methylation at TSS regions recruits MECP2 in complex with co-repressor proteins and HDACs and results in transcriptional repression. In contrast, the DNA of gene bodies of transcribed genes is methylated and recruits MECP2 binding, which in turn prevents the binding of the repressive histone variant H2A.Z (Box 4.1). Thus, **the genomic location of MECP2 binding is of critical importance of its function**.

Chromatin modifiers play central roles in cognitive disorders (Table 9.1). At least seven proteins are known to be mutated in X chromosome-linked intellectual

Table 9.1 Epigenetic mechanisms of neuropathological disorders. Neuronal disorders and functions that are affected by epigenetics are listed, such as mutations or overexpression of MECP2, aberrant DNA methylation, and/or histone modifications

Function or disorder	Mechanism(s) implicated
Rett syndrome	*MECP2* mutations
Autism spectrum disorders	MECP2 overexpression/increased dosage, aberrant DNA methylation
Alzheimer's disease	MEeCP2 decrease, histone modifications, aberrant DNA methylation
Parkinson's disease	Loss of MECP2
Huntington's disease	MECP2 dys-regulation
Fragile X syndrome	Aberrant DNA methylation
Rubinstein-Taybi syndrome	HAT deficiency
Friedreich's ataxia	Reduced histone acetylation
Angelman syndrome	Genomic imprinting (DNA methylation)
Addiction and reward behavior	MECP2 decrease, histone modifications, DNA methylation, miRNAs
Post-traumatic stress disorder	Histone modifications, DNA methylation
Depression and/or suicide	DNA methylation
Schizophrenia	Increased MECP2 binding, histone methylation, aberrant DNA methylation
Epilepsy	MECP2 up-regulation, histone modifications, aberrant DNA methylation

disabilities. These proteins, such as MECP2, are either methyl-binding proteins or methyl-modifying enzymes. The disruption of the *MECP2* gene leads to a special form of autism, referred to as Rett syndrome. More than 95% of all cases of the Rett syndrome can be explained by mutations within the *MECP2* gene; *i.e.*, this special form of autism is largely a monogenetic disease. This means that **the mechanistic understanding of the Rett syndrome is based on dys-functional MECP2 proteins in neurons**. In addition, also functional alternations of MECP2 in astrocytes and in microglia affect the disease phenotype, although MECP2 is much lower expressed in these cell types.

Individuals with Rett syndrome are heterozygous for a mutation in the *MECP2* allele. Since the *MECP2* gene is located on the X chromosome, the disease is mostly embryonal lethal for males and almost exclusively females are affected with an incidence of approximately 1 in 10,000 persons. The syndrome mostly occurs through *de novo* mutations in the paternal germline. Affected females have an apparent normal early post-natal development, which may be due to the fact that during brain maturation mCH levels gradually increase (Sect. 9.1). However, between the age of 6–18 months the syndrome develops, such as that the individuals lose the ability to retain communication function and motor skills. Moreover, physical growth becomes retarded and microcephaly establishes. Thus, Rett syndrome is the **first human neuronal disease, for which a significant impact of epigenetics could be demonstrated**, and it serves as a master example for the impact of neuroepigenetics on disease.

9.4 Epigenetics of Neurodegenerative Diseases

The example of the monogenetic neurodevelopmental disease Rett syndrome (Sect. 9.3) leads to the question, whether epigenetics plays also a role in complex multigenic disorders of the nervous system, such as neurodegenerative diseases. MECP2 is involved in controlling the secretion of the neurotransmitters GABA, dopamine and serotonin from respective neurons and affects the number of synapses of glutamatergic neurons. This is, at least in part, mediated by the interaction of MECP2 with BDNF, which acts as a modulator on glutamatergic and GABAergic synapses. Thus, the tightly regulated expression of MECP2 is critical for neuronal homeostasis. Since both too high as well as too low MECP2 protein levels trigger opposite effects in synaptic transmission, dys-regulation of MECP2 protein expression contributes to many neuro-pathological disorders, such as Alzheimer's disease, Huntington's diseases, schizophrenia and epilepsy (Table 9.1). In parallel, aberrant DNA methylation has been observed in some autistic spectrum disorders, Alzheimer's disease, epilepsy and schizophrenia. In general, **epigenetic mechanisms may be particularly relevant to complex diseases with low genetic penetrance that use epigenetic mechanisms of development and learned behavior**, such as drug addiction, post-traumatic stress disorder, epilepsy and schizophrenia.

Histone acetylation is the best-understood epigenetic modification. Several neurodegenerative diseases involve disruptions in the HAT/HDAC balance; *i.e.*, patients with these disorders have abnormal histone acetylation levels (Fig. 9.4). A master example is the Rubinstein-Taybi syndrome, which is characterized by short stature, mental retardation, moderate to severe learning difficulties, distinctive facial features as well as broad thumbs and big toes. Rubinstein-Taybi syndrome is a monogenetic disease that is based on mutations of the *CREBBP* gene, which encodes for a HAT. The resulting low histone acetylation levels may be counterbalanced by the inhibition of HDACs; *i.e.*, HDAC inhibitors are a therapeutic option for patients with the syndrome.

Another example of a monogenetic neurodegenerative disease is Friedreich's ataxia, which results from the degeneration of nervous tissue in the spinal cord, in particular in sensory neurons that are essential for directing muscle movement of arms and legs. In this disorder the expansion of a triplet repeat region within an intron of the gene *FXN* (frataxin) leads to the loss of H3ac and H4ac marks and gain of H3K9me3 marks, heterochromatin formation and finally transcriptional silencing of the gene (Fig. 9.4). In a mouse model of this disease, HDAC inhibitors increased H3 and H4 acetylation and corrected the *FXN* expression deficiency. This suggests that **HDAC inhibitors may be suitable for the treatment of Friedreich's ataxia**.

Similarly, mental retardation associated with the fragile X syndrome is caused by a CGG-triplet repeat expansion in the 5′-UTR of the gene *FMR1* (fragile X mental retardation 1), which leads to extensive DNA methylation at CpGs close to the gene's TSS and gene silencing (Fig. 9.4). Also in this case a treatment with HDAC inhibitors resulted in reactivation of *FMR1* expression. In particular SIRT1 inhibitors were able to increase acetylation and decrease methylation of histones at the *FMR1* gene locus. Finally, Huntington's disease is based on polyglutamine repeats in the

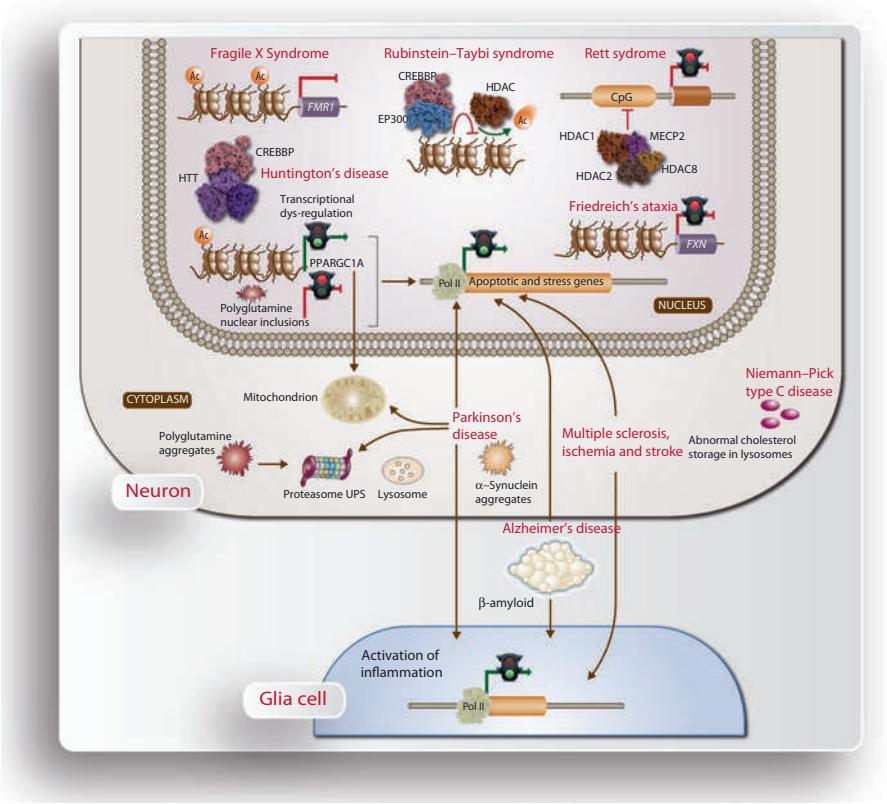


Fig. 9.4 **Role of histone acetylation in neurodegenerative diseases.** The level of histone acetylation depends on the balance of HATs and HDACs and is related to several neurodegenerative diseases. For example, decreased acetylation activity of the HAT CREBBP is associated with the Rubinstein-Taybi syndrome and polyglutamine diseases, such as Huntington's disease. The transcriptional silencing of the *FXN* gene in Friedreich's ataxia or of the *FMR1* gene in fragile X syndrome can be released by HDAC inhibition. Other examples of neuronal disorders are schematically depicted and their possible treatment by HDAC inhibitors is indicated

5′-coding region of the gene *HTT* (huntingtin) causing progressive motor and cognitive decline. The disease involves perturbations in many aspects of neuronal homeostasis, such as abnormal histone acetylation and chromatin remodeling as well as aberrant interactions of the HTT protein, for example affecting the function of the gene PPARGC1A (peroxisome proliferator-activated receptor gamma, co-activator 1 α) (Fig. 9.4). In different **animal models of Huntington's disease HDAC inhibitors showed a neuroprotective effect**.

Targeting histone acetylation *via* the application of HDAC inhibitors had already been introduced in the context of cancer therapy (Sect. 8.5). However, these compounds may also provide benefit for the treatment of complex neuronal diseases, such as Alzheimer's disease and Parkinson's disease as well as depression, schizophrenia, drug addiction and anxiety disorders. For example, in mouse models

the HDAC inhibitor sodium butyrate showed anti-depressant effects. Moreover, HDAC inhibitor-treated animals showed induced sprouting of dendrites, an increased number of synapses, re-established learning behavior and access to long-term memories. This suggests **a possible wide application for HDAC inhibitors in the therapy of cognitive disorders**. For example, valproic acid is approved as mood stabilizer and anti-epileptic.

HDAC inhibitors also widely affect gene expression in the immune system (Sect. 12.3) and showed to be effective in the treatment of inflammation and neuronal apoptosis (Fig. 9.4). **Animal models of ischemia-induced brain infarction indicated that HDAC inhibitors have anti-inflammatory and neuroprotective effects and can be used for the treatment of stroke**.

Neurodegenerative diseases have different underlying causes and pathophysiology, but they all involve impaired cognition, neuronal death and the dys-regulation of the transcription factor REST (also known as NRSF). REST is a repressing transcription factor that has an effect on both acetylation and methylation levels of histones. REST is widely expressed throughout embryogenesis, but at the end of neuronal differentiation it gets downregulated in order to acquire the neuronal phenotype. The transcription factor serves as the DNA-binding platform of a large protein complex containing the co-repressors RCOR1 (Sect. 5.4), HDAC1 and 2, MECP2, the KMT EHMT2, and the KDM LSD1 (KDM1A). In addition, the REST complex can also recruit the DNA methylation machinery. The REST complex binds RE1 (restrictive element 1) sites that locate in the vicinity of TSS regions of REST target genes. This leads to deacetylation and methylation of local nucleosomes at position H3K9 and demethylation at H3K4me2 marks. Thus, the local chromatin at TSS regions of REST target genes gets very effectively silenced. In total there may be up to 2000 primary REST target genes within our genome. However, the set of active REST targets is cell type- and context-dependent and varies with developmental and disease stage. Probably, different epigenetic landscapes in various brain regions and disease states largely influence the selection of REST for a subset of its target genes.

Since most cases of Alzheimer's disease are sporadic and develop over time, there is a significant contribution of environmental factors to the onset of this neurodegenerative disorder. While in neurons of the pre-frontal cortex and the hippocampus of the healthy aging brain REST silences genes involved in apoptosis and oxidative stress, this is lost in patients with mild or severe cognitive impairment and Alzheimer's disease. With the occurrence of mis-folded proteins that are characteristic for the onset of Alzheimer's disease, such as Aβ and tau, the rate of autophagy increases. Under these conditions autophagosomes engulf not only Aβ and tau complexes but also REST, which results in depletion of the transcription factor from the nucleus. The loss of REST results in an increase in expression of previously repressed genes being involved in oxidative stress and neuronal death. Thus, **REST deprivation increases the loss of neurons in the respective brain regions and promotes the progression of Alzheimer's disease**.

Taken together, neuroepigenetics provides alternative and/or additional mechanistic explanations for the onset of neurodegenerative diseases. Thus, **neuroepigenetics has the potential to guide the design of novel therapeutic strategies for improving the harmful consequence of cognitive deficits and neurodegeneration**.

Further Reading

Campbell RR, Wood MA (2019) How the epigenome integrates information and reshapes the synapse. Nat Rev Neurosci 20:133–147

Gräff J, Tsai LH (2013) Histone acetylation: molecular mnemonics on the chromatin. Nat Rev Neurosci 14:97–111

Hwang JY, Aromolaran KA, Zukin RS (2017) The emerging field of epigenetics in neurodegeneration and neuroprotection. Nat Rev Neurosci 18:347–361

Kazantsev AG, Thompson LM (2008) Therapeutic application of histone deacetylase inhibitors for central nervous system disorders. Nat Rev Drug Discov 7:854–868

Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD et al (2013) Global epigenomic reconfiguration during mammalian brain development. Science 341:1237905

Nutritional Epigenetics

10

Abstract

Nutritional epigenetics is a sub-discipline of nutrigenomics and describes how dietary compounds affect our epigenome. Chromatin modifiers use intermediary metabolites, such as acetyl-CoA, α-ketoglutarate, NAD^+, FAD, ATP or SAM, as co-substrates and/or co-factors. In this way, chromatin modifiers act as sensors for the nutritional status of our tissues and cell types leaving respective marks on their epigenome. The thrifty phenotype is a concept of epigenetic programing of metabolic tissues during pre-natal development. Its principles apply also in adult life and may explain the missing heritability of the susceptibility for complex metabolic diseases, such as type 2 diabetes.

Keywords

Nutrigenomics · Energy metabolism · Acetyl-CoA · NAD^+ · SAM · Folate · Thrifty phenotype · Type 2 diabetes · Personalized nutrition

10.1 Epigenetic Mechanisms of Nutrigenomics

Our diet is a complex mixture of biologically active molecules that are either micronutrients (nano- to micromolar concentration range), such as vitamin D, or macronutrients (micro- to millimolar levels), such as fatty acids and cholesterol. Some of these dietary molecules

- have a direct effect on gene expression, for example, by acting as a ligand to a nuclear receptor (Fig. 10.1a)
- modulate, after being metabolized, the activity of a transcription factor (Fig. 10.1b)
- stimulate a signal transduction cascade that ends with the induction of a transcription factor (Fig. 10.1c).

C. Carlberg, F. Molnár, *Human Epigenetics: How Science Works*,
https://doi.org/10.1007/978-3-030-22907-8_10

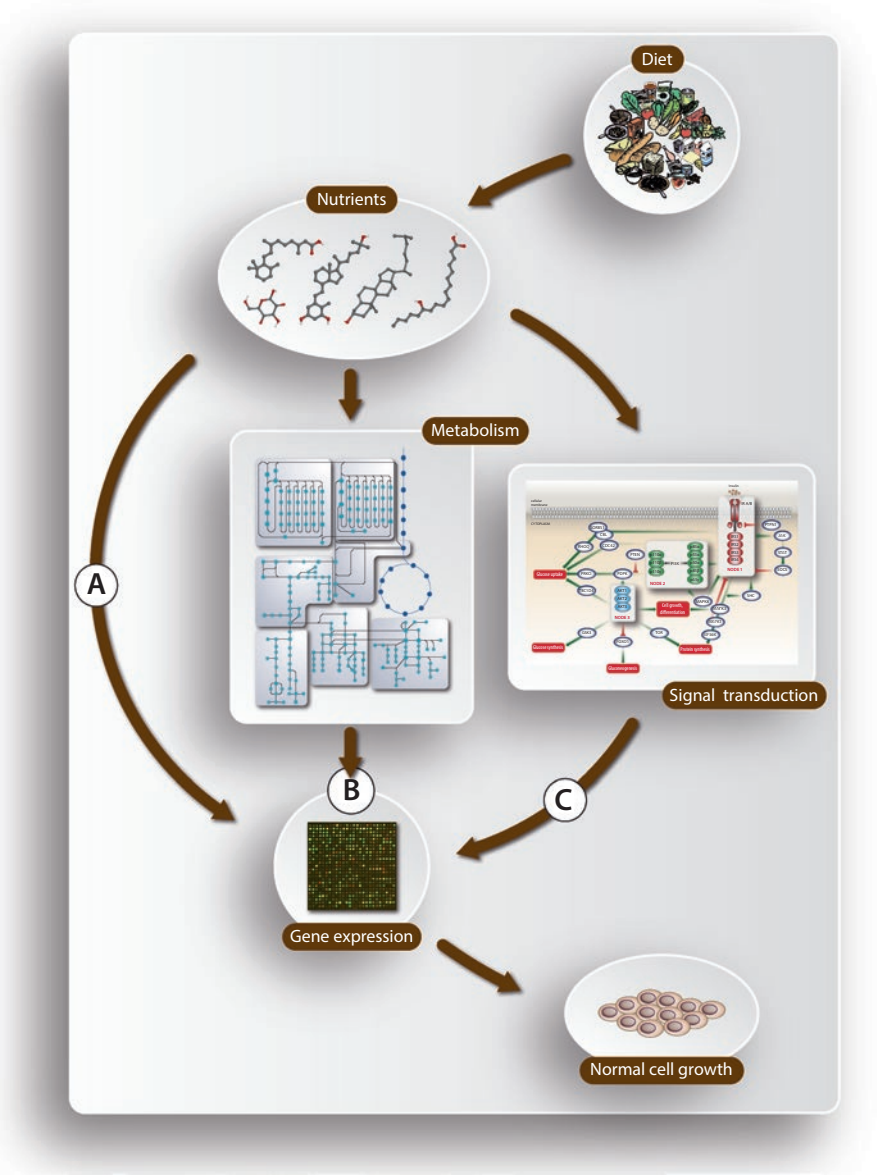


Fig. 10.1 **Basis of nutrigenomics**. Nutrigenomics aims to provide a molecular understanding for how dietary nutrients affect health by altering the expression of a larger set of genes. Nutritional compounds have been shown to alter gene expression in a number of ways. They may (**a**) act as direct ligands for transcription factors, (**b**) be transcription factor modulators after a chemical conversion in metabolic pathways or (**c**) serve as activators of signal transduction cascades that end with the activation of a transcription factor. All three activation pathways modulate physiological effects, such as cellular growth

Nutrigenomics aims to describe, characterize and integrate interactions between diet and gene expression on a genome-wide level. The results of these investigations lead to an improved understanding of how nutrition influences metabolic pathways and homeostatic control. For example, nutrigenomics investigates how nutrition-triggered gene regulation may be disturbed in the early phase of a diet-related disease, such as type 2 diabetes. When individuals are classified according to the interplay of their lifestyle, metabolic pathways and genetic variation, the molecular insight based on nutrigenomic studies can suggest tailored diets, referred to as personalized nutrition, for disease prevention and/or early therapeutic intervention (Sect. 12.2).

Nutrigenomics and pharmacogenomics have some similarities concerning concepts and methodological approaches. However, in principle, in pharmacogenomics the effects of a single clearly defined compound (a drug) of a precise concentration and a specific target can be investigated, whereas nutrigenomics faces the complexity and variability of nutrition. However, when personalized nutrition is used for a specific therapy, the distinction between food and drugs as well as the definition of health and disease dissolves.

The availability of dietary metabolites is essential for all tissues and cell types of our body. Accordingly, this important environmental input results in the activation of signal transduction pathways within cells. These pathways stimulate gene expression programs that integrate the information about our nutritional status with the goal to preserve cellular homeostasis. Interestingly, the activity of most chromatin modifiers critically depends on intra-cellular levels of essential metabolites. Thus, **there are a number of mechanisms how metabolites trigger the epigenome**. For example, α-ketoglutarate is an essential co-substrate of TET enzymes (Sect. 3.1) and Jumonji domain-containing KDMs (Sect. 5.1). Metabolites that are structurally similar to α-ketoglutarate, such as 2-hydroxyglutarate, succinate and fumarate, can block the α-ketoglutarate binding site and inhibit TETs or KDMs. Under hypoxic conditions (S)-2-hydroxyglutarate is produced by the enzyme lactate dehydrogenase, while the isomer (R)-2-hydroxyglutarate is created in cancer cells that have mutant forms of the enzymes IDH1 and IDH2 (Sect. 8.1, Fig. 8.1). In this way, rather different conditions, such as hypoxia and gene mutations, can lead to the production of a metabolite that causes hyper-methylation of both histone proteins and genomic DNA (finally leading to gene silencing). Similarly, the accumulation of succinate or fumarate in tumors that are deficient for the enzymes succinate dehydrogenase or fumarate hydratase can lead to the inhibition of α-ketoglutarate-dependent enzymes and also causes hyper-methylation. Thus, **chromatin modifiers can act as sensors of the metabolic status of a cell** and translate the metabolic information into dynamic post-translational histone modifications that coordinate adaptive transcriptional responses (Fig. 10.2).

Another example how variations in the cellular metabolic state affect the relative concentrations of co-factors and influence the activity of chromatin modifiers is the methyl donor substrate SAM (Sect. 3.1). The molecule connects the processes

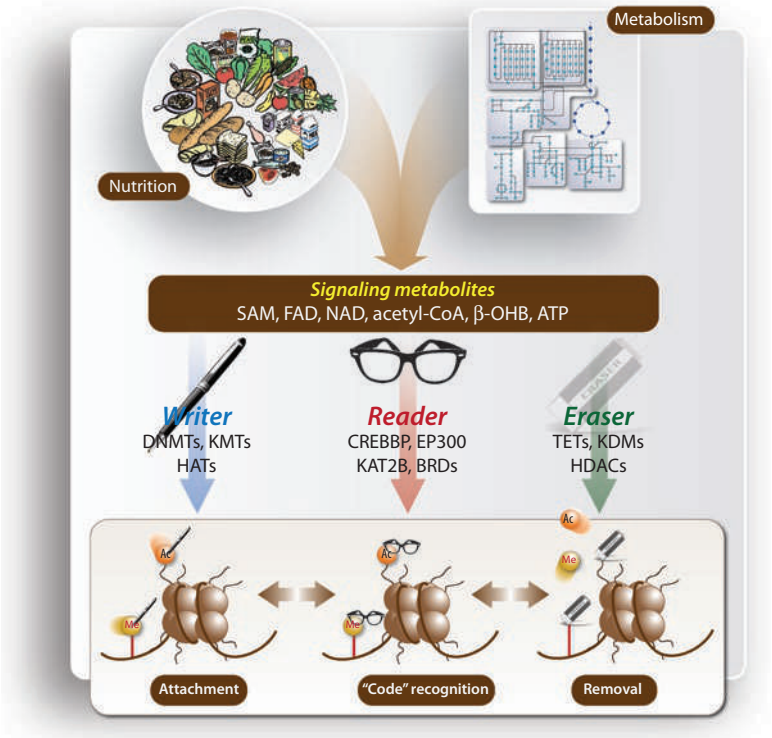


Fig. 10.2 **Nutritional epigenetics**. Changes in nutrition or fluctuations in metabolism affect the transcriptional responses of metabolic tissues. A number of intermediary metabolites are co-substrates and/or co-factors of chromatin modifiers; *i.e.*, these enzymes act as metabolic sensors. Writer enzymes create covalent chromatin marks, reader enzymes recognize these marks, and eraser enzymes remove them (Sect. 5.1). This results in changes of local chromatin structure and has consequences for the activity and regulation of the neighboring genes

intermediary metabolism and DNA methylation (Box 10.1). After the methyl group of SAM is transferred to a histone or to genomic DNA, the product S-adenosylhomocysteine (SAH) is recycled back to SAM. Interestingly, SAH is a negative feedback regulator of KMTs; *i.e.*, the SAM/SAH ratio (also referred to as the "methylation index") is critical for histone and DNA methylation.

Interestingly, a wide spectrum of secondary metabolites that we take up from fruits, vegetables, teas, spices and traditional medicinal herbs, such as genistein, resveratrol, curcumin and polyphenols from green tea, coffee and cocoa, are able to modulate the activity of chromatin modifiers (Fig. 10.3).

Taken together, a number of nutrigenomic approaches aim to maintain wellbeing, promote health and open up new therapeutic strategies, such as **a possible reprograming of the epigenome of metabolic organs through personalized**

Box 10.1 Folate Metabolism and Methylation

A derivative of the B vitamin folate, tetrahydrofolate, feeds the cyclic one-carbon pathway by serving as a methyl group donor (Sect. 7.1). This demonstrates a direct connection between nutrition and epigenetics. **Methyl group donors are critical for epigenetic programing during embryogenesis**. A high homocysteine level is an established biomarker for the disturbance of the one-carbon metabolism and related to low concentrations of folate, vitamins B6 and B12, choline and betaine. This may cause an elevated risk of premature delivery, low birth weight and neural tube defects. Moreover, a low dietary intake of folate or methionine increases the risk of colon adenomas, while *in utero* exposure to higher folate is associated with a reduced risk of childhood acute lymphoblastic leukemia, brain tumors and neuroblastoma. The enzyme methylenetetrahydrofolate reductase, which is encoded by the *MTHFR* gene, catalyzes the conversion of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate. 10–15% of Europeans carry a C677T missense SNP (rs1801133) on both alleles, which reduces the activity of the enzyme by more than 50%. Accordingly, individuals with a TT genotype are affected more by a low folate intake than those with the CC or CT allele.

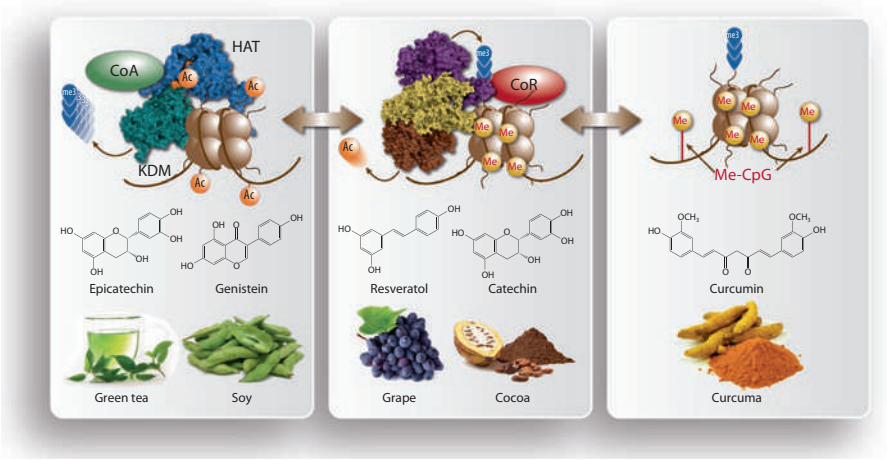


Fig. 10.3 **Natural compounds regulate the activity of chromatin modifiers**. Plant-origin natural compounds modulate the activity of chromatin modifiers, such as HATs and KDMs in open chromatin (**left**), KMTs, DNMTs, and HDACs in facultative heterochromatin (**center**) and methylated CpGs in heterochromatin (**right**). In this way they affect the epigenetic status of most human tissues and cell types

nutrition including natural compounds that modulate the activity of chromatin modifiers and transcription factors. This concept suggests that lifestyle changes, such as increased physical activity and consecutively weight loss can have a beneficial effect on the epigenome and thus lowering the risk for suffering from the metabolic syndrome (Sect. 12.2).

10.2 Energy Metabolism and Epigenetics

The energy status of our tissues and cell types is the most important information for our body in order to interpret and integrate environmental conditions. Nutritional epigenetics investigates how metabolic pathways communicate with chromatin and provide information about nutrient availability and energy status. Since key metabolites, such as AMP, NAD^+, SAM, and acetyl-CoA, act as co-factors and substrates of chromatin modifiers, gene expression programs of many central physiological processes, such as proliferation and differentiation, are modulated by the metabolic status of the cells. The results of these epigenetic events may be memorized in the epigenome of skeletal muscle and adipose tissue. The latter two metabolic organs constitute more than half of our body mass. However, their relative amount is very variable and depends on environmental factors, such as physical activity and nutritional intake. This means that our lifestyle creates a metabolic memory. Thus, not only the tissue mass but also **the epigenome of our muscles and fat memorizes how much we have eaten and moved**.

The ratio of the oxidized (NAD^+) and reduced (NADH) form of the co-factor NAD reflects the cellular redox state and is inversely proportional to the energy state of a cell. During fasting, *i.e.*, at low levels of nutritional metabolites, the intracellular concentration of NAD^+ raises. This leads to an increase in the activity of HDACs of the SIRT family (which use NAD^+ as a co-factor) and the deacetylation of their target proteins (Fig. 10.4, left). The targets are often histones, but also transcription factors or their co-factors, such as p53 and PPARGC1A, are affected in their acetylation status. Calorie restriction, *i.e.*, using only 70–80% of the recommended dietary intake, is beneficial for metabolic health and may slow down aging (Sect. 7.3). Since the NAD^+ concentrations fluctuate in a circadian manner, SIRT-mediated gene regulation is linked to the epigenetic clock (Sect. 7.4). Moreover, the metabolite D-β-hydroxybutyrate (β-OHB) acts as an inhibitor of HDACs 1–3. During fasting or calorie restriction, levels of β-OHB increase and promote histone acetylation.

In contrast, nutrients ingested in the feeding state enter the catabolic pathways of intermediary metabolism and acetyl-CoA is produced. Augmented acetyl-CoA concentrations stimulate HAT activity, so that their target proteins get acetylated (Fig. 10.4, right). When the target proteins are histones, the acetylation of chromatin leads to open chromatin. This stimulates the expression of genes involved in metabolic processes, such as lipogenesis and adipocyte differentiation. Moreover, a

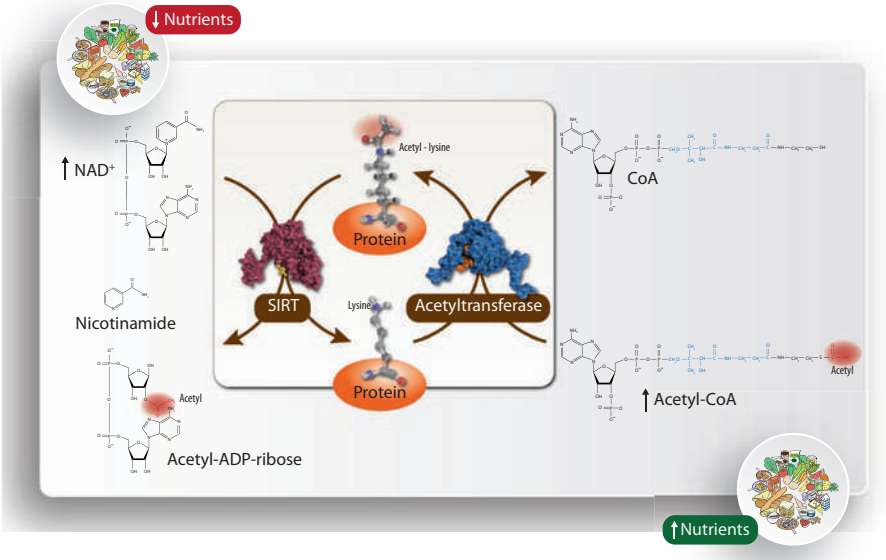


Fig. 10.4 **The relation of protein acetylation and cellular metabolism**. NAD^+ (**left**) acts as a co-factor for HDACs of the SIRT family that deacetylate proteins, which had been acetylated by HATs using acetyl-CoA (**right**). Thus, the acetylation status of key regulatory proteins reflects the cellular concentration of NAD^+ and acetyl-CoA, *i.e.*, of low (**top**) or high (**bottom**) nutritional status, respectively

large proportion of acetyl-CoA-responsive genes are involved in cell cycle progression; *i.e.*, an increase in histone acetylation is associated with cellular proliferation. However, upon induction of cellular differentiation, for example, of ES cells, the acetyl-CoA level decreases significantly. Accordingly, loss of pluripotency is associated with decreased glycolysis and lower levels of acetyl-CoA and histone deacetylation. Moreover, the acetyl-CoA level also affects cell survival and death decisions. For example, a low acetyl-CoA level induces the catabolic process of autophagy, which is crucial for organelle quality control and cell survival during metabolic stress. Thus, the acetyl-CoA/CoA ratio is an important regulator of major cellular decisions.

Another example of metabolite sensing is that of the enzyme AMPK, which is controlled in its activity by the AMP/ATP ratio. When cells consume more ATP than they are producing, *i.e.*, at conditions of low nutrient availability, AMP concentrations raise as a signal of energetic stress. AMP binds to the γ-subunit of the AMPK heterotrimer and activates the kinase. Since histones are AMPK substrates, a low energy status of the cell is marked *via* histone phosphorylation. Thus, **insults to the**

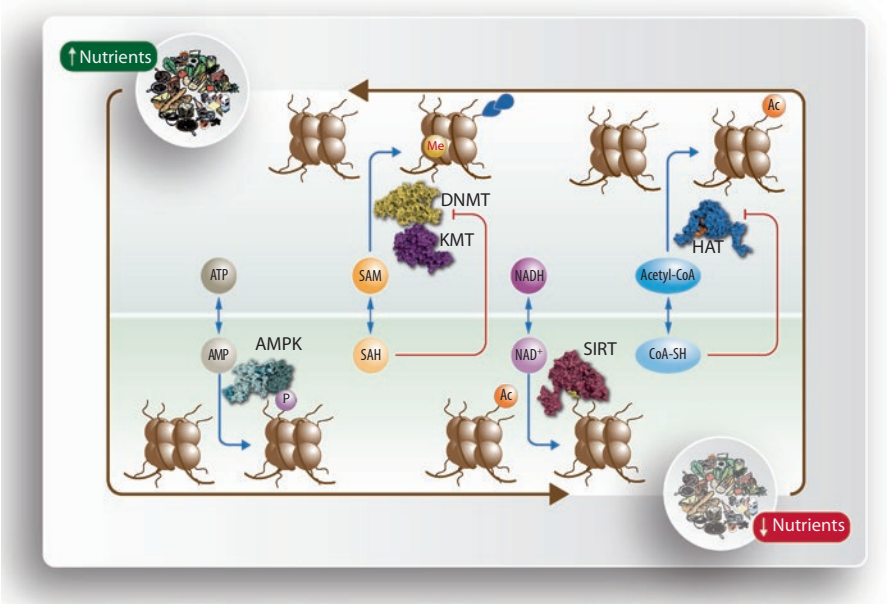


Fig. 10.5 **Epigenetic sensing of the nutritional state**. A high nutritional state of a cell (**top**) is represented by the abundance of the metabolites ATP, SAM, NADH and acetyl-CoA, while in the case of low nutrient levels (**bottom**) the metabolites AMP, SAH, NAD^+, and CoA are predominant. Accordingly, at high nutrient concentrations, KMTs and HATs are stimulated, while at low concentrations AMPK and HDACs of the SIRT family are activated and DNMTs and HATs are repressed. This results in histone methylation and acetylation or histone phosphorylation and deacetylation, respectively

energy status of a cell are memorized on the level of histone modifications and can be translated into functional outputs *via* adaptive gene regulation. In contrast, a high nutritional level results in low AMP levels, no AMPK activity, a modified histone phosphorylation pattern and the activity of a different set of genes. Thus, the metabolic state of a cell can be expressed by the ATP/AMP ratio, the SAM/SAH ratio, the NADH/NAD^+ ratio and the acetyl-CoA/CoA ratio (Fig. 10.5). Under high nutrient concentrations, such as abundant availability of methionine and glucose, SAM activates KMTs and acetyl-CoA stimulates HATs, thus leading to histone methylation and acetylation, respectively. In contrast, at low nutrient levels, such as during fasting, AMP activates AMPK and NAD^+ stimulates SIRTs resulting in histone phosphorylation and deacetylation. Moreover, in parallel SAH inhibits DNMTs and CoA blocks HATs.

10.3 Epigenomics of Intergenerational Metabolic Disease

Lifestyle factors, such as diet and physical exercise, play a more fundamental role in the pathogenesis in metabolic diseases, such as in type 2 diabetes, than a genetic predisposition (Sect. 12.1). The effects of our lifestyle choices concerning eating and moving are memorized in the epigenome of our metabolic organs. For example, epigenome-wide association studies and twin cohort studies have demonstrated differentially methylated CpG islands in metabolic organs when comparing type 2 diabetes patients with healthy individuals. Like in cancer (Sect. 8.2), these epigenomic regions may serve as early prognostic biomarkers for a pre-diabetic state. Lifestyle changes, such as a customized diet and increased physical activity, may reverse the epigenome changes and avoid developing type 2 diabetes. For example, a 6-month program of physical exercise demonstrated epigenome changes at thousands of genomic regions.

The concept of epigenetic programing *via* nutritional compounds suggests that dietary interventions, such as calorie restriction, "Mediterranean" or "Nordic" diet, can affect our chromatin status and lead to the expression of genes being beneficial to metabolic health. This implies that epigenetic states, which are initially fixed during embryogenesis, may shift in response to intrinsic and environmental factors, such as nutritional compounds. The epigenetic drift of metastable epialleles was already discussed in the context of epigenetic memory and transgenerational inheritance (Sect. 7.1). For example, changes in diet or in metabolism being associated with obesity can cause an epigenetic drift that may be inherited to the following generations.

In general, the offspring of mothers being exposed to an adverse environment during embryonic development, such as undernutrition or placental dys-function leading to impaired blood flow, nutrient transport or hypoxia, during adulthood has an increased risk to develop symptoms of the metabolic syndrome, such as obesity, impaired glucose tolerance and finally type 2 diabetes (Sect. 12.1). Thus, **critical disturbances in energy metabolism can lead to stable epigenetic changes that are maintained through the germ line and may affect the health of the next generations**. These observations were the basis for the thrifty phenotype concept (or formulated more general as "developmental origins of health and disease" (DOHaD) hypothesis) suggesting that poor nutrition in early life produces permanent changes in glucose-insulin metabolism (Fig. 10.6).

If this concept holds true, the worldwide growing epidemic of obesity and metabolic diseases may lead to a **significantly increased epigenetic predisposition for the metabolic syndrome in the subsequent generations** resulting in a vicious cycle (Sect. 12.1).

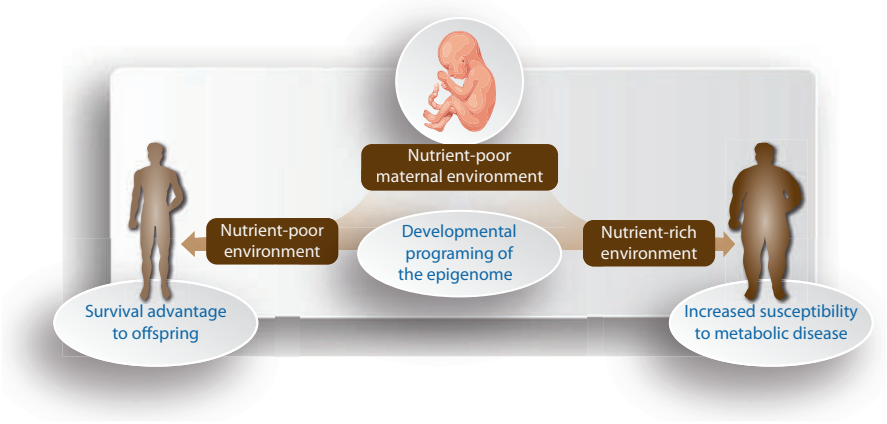


Fig. 10.6 **Thrifty phenotype/DOHaD hypothesis**. Intra-uterine stressors, including maternal undernutrition or placental dys-function (leading to impaired blood flow with consecutively hypoxia or reduced nutrient transport) can initiate abnormal patterns of development, histone modifications, and DNA methylation. Additional post-natal environmental factors, including accelerated post-natal growth, obesity, inactivity, and aging further contribute to the risk for type 2 diabetes potentially *via* changes in histone modifications and DNA methylation patterns of metabolic tissues. Obviously, epigenetic changes during embryogenesis have a much greater impact on the overall epigenetic status of an individual than that of adult stem cells or somatic cells, since they affect a higher number of following cell divisions

Further Reading

Avgustinova A, Benitah SA (2016) Epigenetic control of adult stem cell function. Nat Rev Mol Cell Biol 17:643–658

Barres R, Zierath JR (2016) The role of diet and exercise in the transgenerational epigenetic landscape of T2DM. Nat Rev Endocrinol 12:441–451

Carlberg C, Ulven SM, Molnár F (2016) Nutrigenomics Springer Textbook ISBN: 978-3-319-30413-7

Gut P, Verdin E (2013) The nexus of chromatin regulation and intermediary metabolism. Nature 502:489–498

Kinnaird A, Zhao S, Wellen KE, Michelakis ED (2016) Metabolic control of epigenetics in cancer. Nat Rev Cancer 16:694–707

Sales VM, Ferguson-Smith AC, Patti ME (2017) Epigenetic mechanisms of transmission of metabolic disease across generations. Cell Metab 25:559–571

Epigenetics of Immune Function

11

Abstract

Hematopoiesis is the lifelong regeneration of our blood cells; it produces far more cells than any other tissue of our body. Signal transduction pathways stimulated by growth factors regulate in combination with transcription factors and chromatin modifiers the self-renewal and differentiation of hematopoietic stem cells (HSCs). Most of the approximately 100 different blood cell types, which are generated by hematopoiesis, belong to the immune system and differ in their epigenetic programing, in particular at cell-specific enhancer regions. Epigenetic regulation is fundamental for the differentiation of immune cells, as well as for their adaptive response to several environmental challenges, such as microbe infections. These epigenetic processes are also the basis for trained immunity representing memory function of the innate immune system. In general, epigenetic profiling of immune cells is an important tool for a molecular description of health and diseases as different as allergic reactions and cancer.

Keywords

Hematopoiesis · HSCs · Poised enhancers · Trained immunity · Inflammation · Autoimmunity · Asthma · Epigenetic profiling

11.1 Epigenetics of Blood Cell Differentiation

Hematopoiesis is the process of lifelong regeneration of blood cells. Every day HSCs give rise to some 10^{11} cells; *i.e.*, over our lifespan the bone marrow produces far more cells than any other tissues of our body. This highly dynamic developmental process includes the self-renewal of multipotent HSCs as well as their differentiation in a hierarchical cascade. Hematopoiesis leads to 11 major lineages of mature blood cells including some 100 phenotypically distinct cell sub-types (Fig. 11.1a). Most of these cells belong to the immune system (Sect. 11.2). HSCs differentiate into

C. Carlberg, F. Molnár, *Human Epigenetics: How Science Works*,
https://doi.org/10.1007/978-3-030-22907-8_11

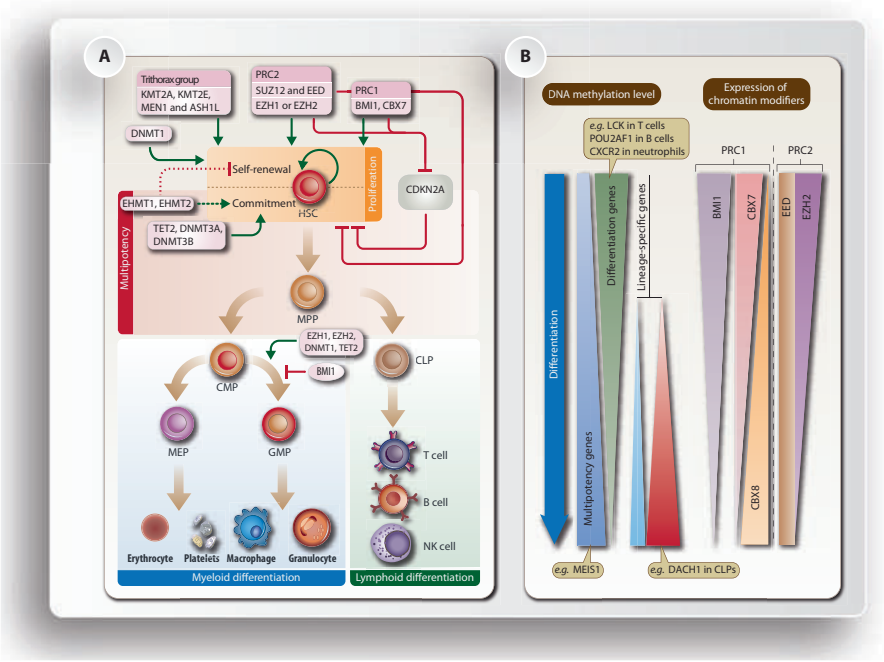


Fig. 11.1 **Epigenetics of hematopoiesis**. Key chromatin modifiers of HSC self-renewal and their differentiation during hematopoiesis are indicated (**a**). During hematopoietic lineage commitment DNA methylation levels change dynamically (**b, left**). The expression of key epigenetic regulators alters during hematopoiesis (**b, right**). More details are provided in the text

immature progenitor cells, such as multipotent progenitors (MPPs), which then give rise to the progenitors of the myeloid or lymphoid lineages, called common myeloid progenitors (CMPs) and common lymphoid progenitors (CLPs), respectively. CMPs further differentiate into megakaryocyte-erythrocyte progenitors (MEPs) and granulocyte-monocyte progenitors (GMPs). The cells of myeloid lineage include erythrocytes, platelets and cells of the innate immune system, such as neutrophils, eosinophils and monocytes (which can further differentiate into dendritic cells or macrophages). On the other side, the lymphoid lineage produces the main cells of the adaptive immune system, B and T cells. Erythrocytes are also referred to as to red blood cells, whereas myeloid and lymphoid cells are summarized as leukocytes or white blood cells.

A number of extrinsic and intrinsic factors, such as growth factor-stimulated signal transduction pathways, transcription factors and chromatin modifiers, regulate the equilibrium between self-renewal and differentiation of HSCs. In contrast, a disruption or mis-regulation of this process can lead to hematological disorders, such as leukemia, lymphoma and myeloma. In these diseases the excessive production of white blood cells in the bone marrow significantly raises their

levels in blood circulation. Moreover, **a disruption in any stage of hematopoiesis affects the production and function of blood cells and may have severe consequences, such as the inability to fight against infections or the risk of uncontrolled bleeding**.

During hematopoiesis the genome-wide DNA methylation pattern of differentiating blood cells changes dynamically and is very locus specific; *i.e.*, at some genomic regions there is a rise in methylation, while it decreases at other regions (Fig. 11.1b, left). The latter correlates with the upregulation of cell-specific genes and their encoded proteins, such as the tyrosine kinase LCK (LCK proto-oncogene, Src family tyrosine kinase) in T cells, the co-factor POU2AF1 in B cells and the chemokine receptor CXCR2 (C-X-C motif chemokine receptor 2) in neutrophils. In general, the commitment to a specific cell lineage increases the level of DNA methylation, since a larger set of genes is not anymore needed in these terminally differentiated cells, such as the transcription factor MEIS1 (meis homeobox 1), which maintains the undifferentiated state, or the myeloid-specific transcription factor DACH1 (dachshund family transcription factor 1) in lymphoid cells. Interestingly, the myeloid lineage is the default outcome of hematopoiesis, since its differentiation requires less correction by increased DNA methylation than that of the lymphoid lineage. Thus, **hematopoietic cells can be easily segregated based on their DNA methylation profile**, *i.e.*, their DNA methylome.

Parallel to alterations in the DNA methylation pattern, also key chromatin modifiers are changing their expression during hematopoiesis. For example, the expression of most members of the Polycomb family changes during HSC differentiation (Fig. 11.1b, right). The PRC1 components CBX7 and BMI1 (BMI1 proto-oncogene, Polycomb ring finger), which are responsible for the recognition and mono-ubiquitination of H2AK119 (Sect. 6.3), are highly expressed in HSCs but are downregulated during lineage commitment. In contrast, the CBX7 competitor CBX8 is upregulated. The PRC2 component EED (embryonic ectoderm development) does not change during hematopoiesis, while the H3K27-specific KMT EZH2 is downregulated. Similarly, members of the Trithorax group family, such as KMT2A, KMT2E, ASH1L (ASH1 like histone lysine methyltransferase) and MEN1 (menin 1), contribute to hematopoiesis. Furthermore, the KMTs EHMT1 and EHMT2 deposit repressive H3K9me2 marks to the epigenome of HSCs (Fig. 11.1a). Accordingly, **a mis-regulation of the genes encoding for these chromatin modifiers can result in hematopoietic failure**, such as HSC cell cycle arrest, pre-mature differentiation, apoptosis and defective self-renewal and finally leads to hematological malignancies.

In addition to chromatin modifiers, some master transcription factors, such as CEBPα (CCAAT/enhancer binding protein α), PU.1 (purine-rich box 1) and GATA2 (GATA binding protein 2), have a key role in hematopoiesis. They act as pioneer factors that directly bind nucleosomal DNA to prime enhancers for activation. These transcription factors then recruit chromatin remodeling and modifier complexes, which in turn facilitate removal and post-translational modification of nucleosomes at these genomic regions. For example, in myeloid progenitors the sustained expression of CEBPα generates macrophages, whereas under sustained expression

of GATA2 mast cells are created. However, when initially CEBPα and afterward GATA2 are expressed, eosinophils are produced, while the reversed order leads to basophils. Furthermore, CEBPα interacts with the DNA demethylating enzyme TET2, so that its target genes get demethylated during hematopoiesis. The activity of TET2 may be the key mechanism why myeloid cells are closer to HSCs than lymphoid cells. This fits with the observation that TET2 is mutated in several myeloid malignancies. Moreover, TET2 could link environmental conditions, such as nutrient availability (Sect. 10.2), to myeloid differentiation, since the metabolite 2-hydroxyglutarate inhibits TET2 activity and leads to DNA hyper-methylation.

Taken together, hematopoiesis is an important process insuring key function of our body by constantly providing new cells needed for oxygen transport (erythrocytes), blood coagulation after injuries (platelets) and immunity (leukocytes). **Transcription factors and chromatin modifiers work together in creating appropriate epigenetic profiles on the level of DNA methylation and histone modifications, which determine the respective functions of the more than 100 different cell types of the hematopoietic system**.

11.2 Role of Epigenetics in Immune Responses

Our immune system is a system of biological structures, such as the lymphatic system, cell types, such as leukocytes (cellular immunity), and proteins, such as antibodies and complement proteins (humoral immunity), protecting us against infectious diseases and cancer. Due to persistent hematopoiesis (Sect. 11.1) most cells of the immune system are replaced every few days to weeks. This also implies that their epigenetic training via encounters with microbes and other antigens is an ongoing learning event leading to **transient epigenetic memory in short-lived cells as well as to persistent memory in long-lived memory B and T cells** (Sect. 11.3).

The immune system detects a wide variety of molecules, known as antigens, of potential pathogenic origin, such as on the surface of microbes or malignant cells, and distinguishes them from our own healthy tissues. The functions of the immune system are classified as innate and adaptive. Innate immunity is evolutionary older, bases on the cell types monocytes/macrophages, neutrophils and natural killer (NK) cells, and uses destructive mechanisms against pathogens, such as phagocytosis, with the support of anti-microbial peptides from the complement system. Adaptive immunity applies more sophisticated defense mechanisms, in which T and B cells use highly antigen-specific surface receptors, such as T cell receptors and B cell receptors, the latter finally turning into secreted antibodies. Moreover, after an initial specific response to a pathogen, the adaptive immune systems creates an immunological memory that leads to an enhanced response to subsequent encounters with that same antigen.

Monocytes are produced from GMPs in the bone marrow (Fig. 11.1), released into the blood stream and migrate within 1-3 days through the endothelium of blood vessels into tissues. In tissues, monocytes differentiate into macrophages or dendritic

cells. This differentiation process is based on **epigenome changes in response to contacts with antigens, such as infectious microbes**. **These epigenetic changes create a memory of the microbe encounter** (Sect. 11.3). In response to their activation by pathogens or metabolites, macrophages secrete a number of signaling proteins, such as cytokines, chemokines and growth factors, which affect the migration and activity of other immune cells. This response is called acute inflammation, no matter whether it is caused by infection or injury. Acute inflammation is often associated with erythema, hyperthermia, swelling and pain, but it resolves within a few days to weeks. Inflammation can also derive from changes in the concentration of nutrients and metabolites. In this case, the immune system cannot cope the primary stimulus, so that chronic inflammation develops.

Most inflammatory lesions are initially dominated by monocyte-derived macrophages. The altered gene expression profile of these macrophages is based on changes of their epigenome in response to extra-cellular signals. The stimuli are classified into pathogen-associated molecular patterns (PAMPs), such as surface proteins of pathogenic bacteria, and damage-associated molecular patterns (DAMPs), such as excess of saturated fatty acids and other molecules representing metabolic stress (often called alarmins). The fact that DAMPs can induce the same cascade of inflammatory reactions like PAMPs explains why a wide variety of molecules and events cause inflammation and respective **specific changes in the epigenome of macrophages**.

Low-grade chronic inflammation is the central cause of many lifestyle-related diseases, such as obesity, insulin resistance, type 2 diabetes and atherosclerosis (Fig. 11.2). Moreover, also neurodegenerative diseases, such as Alzheimer's disease, most types of cancer, allergy, autoimmune diseases and inflammatory bowel diseases, such as Crohn's disease and ulcerative colitis, are closely linked to inflammation. Immune reactions in general and inflammation in particular are related to cellular metabolism. The proliferation of immune cells as well as their action in defense and tissue repair both require high levels of energy metabolism (Sect. 10.2). In turn, metabolic stress, which is often caused by lipid overload in the blood and in adipose tissue, stimulates low-grade chronic inflammation.

The direct or indirect contact of immune cells with microbes and other molecules with antigenic pattern results in effects on gene expression that are often stronger than in any other tissue or cell type of our body. The strong reaction is necessary, since bacteria proliferate far faster than human cells and may represent immediate danger to our body. Most cells of the innate immune system, such as monocytes, NK cells, macrophages or dendritic cells, express some 30 different variants of pattern recognition receptors on their surface that respond to the presence of PAMPs. The strength and specificity of the response of the immune cells, such as different populations of macrophages, depends on their epigenomic profile before encountering microbes (Sect. 11.3). This implies that the proper epigenomic programing of our immune cells before contact with antigen is essential for an optimal response. Thus, **proper epigenomic programing of our immune cells during hematopoiesis and antigen encounter is critical for a well functioning immune system**.

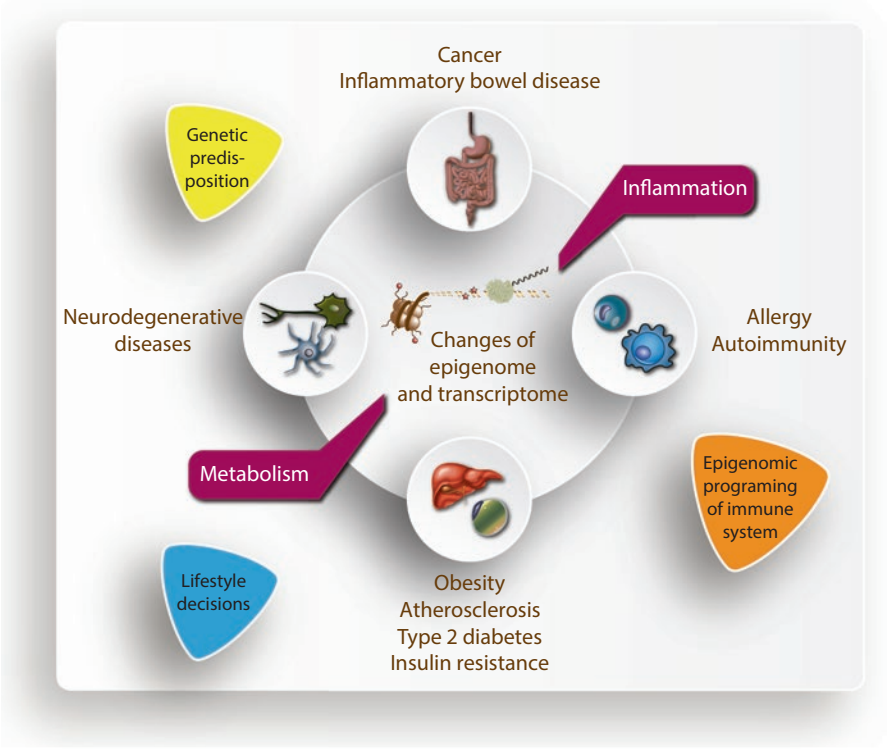


Fig. 11.2 Immune-mediated pathologies as key driver processes of diseases in various target organs. Inflammation and cellular metabolism are closely linked via coordinated changes in the epigenome and transcriptome of target tissues and cell types. More details are provided in the text

The more than 100 genes of the inflammatory cascade differ in their kinetics; *i.e.*, there are fast responding primary target genes, delayed responding secondary targets, and late responding tertiary targets. This is reflected by the underlying epigenetic changes in enhancer and promoter regions of the respective genes of macrophages. The promoter regions of primary target genes typically carry H3K4me3 and H3K27ac marks of active chromatin (Fig. 11.3, center). Moreover, these promoters often carry non-methylated CpGs. In contrast, the promoter regions of secondary target genes are first labeled by repressive H3K27me3 marks. The removal of H3K27me3 marks and CpG demethylation from these promoter regions after PAMP exposure and the introduction of H3K4me3 and H3K27ac marks take time. This explains the delayed response in the expression of the respective genes. The stimulation of macrophages with PAMPs also changes the chromatin status at enhancer regions, such as the *de novo* deposition of H3K4me1 and H3K27ac marks (Fig. 11.3, left). The activation of some of these enhancers is only transient, while others are marked more persistently and retain in this way a memory of the PAMP exposure, *i.e.*, of a contact with microbes.

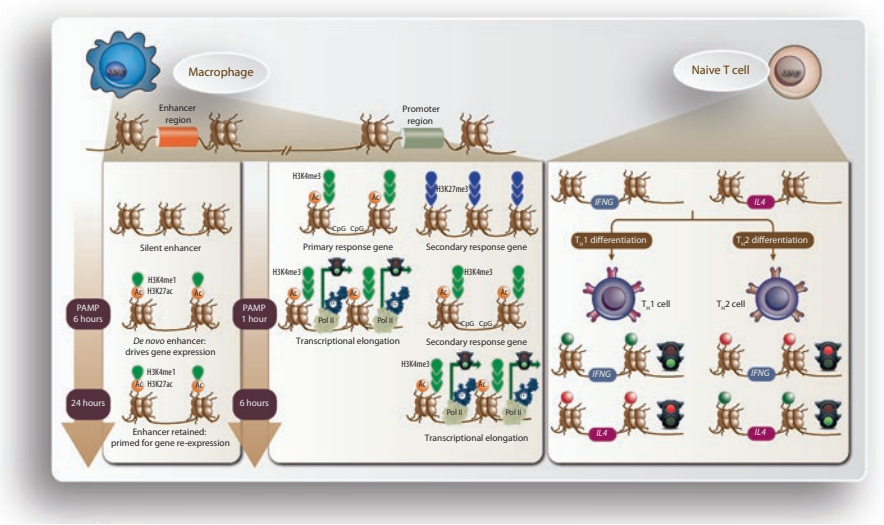


Fig. 11.3 **Epigenetic modifications in immune cells**. Enhancer (**left**) and promoter regions (**center**) of macrophages show a differential response after the exposure with PAMPs, such as lipopolysaccharide. Primary responses (after 1 h) and secondary responses (after 6 h) are distinguished. Even after removal of PAMP stimulation some enhancer regions can keep their activation status for 24 h or longer. This memory effect is part of trained immunity. Differential epigenetic programing of the regulatory regions of key cytokine genes, such as *IFNG* and *IL4*, is the key event in the polarization of T_H cell subtypes (**right**). In T_H1 cells the *IFNG* gene carries marks of active chromatin (green) and the gene is induced after antigen exposure. In contrast, in the same cells the *IL4* gene carries repressive histone markers (red) and stays repressed. In T_H2 cells the reverse process applies; *i.e.*, the *IL4* gene is induced and the *IFNG* gene stays repressed

Antigen binding to receptors of T and B cells, *i.e.*, of cells of the adaptive immune system, activates signal transduction pathways that potently trigger the expression of cytokine genes. Specific epigenetic changes to enhancer and promoter regions are involved in this antigen recognition process (Fig. 11.3, right). For example, the differentiation of T helper (T_H) cells into T_H1 and T_H2 subtypes involves epigenetic programing that primes these cells to increase after antigen exposure either the expression of the genes encoding for the cytokines INFG (interferon γ) or IL4, respectively. While the T_H1 response results in antigen clearance, the T_H2 response is typical for allergic reactions; *i.e.*, the differential epigenetic programing of these cells causes a clearly different physiological response. The epigenetic changes involve histone modifications but also the appearance of 5hmC marks at promoter regions and demethylation *via* TET2. The latter modifications are very stable and can last over 20 and more replication cycles of long-lived memory T cells.

The fact that alternative cell-specific enhancers and cell-specific transcription factors can regulate a given gene explains the differential expression of genes in

cells of the immune system. For example, the *FANTOM5 Project* (Sect. 1.3) identified some 44,000 enhancers from their large collection of primary human tissues and cells types; *i.e.,* there are approximately two enhancers per protein-coding gene. Thus, a given gene can be regulated in B cells *via* B cell-specific transcription factors to a B cell-specific enhancer, while in macrophages macrophage-specific transcription factors controls same gene *via* a macrophage-specific enhancer.

Inappropriate activation of the immune system can lead to a number of diseases, such as the allergic reactions of the respiratory tract in asthma or the autoimmune disease multiple sclerosis (MS). Immune responses often vary in the balance of pro-inflammatory and anti-inflammatory cytokines, *i.e.*, in the amount of possible collateral tissue damage. Thus, **gene expression patterns and the underlying epigenetic programing of immune cells are key factors in immune-mediated diseases**. Accordingly, individuals with autoimmune and/or inflammatory diseases have a clearly different epigenetic profile than healthy controls. For example, altered DNA methylation patterns of immune cells occur in multiple immune-mediated diseases, such as MS, systemic lupus erythematosus (SLE), or Crohn's disease. Patients with MS, in comparison to healthy subjects, show lower levels of 5hmC marks in their immune cells due to low expression of TET2. In contrast, SLE patients often have elevated 5hmC levels due to increased expression of TET2 and TET3 in T_H cells. This parallels with low global H3 and H4 acetylation and high H3K9 methylation levels in these cells.

Epigenome-wide profiling of immune cell subsets can help to identify the epigenomic basis of allergic diseases, such as asthma. For example, when comparing healthy controls with asthmatic patients, a marker for active and poised enhancers, H3K4me2, is significantly increased in T_H2 cells of patients. Asthma is a master example of a disease in which environmental exposure with natural and synthetic compounds causes dynamic changes of the epigenome. In general, immune system-related diseases, such as inflammation and autoimmune diseases, are clinically heterogeneous, but all **develop from the interplay of genetic susceptibility and environmental and/or lifestyle choices, *i.e.*, the balance between genetics and epigenetics** (Sect. 12.1). Thus, epigenetic profiling of cell subsets for various epigenetic markers, such as accessible chromatin, DNA methylation, histone modifications, transcription factor binding, and chromatin modifier association, is an important tool for a molecular understanding of the diseases and can provide hints for a possible therapy, for example, by small-molecule inhibitors of chromatin modifiers (Sect. 12.3).

Antigens stimulate the clonal expansion of naïve cytotoxic T cells, *i.e.*, the growth of antigen-recognizing T cell subsets. This is followed by cellular differentiation, which activates a network of effector genes, such as those encoding for cytokines, *via* the demethylation of their genomic regions. The resulting effector T cells are able to eliminate directly and indirectly antigen-presenting cell. Anergy is an epigenetically induced dys-functional state of T cells that often occurs during

prolonged exposure to an antigen, such as in cancer and in chronic infections. This results in *de novo* DNA methylation of effector genes and their inactivation; *i.e.*, exhausted T cells show no further immune response. The methylation of the effector genes can be blocked and may be reversed by the application of a demethylating agent, such as decitabine (Sect. 3.1).

11.3 Epigenetic Basis of Immunological Memory

T cells constitute up to 30% of all circulating leukocytes and are a major component of the adaptive immune system. They occur in a number of important subtypes, such as T_H, cytotoxic T and regulatory T cells. The impact of epigenetic changes in the context of regulatory T cell proliferation is demonstrated at the example *FOXP3* (forkhead box P3) gene (Fig. 11.4). In this case, the transcription factors REL (REL proto-oncogene, NF-κB subunit), CBFB (core-binding factor subunit β) and RUNX1 (runt-related transcription factor 1) both bind enhancers downstream of the *FOXP3* promoter and stimulate *FOXP3* mRNA expression. The binding of the transcription factors to the enhancers leads to their rapid demethylation allowing FOXP3 protein binding for stable auto-regulation. Thus, **transcription factor-mediated local demethylation of an enhancer region creates epigenomic memory that stabilizes T cell lineage progression over multiple cell divisions**.

Recent studies, in particular of the *BLUEPRINT* (www.blueprint-epigenome.eu) consortium, demonstrated that cells of the innate immune system, such as monocytes/macrophages and NK cells, also have a memory function, referred to as trained immunity (Fig. 11.5). This rather short-term epigenetic memory monitors the close relationship between immune challenges and effects on chromatin. **Trained immunity is based on epigenetic changes, such as DNA methylation and histone modifications, as well as on the actions of miRNAs and long ncRNAs**. The rather long half-life of the latter molecules makes them well suited for a persistent programing of the epigenome.

Trained immunity enables innate immune cells to react with a quantitatively different response, *i.e.*, a higher magnitude of gene expression when they are re-challenged with a pathogen (Fig. 11.6, top left). This response can in part also be qualitatively different, such as *via* the expression of an alternative pattern recognition receptor (Fig. 11.6, bottom left). A key mechanism in trained immunity is enhancer poising (Sect. 4.2), *i.e.*, the addition of persistent histone marks, such as H3K4me1, enabling a strong response after restimulation (Fig. 11.6, right).

Most immune cells are very mobile and experience many different microenvironments throughout our body. This leads to wide range of different signals and respective **adaptive epigenetic programing of the enhancer repertoire of these cells**. There is equilibrium between the persistence of an epigenome instructed by previous stimuli and the reprograming in response to a changing environment.

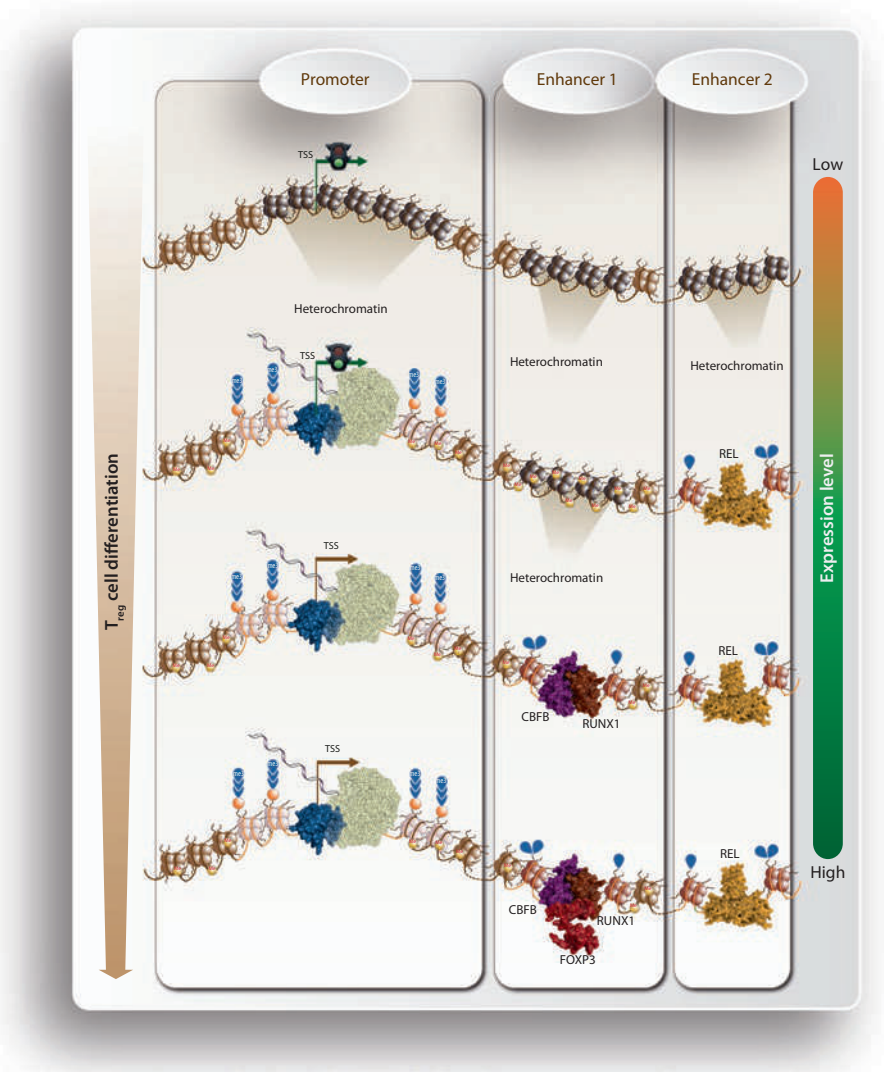


Fig. 11.4 **Epigenetic memory of transcriptional activity**. During differentiation of regulatory T cells the *FOXP3* gene must show a stable and strong expression. A homodimer of the transcription factor REL binds to a downstream enhancer and stimulates *FOXP3* mRNA expression and demethylation of the promoter region. *FOXP3* gene expression is stabilized through the binding of the transcription factors CBFB and RUNX1 to a more proximal enhancer. The latter induces local demethylation and permits binding of the transcription factor FOXP3. Thus, in an auto-regulatory fashion FOXP3 ensures constitutive activity of the promoter of its own gene. The demethylation of enhancer and promoter regions represents epigenetic memory

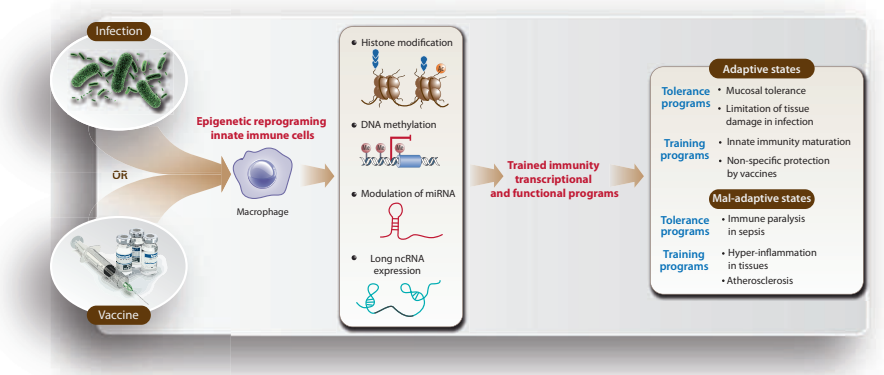


Fig. 11.5 **Epigenetics of innate immune cells**. The activation of the innate immune cells, such as monocytes, macrophages, or NK cells, leads to their epigenetic reprograming, known as trained immunity. This innate immune memory leads to adaptive states that protect the host during and after infections. In certain situations, however, trained immunity can result in mal-adaptive states, such as immune paralysis after sepsis or hyper-inflammation

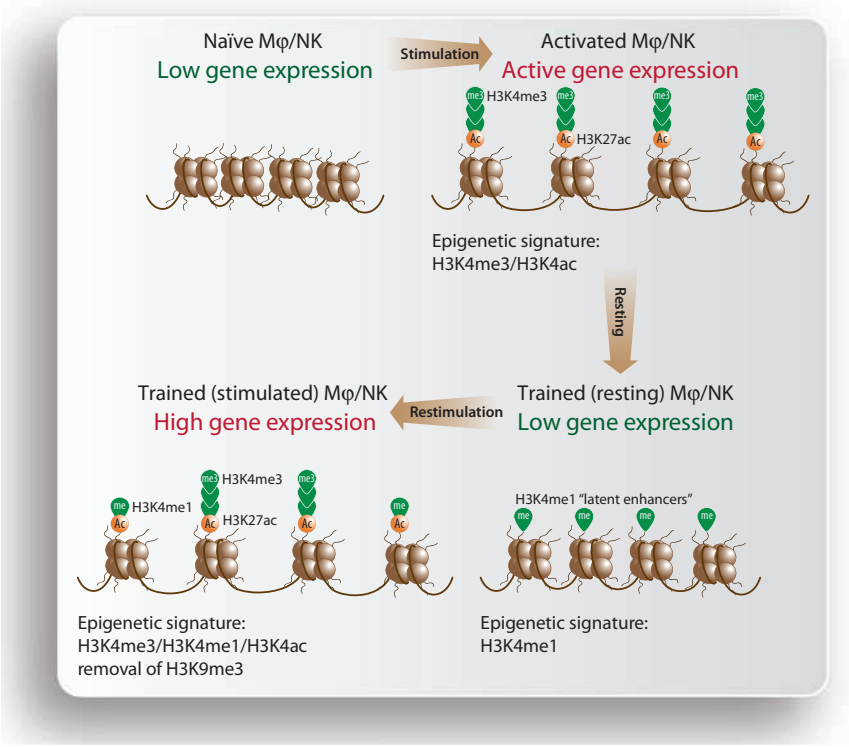


Fig. 11.6 **Trained immunity**. Enhanced inflammatory and anti-microbial properties of innate immune cells (**top left**) are a memory phenomenon that is referred to as trained immunity. It is based on epigenetic reprograming of innate immune cells, such as macrophages (Mϕ) and NK cells, and results, for example, in the increased and/or alternative expression of genes encoding pattern recognition receptors (**bottom left**). The first round of stimulation of the cells leaves persistent H3K4me1 marks on enhancer regions. This enhancer poising enables them to respond faster and stronger to a restimulation (**right**)

Further Reading

Alvarez-Errico D, Vento-Tormo R, Sieweke M, Ballestar E (2015) Epigenetic control of myeloid cell differentiation, identity and function. Nat Rev Immunol 15:7–17

Avgustinova A, Benitah SA (2016) Epigenetic control of adult stem cell function. Nat Rev Mol Cell Biol 17:643–658

Benayoun BA, Pollina EA, Brunet A (2015) Epigenetic regulation of ageing: linking environmental inputs to genomic stability. Nat Rev Mol Cell Biol 16:593–610

Ellmeier W, Seiser C (2018) Histone deacetylase function in $CD4^+$ T cells. Nat Rev Immunol 18:617–634

Henning AN, Roychoudhuri R, Restifo NP (2018) Epigenetic control of $CD8^+$ T cell differentiation. Nat Rev Immunol 18:340–356

Hu D, Shilatifard A (2016) Epigenetics of hematopoiesis and hematological malignancies. Genes Dev 30:2021–2041

Netea MG, Joosten LA, Latz E, Mills KH, Natoli G, Stunnenberg HG, O'Neill LA, Xavier RJ (2016) Trained immunity: a program of innate immune memory in health and disease. Science 352:aaf1098

Vahedi G, Richard AC, O'Shea JJ (2014) Enhancing the understanding of asthma. Nat Immunol 15:701–703

Winter DR, Jung S, Amit I (2015) Making the case for chromatin profiling: a new tool to investigate the immune-regulatory landscape. Nat Rev Immunol 15:585–594

Epigenome-Environment Interactions and Their Therapy

12

Abstract

Despite the huge number of more than 88 million variations of our genome, as revealed by the *1000 Genomes Project*, for most common multigenic diseases only some 20% of the genetic risk can be explained. A part of the missing heritability might be resolved by future identification of rare SNPs, but the major contributions to this phenomenon will be environmental factors and lifestyle choices modulating the epigenome. For the same reason, human populations that made in only few generations a transition from famine to food surplus are under higher risk for developing metabolic diseases than those that improved nutritional conditions over many generations. "Integrative personal omics profiling" (iPOP) represents the most comprehensive assessment of individuals and serves as a master example for the detection of genome-environment interactions. Most epigenetic modifications are reversible, which implies significant therapeutic potential for inhibitors of chromatin modifiers. These compounds are used in the therapy of immune diseases and in particular in the immunotherapy of cancer.

Keywords

Genetics · Environment · GWAS · Missing heritability · Integrative personal omics profiling · Chromatin modifier inhibitors · Cancer immunotherapy

12.1 Environment Versus Genetics

The *1000 Genomes Project* indicated that humans have in total more than 88 million genomic variants (Box 1.1). However, each of us is differing from any unrelated individual only by in average some four million genetic variants, which cover, due to a few thousand larger structural variants, about 12 Mb of DNA sequence. This means that **99.7% of the genome of all humans is identical, whereas only 0.3% explain our individual traits**, such as height, eye and hair color, as well as our

C. Carlberg, F. Molnár, *Human Epigenetics: How Science Works*,
https://doi.org/10.1007/978-3-030-22907-8_12

predisposition for common multigenic diseases, such as type 2 diabetes, atherosclerosis, and Alzheimer's.

Early studies of human genetics established the basis of "Mendelian" disorders, for which, in most cases, a single SNP can explain the occurrence of the disease; *i.e.*, they are monogenetic. During the last 15 years, genome-wide association studies (GWASs) became very popular for the genetic analysis of multigenic diseases. GWASs employ an "agnostic" approach in the search for unknown disease variants; *i.e.*, hundreds of thousands of SNPs are tested for association with a disease in large cohorts of patients versus healthy controls. At present (April 2019) the database GWAS Catalog (www.ebi.ac.uk/gwas) contains more than 130,000 disease-associated loci. The impact of SNPs on the protein coding sequence of our genome (less than 2% of all) is well established. Synonymous mutations do not alter the encoded protein, but non-synonymous mutations cause a change in the amino acid sequence (missense) or introduce a pre-mature stop codon (nonsense). Small insertions or deletions (indels) as well as larger copy number variations in exonic sequences can result in frame-shift mutations. Moreover, variations in intronic sequences may lead to alternative splicing. However, the vast majority of genetic variants are located in regulatory and not in coding regions of genes; *i.e.*, **the phenotypic consequences of most genetic variants are rather based on an epigenetic or gene regulatory processes than on a change in protein function**.

The functionally most relevant epigenomic variations occur at gene regulatory elements, such as at CpGs or transcription factor binding sites in promoter and enhancer regions. For example, a SNP within the DNA binding site of a transcription factor either facilitates, enhances or inhibits the binding of the respective protein. The transcription factor then influences the local chromatin structure *via* the recruitment of chromatin modifying enzymes, leaving marks on the local chromatin region, eventually leading to the activation of Pol II and the transcription of the respective gene. This may have a positive effect on the trait of interest. In contrast, when the transcription factor cannot bind, the genomic region remains inactive and the gene is not transcribed, which may have a negative effect on the studied trait. In total, the DNA methylation levels at thousands of CpGs are affected by genetic variants. These genomic sites are referred to as methylation quantitative trait loci (QTLs). The vast majority of epigenetic variations are in *cis* to their consequences on chromatin activity and gene expression, *i.e.*, they occur within the same genomic region, such as a TAD. In contrast, *trans*-acting epigenetic variations, like epimutations in pluripotency transcription factors, are very sparse suggesting that they are highly deleterious. In fact, they are key mutations in cancer epigenomics (Chap. 8).

GWASs with 2000 to 5000 individuals confidently identified common variants with effect sizes, referred to as odds ratios (ORs), of 1.5 or greater, *i.e.*, a 50% increased risk for the tested disease. Larger sample sizes were achieved by pooling several GWASs through meta-analyses. For example, sample sizes of at least 60,000 subjects provide sufficient power to identify the majority of variants with ORs of 1.1, *i.e.*, a 10% increased risk. However, despite some notable successes in revealing numerous novel SNPs and genomic loci associated with complex phenotypes, **most**

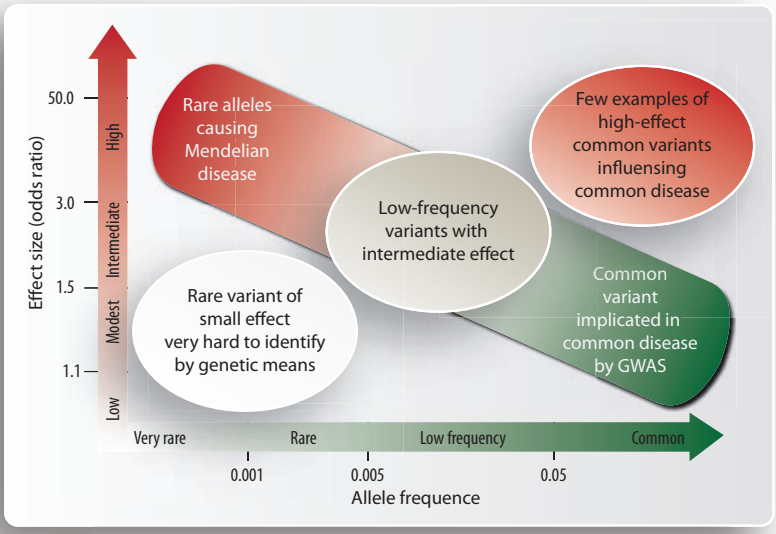


Fig. 12.1 **Identifying genetic variants by risk allele frequency**. The strength of the genetic effect is indicated by odds ratios. Most emphasis and interest lies in identifying associations with characteristics shown within the diagonal box

of polygenic traits have less than 20% of their heritability explained by the common variants.

Common SNPs of a complex multigenic disease are characterized by low ORs (Fig. 12.1, right), while rare monogenetic forms of the disease have high ORs (Fig. 12.1, left). It is expected that whole genome sequencing of large numbers of individuals will identify in the future far more low frequency SNPs with intermediate ORs (Fig. 12.1, center), which may better explain an individual's genetic disease risk. However, the phenomenon of missing heritability will remain, where up to 80% of the risk for a particular disease cannot be reliably estimated. Assessing disease risk based on whole-genome sequencing is limited, since this cannot capture the role of the environment majorly affecting the epigenome. The only well known exceptions are age-related macular degeneration and type 1 diabetes, for which the combinations of common and rare variants can provide a quantifiable risk profile. Thus, **some of the missing heritability may be explained by unidentified rare variants with high ORs but most of it by epigenetic variants**.

Dietary fat composition affects DNA methylation in adipocytes, which is one of multiple indications that the metabolic syndrome and related disorders are linked to epigenetic changes. The lifestyle choice of using a Western diet is the leading cause of type 2 diabetes as well as a major cause of multiple types of cancer. In parallel,

exposure to nicotine and other toxins causes substantial epigenetic changes in various organs explaining why smoking is the leading cause of several cancers and also contributes to respiratory and autoimmune diseases. All of these environment/lifestyle-caused epigenetic changes are associated with low-grade chronic inflammation (Sect. 11.2), which further amplifies the risk for autoimmune disease and cancer. Thus, **most age-related diseases are the result of long-term environmental exposure with food overload and toxins resulting causing epigenome changes in many organs**.

In general, epigenomic variations have different properties, such as being

- heritable or not
- associated with a SNP or not
- associated with environmental factors or not
- in *trans* or *cis* concerning gene regulation
- an epimutation in somatic or germ cells.

Within more or less one generation (33 years), between 1981 and 2014, the worldwide prevalence for obesity doubled. Very obviously the Western diet combined with decreased physical activity are the main environmental contributors to obesity and the subsequent development of the metabolic syndrome. In addition, human populations that made a transition from famine to food surplus just within 1–2 generations are under significantly higher risk for obesity, type 2 diabetes, and the metabolic syndrome, than those that were improving their nutritional conditions over many generations. This means that populations who are born and live in countries that had particularly rapid changes in urbanization and economic development have an increased risk of the metabolic syndrome in the years to come. This suggests that **epigenetic mechanisms rather than genetic variations of the genome play a role in the obesity epidemic and its associated metabolic abnormalities**.

There are increasing epidemiological and clinical evidences that the concept of the thrifty phenotype (Sect. 10.3), *i.e.*, a pre-natal epigenetic programing *in utero*, may be a key cause for metabolic diseases. Individuals that carry an epigenome that during their anthropologic development was programed by suboptimal nutrition *in utero*, despite normal post-natal nutrition, transgenerationally transmits a predisposition for obesity (Fig. 12.2). So far, there is no comprehensive analysis of the epigenome of individuals suffering from the metabolic syndrome, but a high number of genomic regions will be affected in an individual-specific way. Nevertheless, since epigenetic modifications respond dynamically to environmental conditions, there is potential for therapeutic intervention and reversibility (Sect. 12.3).

Dietary reference values are mainly based on requirements from population groups and not from individuals. However, enormous variability exists in individual responses to diet and food components that affect overall health. Both genetic and

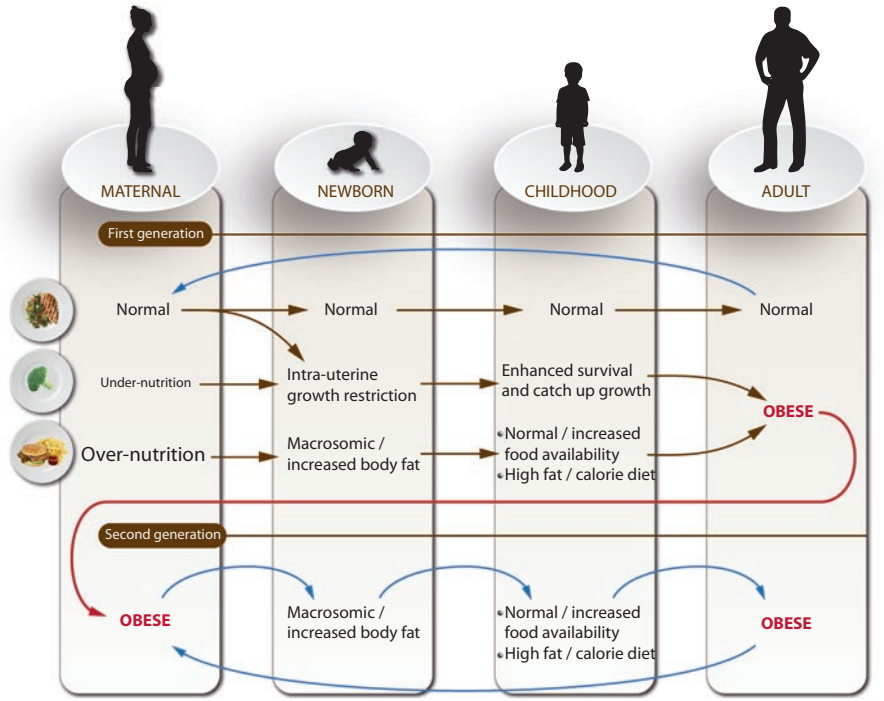


Fig. 12.2 **Epigenetic programing and the shift of populations toward obesity**. Non-obese mothers usually give birth to non-obese children, which develop into adults with a normal metabolic profile and a normal body fat content. However, undernutrition combined with improved neo-natal survival, formula feeding and exposure to a Western post-natal diet increases the incidence of pre-maturity and intra-uterine growth restriction. This results in increased obesity of the offspring and higher risk for obtaining the metabolic syndrome, when pre-natally exposed to Western diet. Some obese mothers may give birth to newborns with increased body fat, as a result of consumption of a high-fat diet. All these processes contribute to a shift of the population toward an obese phenotype. This also includes that second generation obese women have an increased risk to give birth to infants with increased body fat content and a further increased risk to develop obesity and the metabolic syndrome

environmental factors influence the individual's response (Fig. 11.2). Discoveries underpinning this variability will lead to advances in personalized nutrition as well as in improved health and food policies, including dietary reference intakes for nutrient needs and future dietary recommendations. A top priority for future nutrition research is a better understanding of the variability in metabolic responses to diet. Cellular and molecular characterization of disease phenotypes is therefore crucial to understand the role of food components in disease prevention and treatment, which is one of the key concepts of nutrigenomics (Sect. 10.1).

12.2 Epigenome-Wide Diagnosis

The field of personalized medicine significantly advanced, primarily due to the rapid development of next-generation sequencing technologies. Future personalized health care as well as the emerging field of personalized nutrition will benefit from the combination of personal genomic information with global monitoring of the molecular profile that represent physiological states. A prove-of-principle approach is exemplified by iPOP analysis of some 100 individuals (http://snyderlab.stanford.edu/iPOP.html) demonstrating the potential of next-generation technologies (Fig. 12.3). The iPOP study included a whole genome sequencing and sampling

- mRNA and miRNA expression in PBMCs
- proteome profile in PBMCs and in serum
- metabolome and auto-antibodyome in blood plasma.

The study is carried out longitudinally over many months to years. The molecular datasets are complemented by medical lab tests for regular blood biomarkers. Interestingly, the frequent sampling enabled the detection of personalized

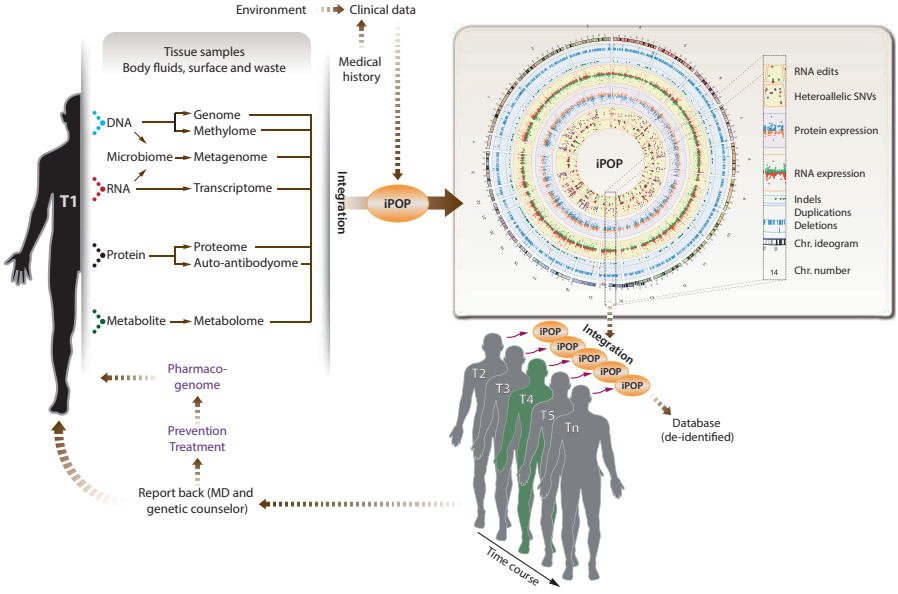


Fig. 12.3 **Implementation of iPOP for personalized medicine**. Tissue samples (for example, PBMCs) of an iPOP participant are collected at time points T1 to Tn, while diet, exercise, medical history and present clinical data are also recorded (**left**). The results of the iPOP analysis can be monitored by Circos plots (**right**), in which DNA (outer ring), RNA (middle ring) and protein (inner ring) data match to chromosome position. The data may be reported back to genetic medical practitioners and/or counselors in order to allow most rational choices for prevention and/or treatment, which may be matched with pharmacogenetic data

physiological state changes, such as markedly elevated glucose levels following two viral infections during the investigation period for the first test persons. The integrative profile monitored both gradual trend changes as well as spike changes in particular at the onset of each physiological state adjustment. Thus, the **iPOP analysis allowed a most comprehensive view on the biological pathways that changed during type 2 diabetes onset** of the study subject including dynamic changes in allele-specific expression and RNA editing events. Importantly, the type 2 diabetes outbreak was detected in a very early stage, so that it could be effectively controlled and reversed by changing both the individual's diet and intensified physical exercise, respectively.

The central aim of an iPOP-type analysis is the accurate assessment of disease risk of investigated individuals. Due to the large number of genetic variants and the fact that most diseases are based on a combination of genetic and environmental factors (Sect. 12.1), this aim is challenging. Results from genotyping approaches summarize the disease risk based on age, gender, and ethnicity as well as multiple independent disease-associated SNPs, in order to calculate the subject's likelihood of developing a disease. **The original iPOP study calculated for the investigated individual a previously unexpected increased risk to develop type 2 diabetes**, which in fact was confirmed experimentally after a viral infection. iPOP-type investigations can be tailored and applied to monitor any disease or physiological state changes of interest. The integrative profile is modular and allows the addition of further omics information, such as epigenome-wide data and the microbiome of skin, oropharynx, nasopharynx, stomach, intestinal mucosa or urine, respectively, as well as quantifiable environmental factors. In this way, iPOP-like analyses may become also central to nutrigenomics (Sect. 10.1).

In a few years, whole genome sequence information will be available for millions of individuals. This will allow a deeper understanding of the processes of human evolution and the causes of patterns of genetic variations for all human populations. The rapid development of omics technologies in combination with decreasing costs will allow collecting iPOP-style large-scale datasets on many individuals, the integration of which will allow further exploring the relationship between human (epi)genetic variations and complex diseases and respective traits. In particular, **the systematic exploration of epigenomics will provide critical insights into disease susceptibility**. The ability to stratify individuals according to their genotype will make clinical trials more efficient by enrolling a lower number of subjects with an anticipated larger effect when personalizing the intervention. Diseases, such as type 2 diabetes, will be classified into sub-phenotypes based on the genotype and the dynamic reply of the individual, for example in response to a personalized diet.

Interestingly, each individual is heterozygous for 50–100 genetic variants that may cause inherited disorders in homozygous offspring. This will provide a large demand and challenge for genetic counseling based on whole-genome sequencing. Moreover, (epi)genome-environment interactions provided by lifestyle choices will create an additional level of complexity. Within the next few years a number of next-generation sequencing applications will be incorporated into clinical (epi)genetic

diagnostics, but it is yet unclear, how this will be financed. Nevertheless, **further developments of next-generation sequencing will stay a driving force in basic biomedical research**.

12.3 Epigenetic Therapy of Diseases

During aging histone acetylation and methylation of many genomic regions changes, most likely because SIRTs can promote gene silencing and longevity (Sect. 7.3). Similarly, the epigenome of cancer cells is also reprogramed during the transformation process from normal cells. The mapping of active and repressed chromatin regions in cancer cells allows more accurate prognosis and even may facilitate therapy (Sect. 8.5). For example, DNMT and HDAC inhibitors have already been approved for cancer treatment. In addition, numerous psychiatric disorders, such as anxiety and depression, can be treated with HDAC inhibitors (Sect. 9.3). Interestingly, these small-molecule inhibitors also enhance the efficacy of immunotherapeutic agents, such as a blockage of the interaction between the surface inhibitory receptor PDCD1 (programmed cell death 1, also called PD1) on cytotoxic T cells and CD274 (CD274 molecule) on cancer cells (Fig. 12.4). At present, immunotherapy is the most promising therapy of cancer, since it takes advantage of the general immune surveillance function of cytotoxic T cells. In these cells a PDCD1-induced signal transduction cascade would inhibit their activation, which is prevented *via* blocking of PDCD1. This so-called immune checkpoint blockage can boost the anti-tumor immune response of the host, when the cytotoxic T cells start again their growth and effector function. Moreover, the DNMT inhibitors azacitidine and decitabine induce the expression of genes encoding for major histocompatibility complexes (MHCs) or tumor antigens. This increases the visibility of the cancer cell to cytotoxic T cells and their subsequent elimination. In addition, decitabine increases the sensitivity of cancer cells to growth inhibition by type 1 interferons. This results in a "viral mimicry", in which DNA demethylation activates the transcription of endogenous retroviral elements in the cancer cells leading to a double-stranded RNA-mediated immune response. The combination with an epigenetic mediator, such as chromatin modifier inhibitors, gives immune checkpoint blockage of T cells a wider approach for treating both cancer and chronic infections. For example, Hodgkin's lymphoma patients, who received a DNMT inhibitor before a treatment with immune checkpoint inhibitors showed a higher rate of complete remission in their cancer.

Importantly, most epigenomic modifications are reversible, which implies significant therapeutic potential. Therefore, **epigenomics is one of the most innovative research areas in modern biology and biomedicine, where the molecular hallmarks of epigenetic control can be used as targets for medical interventions and treatments** (Fig. 12.5). For example, iPS cells have the potential to regenerate damaged tissues: they originate from differentiated adult cells, which were overexpressed with pluripotency transcription factors binding to super-enhancers, however, their epigenome differs from ES cells. The latter cells are more depleted with

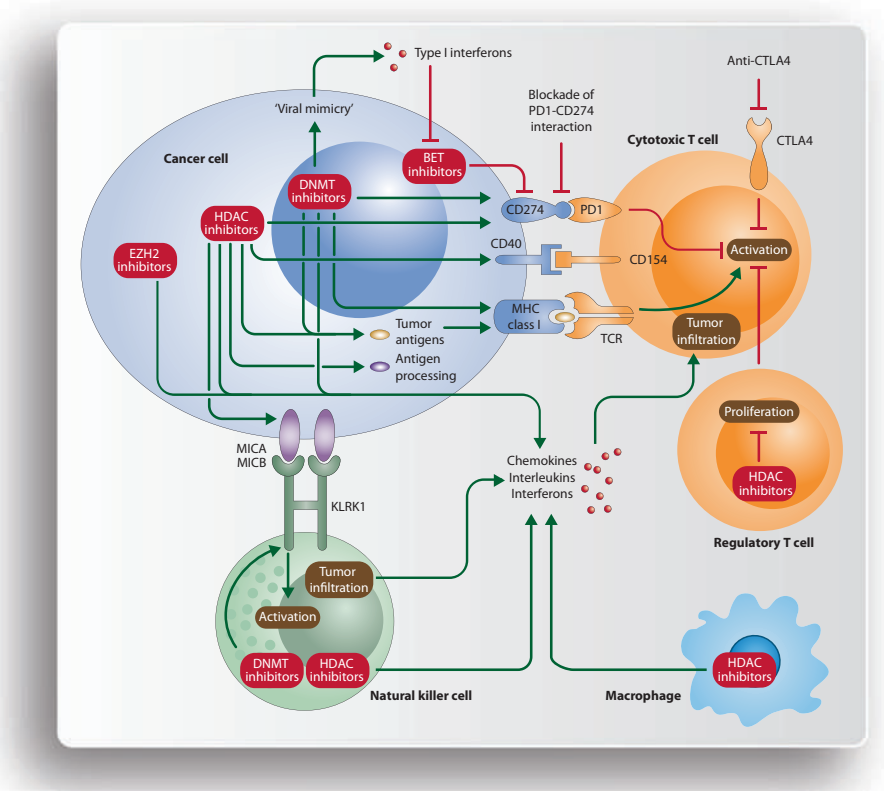


Fig. 12.4 **Inhibitors of chromatin modifiers in immunotherapy**. Epigenetic inhibitors also may play an important role in immuno-oncology. HDAC inhibitors also modulate the expression of MHC proteins, co-stimulatory CD40 molecules and tumor antigens. Moreover, they affect the antigen-processing machinery and change in chemokine expression in both cancer and immune cells. Furthermore, they suppress T_H cells and induce of the NK cell receptor ligands MICA and MICB. Inhibitors of the PRC2 component EZH2 increase the expression of chemokines CXCL9 and CXCL10 attracting T cells and improving tumor clearance. BET, bromodomain and extra-terminal; CTLA4, cytotoxic T-lymphocyte associated protein 4; KLRK1, killer cell lectin like receptor K1; MHC, major histocompatibility complex; MIC, MHC class I polypeptide-related sequence; TCR, T cell receptor

respect to marks of repressive chromatin and are more responsive to chromatin remodeling (Sect. 6.3).

The promises of epidrugs increased the number of compounds that are in pre-clinical or clinical trials. In addition to writers and erasers, now also chromatin modifiers of the reader class, such as chromatin-remodeling proteins with bromodomain or methyl-binding proteins, are addressed. Oncology is currently the main focus of clinical epigenetics, but the large number of clinical trials or pre-clinical studies in other medical areas indicates that in the near future **clinical epigenetics will expand far beyond cancer**.

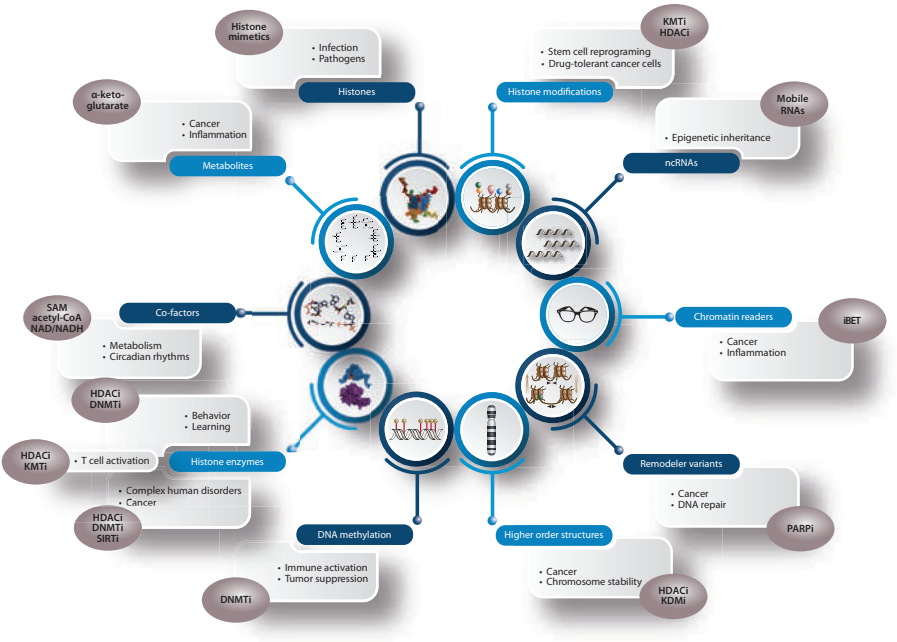


Fig. 12.5 **Prognostic and therapeutic potential of epigenomics**. Examples for the impact of epigenomics in normal development and in disease are indicated. The circles each represent key epigenomic mechanisms and the mainly associated nuclear proteins. Dys-regulated epigenomics may be reversed by pharmacological intervention with small-molecule inhibitors, such as HDAC inhibitors (HDACi), DNA methylation inhibitors (DNMTi), SIRT inhibitors (SIRTi), metabolic co-factors, such as SAM and α-ketoglutarate, histone lysine methyltransferase inhibitors (KMTi), bromodomain and extra-terminal inhibitors (iBET), poly(ADP-ribose) polymerase inhibitors (PARPi), and histone lysine demethylase inhibitors (KDMi)

Further Reading

Berdasco M, Esteller M (2019) Clinical epigenetics: seizing opportunities for translation. Nat Rev Genet 20:109–127

Feinberg AP (2018) The key role of epigenetics in human disease prevention and mitigation. N Engl J Med 378:1323–1334

Ghoneim HE, Fan Y, Moustaki A, Abdelsamed HA, Dash P, Dogra P, Carter R, Awad W, Neale G, Thomas PG et al (2017) De novo epigenetic programs inhibit PD-1 blockade-mediated T cell rejuvenation. Cell 170:142–157

Li-Pook-Than J, Snyder M (2013) iPOP goes the world: integrated personalized omics profiling and the road toward improved health care. Chem Biol 20:660–666

Tough DF, Tak PP, Tarakhovsky A, Prinjha RK (2016) Epigenetic drug discovery: breaking through the immune barrier. Nat Rev. Drug Discov 15:835–853

Glossary

Acute inflammation a short-term immunological process occurring in response to tissue injury or microbe infection usually appearing within minutes or hours. It is characterized by pain, redness, swelling and heat.

Anergy a tolerance mechanism of T cells, in which after an antigen encounter cells are inactivated *via* epigenomic reprograming.

Assay for transposase accessible chromatin using sequencing (ATAC-seq) a method similar to DNase I hyper-sensitivity and FAIRE-seq mapping, which is used to identify active regulatory sites characterized by lower density of nucleosomes. ATAC-seq uses the Tn5 transposase, which can insert sequencing adaptor sequences only into regions free of nucleosomes.

Autistic spectrum disorders a group of neurodevelopmental diseases characterized by deficits in social and communicative interaction and stereotypic behaviors.

Basal transcriptional machinery a large number of basal transcription factors (many of which are summarized as the TFIID complex) located at the TSS and using Pol II as its core. Via another multi-protein complex of co-activators (termed the Mediator complex) the basal transcriptional machinery is connected with activating and repressing cell- and site-specific transcription factors binding to enhancer regions.

Beckwith-Wiedemann syndrome a predominantly maternally transmitted disorder, involving fetal and post-natal overgrowth and a predisposition to embryonic tumors. The disease locus includes several imprinted genes, including *IGF2*, *H19* and *KCNQ1*, and loss of imprinting at *IGF2* is seen in ~20% of cases.

Bisulfite sequencing a method to study 5mC DNA methylation. Native DNA is exposed to sodium bisulfite, as a result of which non-methylated cytosines undergo deamination and are converted to uracils (which are read as thymines), whereas methylated cytosines remain unconverted. Sequencing libraries are generated from the converted template and they allow the study of methylation at single-base resolution.

Bivalent chromatin chromatin regions that harbor active and repressive histone modifications. Bivalent chromatin domains mark genes that are expressed at low levels only but are poised for activation upon an intra- or extra-cellular signal.

C. Carlberg, F. Molnár, *Human Epigenetics: How Science Works*,
https://doi.org/10.1007/978-3-030-22907-8

Blastocysts early stage embryos that have undergone the first cell lineage specification, which results in two primary cell types: cells of the inner cell mass and trophoblasts.

Bromodomain a protein module of ~110 amino acids that mediates interaction with acetylated lysines and is often found in HATs and ATP-dependent chromatin remodeling proteins.

Cellular reprograming conversion of a differentiated cell to an embryonic state.

Chromatin the molecular substance of chromosomes being a complex of genomic DNA and histone proteins.

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) a method for genome-wide mapping of the distribution of histone modifications and chromatin associated proteins that relies on immunoprecipitation with antibodies to modified histones or other chromatin proteins. The enriched DNA is sequenced to create genome-wide profiles.

Chromatin conformation capture (3C) a method for studying chromosomal 3D structure by proximity ligation. The assay relies on cross-linking chromatin with a fixing agent (usually formaldehyde), digestion of the DNA with a restriction enzyme and ligation of the fixed chromatin. In the resulting chimeric DNA template, regions that were close spatially are now closed linearly.

Chromatin modifying enzyme an enzyme either recognizing (reading) chromatin (*i.e.*, post-translationally modified histones and methylated genomic DNA), adding (writing) marks or removing (erasing) them.

Chromatosome the result of histone H1 binding to a nucleosome. It contains 166 bp of DNA, 147 of which wrapped around the histone core of the nucleosome.

Chromodomain a modular methyl-binding domain of 40-50 amino acids that is commonly found in chromatin remodeling proteins.

Chronic inflammation long-term inflammation lasting for prolonged periods of several months to years. Chronic inflammation plays a central role in most common non-communicable diseases, such as cancer, type 2 diabetes, asthma and Alzheimer's.

Constitutive heterochromatin a subtype of heterochromatin that is present at the highly repetitive DNA sequences found at the centromeres and telomeres of chromosomes, where it hinders transposable elements from becoming activated and thereby ensures genome stability and integrity.

CpG CG dinucleotides (the "p" indicates the phosphate linking the two nucleosides). Out of 16 possible dinucleotides, CpGs are the only ones that can be methylated symmetrically; *i.e.*, DNA methylation can be inherited only via CpGs to both daughter cells.

CpG island a genomic region of at least 200 bp showing a CG percentage of higher than 55%. However, typically CpG islands are 300-3,000 bp long.

CTCF a transcription factor with an 11-zinc finger DNA-binding domain that is involved in many cellular processes, such as transcriptional regulation, insulator activity and regulation of chromatin architecture.

Damage-associated molecular patterns (DAMPs) also known as alarmins, are molecules often released by stressed cells undergoing necrosis that act as

endogenous danger signals to promote and exacerbate inflammatory responses. Examples of non-protein DAMPs include cholesterol crystals and saturated fatty acids. DAMPs are associated with many inflammatory diseases, including arthritis, atherosclerosis, Crohn's disease and cancer.

DNA methylation the covalent addition of a methyl group to the C5 position of cytosine.

DNA methyltransferases (DNMTs) family of enzymes catalyzing the transfer of a methyl group to cytosines of genomic DNA.

Ectoderm the outermost layer of the three embryonic germ layers that gives rise to the epidermis, like skin, hair and eyes, and the nervous system.

Embryogenesis also called embryonic development, *i.e.*, the process by which the embryo forms and develops. In mammals, the term is used exclusively to the early stages of pre-natal development, whereas the terms fetus and fetal development describe later stages.

Embryonic stem (ES) cell a pluripotent stem cell that is derived from the inner cell mass of the early embryo. Pluripotent cells are capable of generating virtually all cell types of the organism.

Endoderm the innermost layer of the three embryonic germ layers that gives rise to the epithelia of the digestive and respiratory systems, such as liver, pancreas and lungs.

Enhancer a stretch of genomic sequence that (like a promoter) contains clusters of transcription factor binding sites that regulate a gene within the same TAD.

Enhancer RNAs (eRNAs) a class of relatively short ncRNA molecules (50-2000 nt) transcribed from enhancer regions. The expression of a given eRNA correlates with the activity of its corresponding enhancer.

Epidrugs small-molecule inhibitors that target chromatin modifying enzymes, such as DNMTs, HATs, HDACs, KMTs or KDMs.

Epigenetic clock a term used to measure age based on DNA methylation levels.

Epigenetic drift a divergence of the epigenome as a function of age due to stochastic changes in DNA methylation or stable histone modifications.

Epigenetic epidemiology the study of the relationship between epigenetic variants and disease phenotype in the population.

Epigenetic landscape a metaphor of cellular development, in which a landscape valleys and ridges illustrate how a pluripotent cell is guided to a well defined differentiated state, represented by a ball rolling down the landscape.

Epigenetic mediators genes whose products are the targets of the epigenetic modifiers.

Epigenetic memory a heritable change in gene expression that is induced by a previous developmental or environmental stimulus. It requires chromatin-based changes, such as DNA methylation, histone modifications or incorporation of variant histones.

Epigenetic modifier mostly identical to a chromatin modifying enzyme or the genes encoding for them.

Epigenetic modulator genes upstream of the epigenetic modifiers and mediators. The products of these genes mediate injury, inflammation and other forms of cellular stress.

Epigenetic programing the process leading to stable and long-lasting alterations of the epigenome based on specific covalent modifications of the DNA and histones.

Epigenetics the study of heritable changes in gene function that do not involve changes in the DNA sequence. Epigenetic mechanisms include the covalent modifications of DNA and histones.

Epigenome the complete set of epigenetic modifications across an individual's genome.

Epigenomics studies of the epigenome.

Epimutation heritable change in the chromatin state at a given position or region. In the context of cytosine methylation, epimutations are defined as changes in the methylation status of a single cytosine or of a region or cluster of cytosines. Epimutations do not necessarily imply changes in gene expression.

Erasers enzymes that remove histone modifications from chromatin, such as HDACs or KDMs.

Euchromatin light-staining, decondensed and transcriptionally accessible regions of the genome.

Facultative heterochromatin a dynamic form of heterochromatin that can change its density and activity in response to intra- and extra-cellular signals.

Gene body DNA sequence of a gene from the TSS to the end of the mRNA transcript.

Gene expression process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but can also be ncRNAs.

Gene regulatory networks represent units of interacting proteins that are functionally constrained by defined regulatory relationships. These interactions provide a structure and determine an output in the form of a pattern of gene expression. The networks are usually visualized by nodes (proteins) and edges (their interactions).

Genome the complete haploid DNA sequence of an organism comprising all coding genes and far larger non-coding regions. The genome of all 400 tissues and cell types of an individual is identical and constant over time (with the exception of cancer cells).

Genome-wide association study (GWAS) studies that aim to identify genetic loci (mostly SNPs) associated with an observable trait, disease or condition.

Genomic imprinting an epigenetic phenomenon in which expression of a gene is restricted to a single allele based on parental origin.

Genotype complete heritable genetic identity.

Healthspan the duration of disease-free physiological health within the lifespan of an individual. In humans, for instance, this corresponds to the period of high cognitive abilities, immune competence and peak physical condition.

Hematopoietic stem cells (HSCs) stem cells located in the bone marrow that can develop into all types of blood cells.

Heterochromatin dark-staining, condensed and gene-poor regions of the genome.

Hi-C a 3C-based method for genome-wide analysis of chromosome conformation. Hi-C involves massive parallel sequencing of chimeric 3C DNA templates and subsequent statistical analysis of the distribution of ligation junctions over two-dimensional contact matrices.

Histone acetyltransferases (HATs) enzymes that acetylate lysine amino acids on histone proteins by transferring an acetyl group from acetyl-CoA to form ε-N-acetyl-lysine.

Histone code an epigenetic code that is based on post-translational modifications of histone proteins. The histone modifications serve to recruit other proteins by specific recognition of the modified histone via specialized protein domains. The code comprises more than 130 post-translational modifications serving as an "alphabet" for the instructions, how the epigenome directs transcriptional regulation and stores information.

Histone deacetylases enzymes that remove acetyl groups from an ε-N-acetyl lysine amino acid on a histone.

Histone proteins lysine- and arginine-rich, positively charged proteins forming octamer cores on which genomic DNA is wrapped. They are the key protein components of chromatin.

Histone modification a covalent post-translational modification to histone proteins, which includes methylation, phosphorylation, acetylation, ubiquitylation, and sumoylation. They can impact gene expression by altering chromatin structure or recruiting histone modifiers.

Homodimer a complex formed by two identical proteins.

Hutchinson-Gilford progeria syndrome an extremely rare autosomal dominant genetic disorder (1 in 4 million), in which symptoms resembling aspects of aging are manifested at a very early age.

Immune checkpoint blockage immune checkpoints are regulators of the immune system being crucial for self-tolerance. Immune checkpoint blocking molecules, such as CTLA4 and PD1, are targets for cancer immunotherapy.

Implantation an early developmental stage at which the embryo adheres to the endometrium.

Imprinting a chromatin state defined by whether the gene or genetic locus is inherited from the male or the female germ line.

Inducible pluripotent stem (iPS) cells pluripotent stem cells that can be generated directly from terminally differentiated adult cells.

Inner cell mass a group of cells inside a mammalian blastocyst that gives rise to the embryo.

Integrative personal omics profiling an analysis method that combines genomic, transcriptomic, proteomic, metabolomic and autoantibody profiles from individuals in a longitudinal over a period of multiple months to years.

Interphase the resting phase between successive mitotic divisions of a cell.

Insulator a chromatin element that acts as a barrier against the influence of positive signals from enhancers or negative signals from silencers and heterochromatin.

Lamin-associated domains (LADs) heterochromatin interacting with lamina at the nuclear membrane.

Lineage descent in a line from a common progenitor cell.

Large organized chromatin K9-modifications (LOCKs): a subtype of LADs.

Long-term potentiation a persistent strengthening of synapses based on recent patterns of activity.

Lysine methyltransferases (KMTs) chromatin modifying enzymes that catalyze the transfer of one, two or three methyl groups to lysine residues of histone proteins.

Massive parallel sequencing high-throughput approach to DNA sequencing using the concept of massively parallel processing. It is also called next-generation sequencing (NGS) or deep sequencing.

Mesoderm the middle layer of the three embryonic germ layers that gives rise to the muscle, cartilage, bone, blood, connective tissue etc.

Metabolic syndrome a cluster of conditions, such as increased blood pressure, high blood sugar, excess body fat around the waist and abnormal cholesterol or triglyceride levels, that occur together, increasing the risk of heart disease, stroke and type 2 diabetes.

Metaphase the second stage of mitosis, between prophase and anaphase, during which the chromosomes become attached to the spindle fibers.

Metastable epiallele alleles that are variably expressed in genetically identical individuals due to epigenetic modifications established during early development and are particularly vulnerable to environmental influences.

Missing heritability the fact that genetic variations cannot account for all of the heritability of diseases, behaviors and other phenotypes.

Mitosis cell division that results in two daughter cells each having the same number and kind of chromosomes as the parent nucleus.

Multigenic disease a disorder that, in contrast to a monogenetic disease, is caused by changes in multiple genes. Most common diseases, such as type 2 diabetes, atherosclerosis and Alzheimer's, belong to this category.

Multipotent the ability of a cell to differentiate into multiple but a limited range of cell types. For example, cells of the embryonic germ layers and adult stem cells are multi-potent.

Non-coding RNA (ncRNA) an RNA molecule that is not translated into a protein.

Nucleosome a basic unit of DNA packaging in eukaryotes, consisting of 147 bp of genomic DNA wound around a histone octamer.

Odds ratio (OR) the mathematical expression of the relation between the presence or absence of a variant, for example, a SNP, and the presence or absence of a trait, for example, a disease, in the population.

Pathogen-associated molecular patterns (PAMPs) small molecular motifs derived from microbes, such a lipopolysaccharides. They are recognized by toll-like receptors and other pattern recognition receptors on the surface of cells of the innate immune system.

PHD finger a protein domain of 50-80 amino acids in length that recognizes trimethylated lysines. It is found in more than 100 human proteins, such as coactivators, Polycomb proteins, Trithorax group proteins and KDMs.

Personalized nutrition a conceptual analog to personalized medicine, where individuals are recommended to take certain food products based on nutrigenomics approaches.

Phenotype the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment.

Pioneer factors transcription factors that can directly bind to heterochromatin. They can have positive and negative effects on transcription and are important in recruiting other transcription factors and histone modification enzymes as well as controlling DNA methylation.

Plasticity the reversibility of epigenetic marks on DNA and proteins.

Pleiotropic genetic or epigenetic changes that affect multiple unrelated phenotypic traits.

Pluripotency the ability of a cell to differentiate into all three germ layers and to give rise to all fetal or adult cell types. For example, cells of the inner cell mass of blastocysts are pluripotent.

Poised promoter or enhancer an inactive status of a genomic region carrying histone markers, such as H3K4me1, that enable, after appropriate stimulation, the rapid reactivation of the region.

Polycomb repressive complexes large protein complexes limiting access of chromatin to transcription factors and therefore limit gene expression.

Post-translational modifications covalent modifications by which most proteins reach their full functional profile. Due to post-translational modifications the proteome is far more complex than the transcriptome and also varies a lot in response to extra- and intra-cellular signals.

Primordial germ cells the common origins of oocytes and spermatozoa; *i.e.*, they represent the ancestors of the germline. They occur in primary ectoderm already in the second week of embryogenesis.

Promoter stretches of genomic DNA for productive transcription initiation encompassing at least one TSS.

Proteome in analogy to the transcriptome, the complete set of all expressed proteins in a given tissue of cell type. The proteome depends on the transcriptome, but is not its 1:1 translation; *i.e.*, transcriptome analyses provide only a very rough description of the resulting proteome.

Quantitative trait loci (QTLs) genomic regions at which genetic variation is associated with molecular variation across individuals. For example, individuals with a particular single nucleotide variant have altered expression levels of a gene (eQTL), altered DNA methylation (meQTL; also known as mQTL) or altered chromatin state (chromQTL).

Readers nuclear proteins that recognize and bind chromatin through histone modification recognition domains.

RNA sequencing (RNA-seq) a method using massive parallel sequencing to reveal the presence and quantity of RNA in a biological sample at a given moment.

(Retro)transposon: A transposon (also called transposable element or "jumping DNA") is a DNA sequence that can change its position within a genome. When this transposition is mediated *via* an RNA intermediate, the term retrotransposon is used.

Senescence also called biological aging, the gradual deterioration of functional characteristics. It can refer either to cellular senescence or to senescence of the whole organism.

Signal transduction cascade the process by which a chemical or physical signal is transmitted through a cell membrane as a series of molecular events, such as protein phosphorylation catalyzed by protein kinases. Mostly, signal transduction cascades end in the activation of a transcription factor or a chromatin modifier.

Single nucleotide polymorphism (SNP) a substitution of a single nucleotide at a specific position in the genome, which is present to some appreciable degree within a population (for example, more than 1%).

Silencer genomic region that causes reduced expression of their target gene(s).

Sirtuins (SIRTs) a family of seven NAD^+-dependent HDACs that are structurally and mechanistically distinct from Zn^{2+}-dependent HDACs. Sirtuins influence a wide range of cellular processes such as aging, transcription, apoptosis, inflammation and stress resistance.

Stem cells can differentiate into other cell types and also divide in self-renewal to produce more of the same type of stem cells. There are embryonic stem cells, which are isolated from the inner cell mass of blastocysts, and adult stem cells, which are found in various tissues.

Super-enhancer a genomic region comprising multiple enhancers that is collectively bound by multiple transcription factor proteins driving transcription of genes involved in cell identity.

SWI/SNF complex protein complex that uses the energy of ATP hydrolysis to mobilize nucleosomes and remodel chromatin.

TET family α-ketoglutarate-dependent dioxygenases that catalyze the oxidation of 5mC to 5hmC and further products. Genes encoding these enzymes are frequently mutated in human cancers.

Tn5 transposase a member of the RNase superfamily of proteins, which includes retroviral integrases. Tn5 is utilized for fragmentation of the DNA, in next generation sequencing methods, such as ATAC-seq.

Topologically associated domain (TAD) large genomic region promoting regulatory interactions by forming higher-order chromatin structures separated by boundary regions.

Totipotent the ability of a cell to give rise to differentiated cells of all tissues, including embryonic and extra-embryonic tissues, in an organism. For example, a zygote is totipotent.

Trained immunity a memory system of innate immunity which is based on epigenetic programing.

Trait a distinguishing quality or characteristic belonging to a person.

Transcription factors proteins that sequence-specifically bind to genomic DNA. Our genome encodes approximately 1,600 transcription factors, referred to as *trans*-acting factors, since they are not encoded by the same genomic regions, which they are controlling. Accordingly, the process of transcriptional regulation by transcription factors is often called *trans*-activation.

Transcription factor binding motif a short (4-12 bp) DNA sequence pattern that summarizes the DNA sequence binding preference of a transcription factor. These motifs are usually represented as sequence logos based on position weight matrices.

Transcription factory discrete sites within the nucleus where transcription occurs. The factories contain Pol II, transcription factors, chromatin modifiers and modelers.

Transcription start sites (TSSs) Nucleotides within a promoter that are the first to be transcribed by Pol II into a particular RNA.

Transcriptome the complete set of all transcribed RNA molecules of a tissue or cell type. It significantly differs between tissues and depends on extra- and intracellular signals.

Transgenerational epigenetic inheritance transmission of epigenetic information that is passed on to gametes without alteration of the DNA sequence.

Trithorax group proteins large protein complexes maintaining the stable and heritable expression of certain genes throughout development.

Trophoblast the outer layer of the mammalian blastocyst that eventually develops to form part of the placenta.

Werner syndrome a rare (1 in 100,000), autosomal recessive disorder, which is characterized by the appearance of premature aging and caused by mutations in the *WRN* gene encoding for an ATP-dependent DNA helicase.

Western diet a dietary pattern characterized by high intakes of red meat, processed meat, pre-packaged foods, butter, fried foods, high-fat dairy products, eggs, refined grains, potatoes, corn and high-sugar drinks.

Writers enzymes that add modifications to chromatin, such as DNMTs, HATs and KTMs.

X chromosome inactivation a process in which one of the two X chromosomes is randomly inactivated in female mammalian cells early in development.

Zygote fertilized egg before cleavage occurs, *i.e.*, the one-cell stage embryo.