V25: the histone code

The DNA of eukaryotic organisms is packaged into chromatin, whose basic repeating unit is the **nucleosome**.

A nucleosome is formed by wrapping 147 base pairs of DNA twice around an octamer of four core histones, H2A, H2B, H3 and H4 (2 copies of each one).

X-ray structure of the nucleosome core particle consisting of core histones, and DNA. Top view.



Side view shows two windings of DNA and two histone layers



www.wikipedia.org

Basic principles of epigenetics: DNA methylation and histone modfications

The human genome contains ~20 000 genes that must be expressed in specific cells at precise times.

In cells, DNA is wrapped around clusters (octamers) of globular **histone** proteins to form **nucleosomes**.

These nucleosomes of DNA and histones are organized into **chromatin**, the building block of a chromosome.



Fig. 1. Carriers of epigenetic information: DNA and nucleosome. The left panel shows a DNA double helix that is methylated symmetrically on both strands (orange spheres) at its center CpG (PDB structure: 329d). DNA methylation is the only epigenetic mechanism that directly targets the DNA. The right panel shows a nucleosome spindle consisting of eight histone proteins (center), around which two loops of DNA are wound (PDB structure: 1KX5). The nucleosome is subject to covalent modifications of its histones and to the binding of non-histone proteins.

Rodenhiser, Mann, CMAJ 174, 341 (2006)

Bock, Lengauer, Bioinformatics 24, 1 (2008)

25. lecture SS 2018

Bioinformatics III

Epigenetic modifications



Reversible and site-specific **histone modifications** occur at multiple sites at the unstructured histone tails through **acetylation**, **methylation** and **phosphorylation**.

DNA methylation occurs at 5-position of cytosine residues within CpG pairs in a reaction catalyzed by DNA methyltransferases (DNMTs).

Post-translational modifications of histone tails

The disordered histone tails comprise 25-30% of the histone mass.

They extend from the compact histone multimer to provide a platform for various **posttranslational modifications** (PTMs).

These modifications affect the histones' ability to bind DNA and to other histones.

This, in turn affects **gene** expression.

Strahl BD and Allis CD, 2000. Nature 403:41-45



Mode of action of histone PTMs

Histone PTMs exert their effects via two main mechanisms.

(1) PTMs directly influence the overall structure of chromatin, either over short or long distances.

(2) PTMs regulate (either positively or negatively) the binding of effector molecules.

Bannister, Kouzarides, Cell Res. (2011) 21: 381-395.

PTMs of histone tails

Histone acetylation and phosphorylation effectively reduce the positive charge of



This potentially disrupts electrostatic interactions between histones and DNA.

This presumably leads to a less compact chromatin structure, thereby facilitating DNA access by protein machineries such as those involved in transcription.

Histone **methylation** mainly occurs on the side chains of lysines and arginines.

Unlike acetylation and phosphorylation, however, histone methylation does not alter the charge of the histone protein.

Bannister, Kouzarides, Cell Res. (2011) 21: 381–395. By Ybs.Umich - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=31240656

25. lecture SS 2018

histones.

H4 tail : conformational dynamics

All histone tails can influence chromatin compaction and accessibility, depending on

- salt concentration,
- construction of the nucleosome arrays, and
- the type of assembly process.

The **H4 tail** probably plays the most important role in inter-nucleosome interaction. Its middle part, the K¹⁶RHRK²⁰ segment forms a positively charged "**basic patch**".

On the H2A-H2B dimer, the glutamic acid and aspartic acid residues H2A E56, E61, E64, D90, E91, E92, and H2B E102 and E110 build up a negatively charged area, called the "**acidic patch**".

Due to the spatial proximity and the electrostatic attraction, stable salt bridges can be formed between these two parts from neighboring nucleosomes

Molecular dynamics simulations of H4-H2A/H2B-DNA system



Left: Structure of two nucleosomes from crystal packing Right: the model structure used in atomistic MD simulations. Water not shown.

(Green) DNA; (yellow) H3; (gray) H4; (pink) H2A; and (blue) H2B.

http://www.cell.com/biophysj/abstract/S0006-3495(16)31043-8

Bioinformatics III



Acetylation effects

Distribution of the distance between the H4 tail and the neighboring H2A-H2B dimer in the MD simulations. The center of mass of the backbone atoms of H4 tail residues 7–17 and H2A-H2B dimer are used for distance measurement. The middle part of the **AC H4 tail** is generally **further away** from the

adjacent H2A-H2B dimer.

The major population of WT is located at ~2.3–2.5 nm, indicating close contact between H4 tail middle part and the neighboring H2A-H2B dimer.

The distribution of AC is broader, ranging from 2.2 to 3.1 nm, and the multiple peaks refer to diverse conformation clusters. The center of the major peak of the AC population (AC-3 and AC-6) is shifted 0.2 nm to the right of the WT center.

http://www.cell.com/biophysj/abstract/S0006-3495(16)31043-8

The H4 tail is basically **disordered** due to active electrostatic interaction with outside partners.

Only some low-frequency 3_{10} -helix structures (formed by i+3 \rightarrow i hydrogen bonds) were found in the WT system.

In the AC system, the occupancy of 3_{10} -helix structures is twice as high.

Acetylation effects



Acetylation disrupts the interaction between the H4 tail and the acidic patch.

This gives the H4 tail the flexibility to form intratail hydrogen bonds.

The increasing intratail interaction helps to stabilize these structures.

Protein domains bind to modified histones



H3K4me3 – a mark associated with active transcription – is recognized by a PHD finger within the ING family of proteins (ING1-5). The ING proteins in turn recruit additional chromatin modifiers such as HATs and HDACs. Examples of proteins with domains that specifically bind to modified histones. There are more domain types recognizing lysine methylation than any other PTM.

H₃

H4

Bannister, Kouzarides Cell Res. (2011) 21: 381–395.

Chromodomain

Histone modification crosstalk



Histone PTMs can positively or negatively affect other PTMs.

A positive effect is indicated by an arrowhead and a negative effect is indicated by a flat head

The large number of histone PTMs enables tight control of chromatin structure. An extra level of complexity exists due to cross-talk between different modifications, which presumably helps to fine-tune the overall control.

Bannister, Kouzarides Cell Res. (2011) 21: 381–395.

Euchromatin vs. Heterochromatin structure

Eukaryotic genomes can be divided into two geographically distinct environments. (1) a relatively relaxed environment, containing most of the active genes and undergoing cyclical changes during the cell cycle. These **'open'** regions are referred to as **euchromatin**.

(2) other genomic regions, such as centromeres and telomeres, are relatively compact structures containing mostly inactive genes. These more **'compact'** regions are referred to as **heterochromatin**.

Both heterochromatin and euchromatin are enriched, and indeed also depleted, of certain **characteristic histone PTMs**.

However, there appears to be no simple rules governing the localization of PTMs. There is a high degree of overlap between different chromatin regions.

Nevertheless, there are regions of demarcation between heterochromatin and euchromatin. These **'boundary elements**' are bound by specific factors such as the "insulator" CTCF.



(Left) The scSet1 H3K4 methyltransferase binds to serine5 phosphorylated C-terminal domain (CTD) of RNAPII, the **initiating form** of polymerase situated at the TSS.

(Right) In contrast, the **scSet2** H3K36 methyltransferase binds to serine 2 phosphorrylated CTD of RNAPII, the transcriptional **elongating form** of polymerase.

Thus, the two enzymes are recruited to genes via interactions with distinct forms of RNAPII

 \rightarrow the location of the different forms of RNAPII define where the PTMs are placed

Epifactors database



DNA and two histone layers

The database EpiFactors stores detailed and curated information about 815 proteins and 69 complexes involved in epigenetic regulation.

http://epifactors.autosome.ru/protein_complexes

MSc thesis topic!

Database (Oxford). 2015; 2015: bav067.

Frequency of main annotation terms of epifactor proteins

Function	Count	Modification	Count
DNA modification	22	DNA methylation	7
RNA modification	30	DNA demethylation	12
Chromatin remodeling	101	DNA hydroxymethylation	5
Chromatin remodeling cofactor	41	RNA degradation	9
Histone chaperone	26	mRNA editing	10
Histone modification	15	Histone methylation	127
Histone modification cofactor	12	Histone acetylation	139
Histone modification read	90	Histone phosphorylation	55
Histone modification write	158	Histone ubiquitination	61
Histone modification write cofactor	95	Histone sumoylation	2
Histone modification erase	66	Histone citrullination	4
Histone modification erase cofactor	58	TF activator	18
Polycomb group (PcG) protein	29	TF repressor	27
Scaffold protein	12		
TF	53	Database (Oxford).	2015; 2015: bav067.

Most frequently occurring Pfam domains



Database (Oxford). 2015; 2015: bav067.

ENCODE



The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI).

The goal of ENCODE is to build a comprehensive parts list of **functional elements** in the human genome, including elements that act at the protein and RNA levels, and **regulatory elements** that control cells and circumstances in which a gene is active.

ENCODE: gene expression – TF binding sites



Correlative models between TF binding and RNA production in K562 cells.

(Left) output of the correlation models (x axis) compared to observed values (y axis).

(Right) The bar graphs show the most important TFs.

ENCODE consortium Nature 489, 57 (2012)

ENCODE: gene expression – histone marks



Correlative models between histone marks and RNA production in K562 cells.

ENCODE consortium Nature 489, 57 (2012)

ChromHMM

- ChromHMM is a software based on a multivariate **Hidden Markov Model** for learning and characterizing **chromatin states**.

- Input data can be multiple chromatin datasets such as ChIP-seq data of various histone modifications.

- The trained ChromHMM model can be used to systematically annotate a genome in one or more cell types.

Inputs to the initialization procedure: Let M be the number of marks in the model Let K be the number of states in the model Let C denote the set of chromosomes Let T_c denote the number of bins in chromosome c in CLet c_t denote the bin t on chromosome cLet v_{c_t} denote the observation vector at position t on the chromosome cLet $v_{c_t,m}$ denote the binary 0/1 observation for the m^{th} mark Let α be a smoothing constant with a default value of 0.02

Outputs of the parameter initialization procedure:

Let $p_{k,m}$ denote the emission probability in state k for mark m. Let $b_{i,j}$ denote the transition probability from state i to state j. Let a_i denote the probability of starting in state i.



Manolis Kellis MIT

Ernst, Kellis, Nature Methods 9, 215 (2012)

ChromHMM



Example of chromatin-state annotation tracks produced from ChromHMM and visualized in the UCSC genome browser.

Shown as example is the NFKB1 (subunit of nuclear factor kappa B, this TF controls more than 200 genes).

Active promoter, transcription transcription + elongation,

insulator before next gene MANBA

Ernst, Kellis, Nature Methods 9, 215 (2012)

ChromHMM



(left) which PTMs

contribute to which states.

```
1_Active_Promoter

2_Weak_Promoter

3_Poised_Promoter

4_Strong_Enhancer

5_Strong_Enhancer

6_Weak_Enhancer

7_Weak_Enhancer

8_Insulator

9_Txn_Transition

10_Txn_Elongation

11_Weak_Txn

12_Repressed

13_Heterochrom/lo

14_Repetitive/CNV

15_Repetitive/CNV
```



(right) Relative percentage of the genome represented by each chromatin state.

TSS, transcription start site; TES, transcript end site; GM12878 is a lymphoblastoid cell line.

> Ernst, Kellis, Nature Methods 9, 215 (2012)

Relate histone modifications to expression

(*i*) Is there a quantitative relationship between histone modifications levels and transcription?

(*ii*) Are there histone modifications that are more important than others to predict transcript levels?

(*iii*) Are there different requirements for different promoter types?

(*iv*) Are the relationships general?

The numbers of tags for each histone modification or variant, found in a region of 4,001 base pairs surrounding the transcription start sites of 14,802 RefSeq genes, was used as an estimation of the level of histone modifications.



Martin Vingron MPI Berlin

Relate histone modifications to expression

One-modification model (41 models)



Two-modifications model (820 models)



Three-modifications model (10,660 models)



Full model (1 model)

$$f(N'_{1, \dots, N'_{41}}) = a + b_1 N'_1 + \dots + b_{41} N'_{41}$$



Models are formulated as equations that **linearly** relate the levels of histone modifications to the measured expression value.

 N'_i : transformed levels of histone modification i

 $N'_i = \log(N_i + \alpha_i)$ (vector of length L)

- N_i : number of tags in each promoter
- y : expression values (vector of length L).

In the **one-modification models**, i can be any of the 39 modifications or two control IgG antibodies.

In the **two-modifications models**, i and j cover all combinations of two modifications without repetition.

In the **three-modifications models**, i, j, and k cover all combinations of three modifications without repetition.

The full model incorporates all 41 variables.

Karlic et al., PNAS 107, 2926 (2010)

25. lecture SS 2018

Bioinformatics III

Linear model for expression



Predicted expression values in CD4+ T-cells using the **full linear model** on the *x* axis and the measured expression values in CD4+ T-cells on the *y* axis.

The shades of blue indicate the density of points; the darker color, the more points.

Red line : linear fit between predicted and measured expression (y = 0.99x + 0.02), which are highly correlated (r = 0.77)

 \rightarrow a quantitative relationship exists between levels of histone modifications at the promoter and gene expression levels

(see slide 18 from ENCODE project)





models ranked according to prediction accuracy

Comparison of prediction accuracy between all possible one-modification, two-modifications, three-modifications models, and the full model for CD4+ T-cells.

Models are sorted by ascending prediction accuracy along the *x* axis. The best models using only a small subset of modifications almost reach the prediction accuracy of the full linear model.

The top one-modification (r_{max} = 0.72, H3K27ac), two-modifications (r_{max} = 0.74, H3K27ac + H4K20me1) and three-modifications models (r_{max} = 0.75, H3K27ac + H3K4me1 + H4K20me1) are very well correlated to expression.

Linear model for expression



Bar plot showing the frequency of appearance of different histone PTMs in best scoring three-modifications models (142 models) for CD4+ T-cells.

Best scoring models are defined as reaching at least 95% of prediction accuracy of the full linear model.

Not all modifications are equally important, possibly because of a high degree of redundancy.

Promoter methylation

Next, the authors separated the promoters into 2779 LCPs (**low CpG-content** promoters) and 7089 HCPs (**high CpG-content** promoters). Promoters with normalized CpG content > 0.4 are classified as HCP and the others as LCP.

This was motivated by the fact that the nucleosomes in HCPs are almost always decorated with **H3K4me3**, whereas nucleosomes in LCPs carry this modification only when they are expressed.

H3K4me3 is thought to be a mark of transcription initiation.

The authors reasoned that if these promoters are differently marked by histone modifications then the predictive power of histone modifications should also differ between these two groups of promoters.

Derive separate linear models for both groups.

Linear model for expression



(A) H4K20me1 and H3K27ac (and possibly H2BK5ac) are significantly over-

represented among the best scoring models for HCPs

(p-values hypergeometric test 9.97e-43, 2.58e-31, and 0.003)

(B) H3K4me3 and H3K79me1 are significantly overrepresented in the LCPs

(p-values of the hypergeometric test 9.71e-36 and 2.1e-34)

 \rightarrow different modifications are important for the prediction of expression of genes in these two groups.

Linear model for expression



Normalized cumulative tag counts in the region of -500 base pairs to 3,000 base pairs surrounding the transcription start site of RefSeq genes in CD4+ T-cells for the 5 important modifications identified by our analysis.

H3K4me3, H3K27ac, and H2BK5ac have the highest levels at the promoter, with the highest peaks around 100 base pairs downstream of the TSS.

H3K79me1 is enriched along the gene body, and H4K20me1 shows two distinct patterns: a peak close to the promoter at a similar position to H3K4me3 and H3K27ac, and a further enrichment across the gene body region.

Test whether model is transferable to other cell types

Apply trained CD4+ model to CD36+ and CD133+ cells.

The gene expression profiles of CD36+ and CD133+ cells are highly correlated to CD4+ T-cells (r = 0.79 and r = 0.82, respectively).

Thus, he prediction was restricted to genes with a fold change higher than five.

They found high correlation of predicted and measured expression values for both CD36+ (r = 0.75) and CD133+ (r = 0.63) cells.

This suggests that the relationship between histone modifications and gene expression is general and not dependent on the cellular context.

Roadmap: Integrative analysis of 111 epigenomes

How does the epigenomic landscape contribute to cellular circuitry, lineage specification, and the onset and progression of human disease?



Roadmap Epigenomics Consortium Nature 518, 317 (2015). 33

Mapped modifications

H3K4me3 - associated with promoter regions

H3K4me1 - associated with enhancer regions

H3K36me3 - associated with transcribed regions

H3K27me3 - associated with Polycomb repression

H3K9me3 - associated with heterochromatin regions

H3K27ac and H3K9ac, associated with increased activation of enhancer and promoter regions

DNase hypersensitivity denoting regions of accessible chromatin commonly associated with regulator binding

DNA methylation, typically associated with repressed regulatory regions or active gene transcripts Roadmap Epigenomics Consortium

25. lecture SS 2018

Bioinformatics III

Nature 518, 317 (2015).

Data sets available for 111 epigenomes

а	b	С	d	е		f	g	h	i	jk
Sample type	Cell type/ tissue group	EID F017	Epigenome name	H3K4me1	H3K4me3 H3K36me3 H3K27me3	H3K9me3 H3K27ac	DNase-Seq	DNA methyl	Gene expr.	Addtl marks Chrom. states
ry cultures	ES cell	E002 E008 E001 E015 E014 E016 E003 E024	ES-WA7 cells H9 cells ES-13 cells HUES6 cells HUES64 cells HUES64 cells H1 cells ES-UCSF4 cells							21
Primary	iPSC	E020 E019 E018 E021 E022	IPS-20b cells IPS-18 cells IPS-16 cells IPS DF 6.9 cells IPS DF 19.11 cells							
ES cell derived	ES-deriv.	E007 E009 E010 E013 E012 E011 E004 E005 E006	H1 derived neuronal progenitor cultured cells H9 derived neuronal progenitor cultured cells H9 derived neuron cultured cells HUES64 derived CD56 ⁺ mesoderm HUES64 derived CD56 ⁺ ectoderm HUES64 derived CD184 ⁺ endoderm H1 BMP4 derived mesendoderm H1 BMP4 derived trophoblast H1 derived mesenchymal stem cells					ľ		13 1 1 1 15 13
Primary cells	Blood & T cell	E062 E034 E045 E033 E044 E043 E049 E041 E040 E040 E037 E048 E038 E047	Primary mononuclear cells (from PB) Primary T cells from primary blood (from PB) Primary T cells effector/memory enriched (PB) Primary T cells from cord blood Primary T helper cells (from PB) Primary T helper cells (from PB) Primary T helper naive cells (from PB) Primary T helper envoluence (from PB) Primary T helper memory cells (from PB) Primary T helper memory cells (from PB) Primary T helper memory cells (from PB) Primary T Delper naive cells (from PB) Primary T helper naive cells (from PB) Primary T helper naive cells (from PB) Primary T CD8* memory cells (from PB)							
	HSC & B cell	E029 E031 E035 E051 E050 E036 E032 E046 E030	Primary monocytes (from PB) Primary B cells from cord blood Primary haematopoietic stem cells (HSCs) Primary HSCs G-CSF-mobilized male Primary HSCs short term culture Primary B cells (from PB) Primary natural killer cells (from PB) Primary neutrophils (from PB)		-					
'es	Mesench.	E026 E049 E025 E023	Bone marrow derived MSCs Mesenchymal stern cell deriv. chondrocyte Adipose-derived mesenchymal stern cells Mesenchymal stern cell derived adipocyte							, <u>, , , , , , , , , , , , , , , , , , </u>
릒	Myosat.	E052	Muscle satellite							1
Primary cu	Epithelial	E055 E056 E059 E061 E057 E058 E028 E027	Foreskin fibroblast Foreskin melanocyte Foreskin keratinocyte Foreskin keratinocyte Breast VHMEC mammary epithelial Breast myoepithelial							
	Neurosph.	E054 E053	Ganglion eminence derived neurospheres Cortex derived neurospheres							
	Thymus	E112 E093	Thymus Fefal thymus							
		1 5074				and the second se	1.00			1000

	Brain	E071 E074 E068 E069 E072 E067 E073 E070 E082 E081	Brain hippocampus middle Brain substantia nigra Brain anterior caudate Brain nirgulate gyrus Brain inferior temporal lobe Brain angular gyrus Brain dorsolateral prefrontal cortex Brain germinal matrix Fetal brain female Fetal brain male	
	Adipose	F063	Adipose nuclei	
iues	Muscle	E100 E108 E107 E089 E090	Soas muscle Skeletal muscle female Skeletal muscle male Fetal muscle trunk Fetal muscle leg	
	Heart	E083 E104 E095 E105 E065	Fetal heart Right atrium Left ventricle Right ventricle Aorta	
ary tis	Smooth muscle	E078 E076 E103	Duodenum smooth muscle Colon smooth muscle Rectal smooth muscle	
Frimar	Digestive	E111 E092 E085 E084 E109 E106 E075 E101 E102 E110 E077 E079	Stomach smooth muscle Fetal stomach Fetal intestine small Fetal intestine large Small intestine Sigmoid colon Colonic mucosa Rectal mucosa donor 29 Rectal mucosa donor 31 Stomach mucosa Duodenum mucosa Oesophagus	
	Other	E094 E099 E086 E088 E097 E087 E080 E091 E091 E096 E098 E096 E113	Placenta amnion Fetal kidney Fetal lung Ovary Pancreatic islets Fetal adrenal gland Placenta Liver Pancreas Lung Soleen	
	ENCODE 2012	E113 E114 E115 E116 E117 E118 E120 E121 E122 E123 E124 E125 E126 E127 E128 E129	AS49 EYOH 0.02pct lung carcinoma Drd41 T cell leukaemia GM12878 lymphoblastoid HeLa-S3 cervical carcinoma HepG2 hepatocellular carcinoma HMEC mammary epithelial HSMM skeletal muscle myoblasts HSMM-derived skeletal muscle myotubes HUVEC umbilical vein endothelial K562 leukaemia Monocytes-CD14 ⁺ RO01746 NH-A astrocyte NHDF-ad adult dermal fibroblast NHEK-epidermal keratinocyte NHLF lung fibroblast Osteoblast	
dai lipie type	Prim. culture ES cell derived Primary cell Prim. tissue Cell line	Signal-to-noise	$ \begin{array}{c} 100\% \\ 50\% \\ 0\% \\ 0\% \\ 0\% \\ 0\% \\ 0\% \\ 0\% $	est-quality epigenomes ($n = 60$) omHMM model trained + applied naining epigenomes ($n = 67$) omHMM model only applied)

Roadmap Epigenomics Consortium Nature 518, 317 (2015). 35

Integrative analysis of 111 epigenomes





Chromatin state annotations across 127 reference epigenomes (rows) in a ~3.5-Mb region on chromosome 9.

Promoters are

primarily constitutive (i.e. unchanged) (red vertical lines), while enhancers are highly dynamic (dispersed yellow regions).

Roadmap Epigenomics Consortium Nature 518, 317 (2015). 36

Signal tracks for IMR90

Cł	nrom. states	FALIORE							
	RetSeq genes	FAM205B	ATP885P STITTNPR2	HECK HIM HNF38	MAN MELK MAN PAXS	GRHP	R	SHBRANCE	ALDHIBI
h	RINA-seq								
	H3K36me3						A.M		
	H4K20me1				A				
	H3K79me2								
	H3K79me1			-	-		A		
	H3K9me1	-							
S	DNase	4.44				-	-	- Ander	
st	DGF	4.4	at the table	makes-			-		
ð	input					and move			
q	H3K4me3		La shika				· · · · · ·		
0	H3K9ac		an all has	and and a				-	
ō	H3K56ac			the dealers				-	
÷	H2A.Z	-		And the sec			1 th al +	Inch	
D	H2AK9ac	-	and all an	han all					
č	H2BK5ac			-				-	
	H3K4me2		and a state of the state	- frate-					
-	H3K18ac		and the set is an	In Adda					
ta	H3K4me1	- Bank		Sec. and a				-	
Ð	H3K27ac		and the state of the second se	In Adam.					
T	H4K5ac			A Adam				-	. A.
8	H4K8ac			State				-	
ř	H3K4ac			the second				-	
1	H3K23ac	-	4	the Areas				-	-
\leq	H2AK5ac			6.4					
	H4K91ac		a de miter	A					
	H2BK120ac		hand -	A Ann					
	H2BK12ac			A A A A					
	H2BK15ac		Al-	-					
	H2BK20ac					*			
	H3K27me3						- And		A
	H3K9me3	-	And	mather the law	and a stand of the second	And an one			da amanda an
	WGBS	Ind addition of the			In the second states of	AR	I Marcal	and the second s	Service of the servic
	Hi-C	10000000	((()))))))))))))))))))))))))))))))))))		000000000000000000000000000000000000000	000000000	×		
						and the second s			

Signal tracks for IMR90 (fetal lung fibroblast) showing RNA-seq, a total of 28 histone modification marks, whole-genome bisulfite DNA methylation, DNA accessibility, digital genomic footprints (DGF), input DNA and chromatin conformation information.

> Roadmap Epigenomics Consortium Nature 518, 317 (2015). 37

Bioinformatics III

Training of recurring 15-states chromatin model

a	1			b		C	
С	hromatin state	Abbreviation	emissions	Cov.	Annotation overlap	Expr.	Repr.
1	Active TSS	TssA		0.7%			
2	Flanking active TSS	TssAFInk		0.5%			
3	Transcr. at gene 5' and 3'	TxFlnk		0.1%			
4	Strong transcription	Tx		3.6%			
5	Weak transcription	TxWk		11.6%			
6	Genic enhancers	EnhG	1.5	0.4%		100 E	
7	'Enhancers	Enh		2.8%			
8	ZNF genes + repeats	ZNF/Rpts		0.2%			
9	Heterochromatin	Het		2.6%			
10	Bivalent/poised TSS	TssBiv		0.1%			
11	Flanking bivalent TSS/Enh	BivFlnk		0.1%			
12	Bivalent enhancer	EnhBiv		0.1%			
13	Repressed Polycomb	ReprPC		1.2%			
14	Weak repressed Polycomb	ReprPCWk		8.3%			
15	Quiescent/low	Quies		67.8%			
		Relative enrichment	H3K4me3 H3K4me1 H3K36me3 H3K9me3	Genome% (average)	Cpus Exons Genes Introns TES TSS (2 Kb) TSS TSS (2 Kb)	Expr. genes Expr. TES Expr. TSS	Repr. genes Repr. TES Repr. TSS

Roadmap Epigenomics Consortium Nature 518, 317 (2015). 38

Consisteny of chromatin states across genomic positions



Roadmap Epigenomics Consortium Nature 518, 317 (2015).

H3K4me1-associated states (including TxFlnk, EnhG, EnhBiv and Enh) are the most **tissue specific**, with 90% of instances present in at most 5–10 epigenomes, followed by bivalent promoters (TssBiv) and repressed states (ReprPC, Het).

In contrast, active promoters (TssA) and transcribed states (Tx, TxWk) were highly **constitutive**, with 90% of regions marked in as many as 60–75 epigenomes.

Quiescent regions were the most constitutive, with 90% consistently marked in most of the 127 epigenomes.

25. lecture SS 2018

Relative switching between states



More frequent switching found between active states and repressed states.

This is consistent with activation and repression of regulatory regions.

Summary

Combinations of histone modification marks are highly informative of the methylation and accessibility levels of different genomic regions, while the converse is not always true.

Genomic regions vary greatly in their association with active marks.

Approximately 5% of each epigenome is marked by enhancer or promoter signatures on average, which show increased association with expressed genes, and increased evolutionary conservation.

Two-thirds of each reference epigenome on average are quiescent, and enriched in gene-poor and stably repressed regions.

Even though promoter and transcription associated marks are less dynamic than enhancer marks, each mark recovers biologically meaningful cell-type groupings when evaluated in relevant chromatin states.