Introduction to (epi)Genomic data integration

Shamith Samarajiwa

CRUK Bioinformatics Summer School

July 2019

MRC Medical Research Council

UNIVERSITY OF CAMBRIDGE

(epi)Genomic Data Integration

- Combining different biological layers to understand phenotypes
- Computational or Statistical approaches

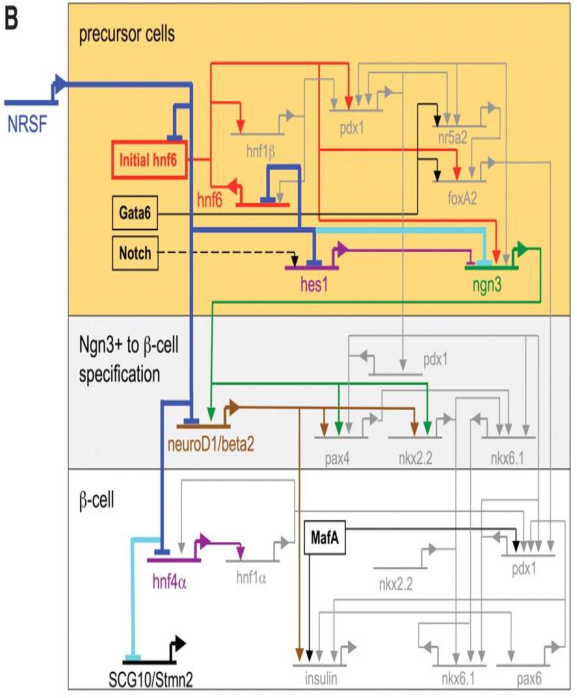
Some examples:

1. TF direct targets: TF ChIP-seq + RNA-seq
2. Enhancer-Promoter Interactions
    - Chromosomal Conformation Capture (Hi-C) + Histone mark ChIP-seq + RNA-seq
3. Impact of chromatin accessibility on transcription
    - ATAC-seq + RNA-seq +DRIP-seq
4. Translational regulation
    - iCLIP + Ribo-seq +RNA-seq
5. Epigenetic silencing: DNA methylation + RNA-seq
6. Regulatory elements: multiple Histone mark ChIP-seq

Identifying direct targets of TFs

Network Biology: reverse engineer regulatory networks by integrating TF binding and gene expression

- Not all TF binding sites are transcriptionally active. The collection of TF binding sites are called the Cistrome and the collection of transcriptionally active targets (regulons) of a TF is its Regulome.
- Regulomes can be used to "explain" the phenotype under consideration and understand aspects of biological systems.
- Regulomes in combination with pathway and network modelling approaches can then be used reverse engineer the networks underlying phenotypes.
- These networks provide information on connectivity, information flow, and signaling, regulatory, metabolic and other interactions between cellular components and processes.



TF Direct Target detection

Rcade (R-based analysis of ChIP-seq And Differential Expression)

- Rcade is a Bioconductor package we (Cairns *et al.*) developed that utilizes **Bayesian** methods to integrates TF binding ChIP-seq, with transcriptomic Differential Expression.
- The method is “read-based” and independent of peak-calling, thus avoids problems associated with peak-calling methods.

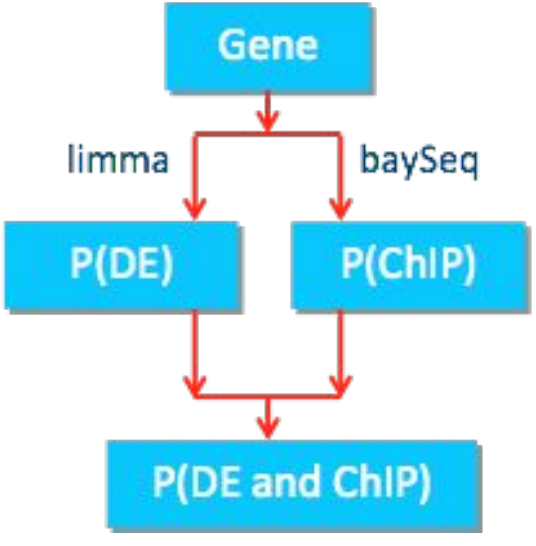
- A key application of Rcade is in inferring the direct targets of a transcription factor (TF).
- These targets should exhibit TF binding activity, and their expression levels should change in response to a perturbation of the TF.

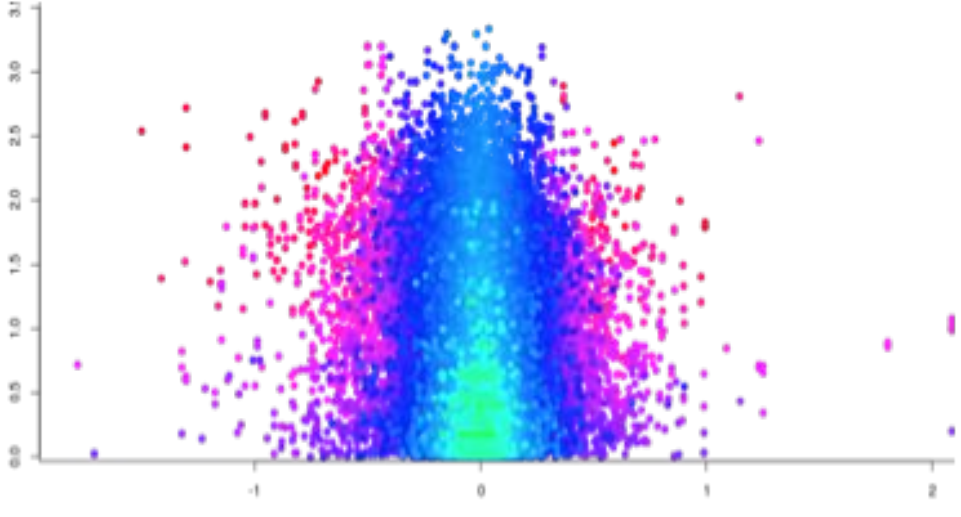
Statistical approaches to data integration

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Rcade

- **Rcade: R based analysis of ChIPseq And Differential Expression**
- Bayesian approach used to integrate ChIP-seq with differential expression to identify direct transcriptional targets of transcription factors.





Rcade

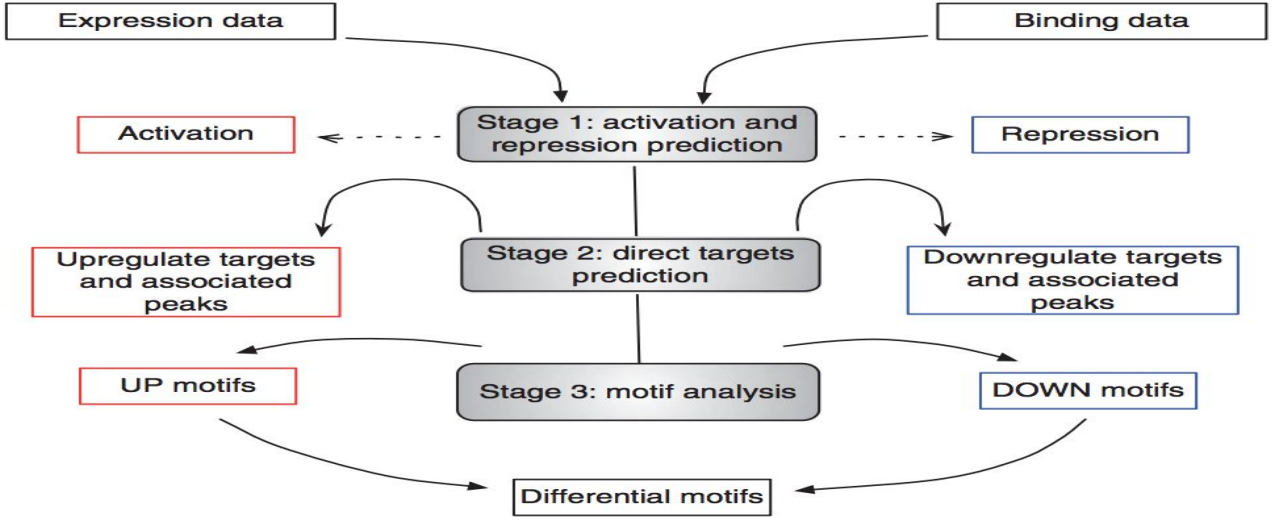
- Rcade integrates posterior probabilities of binding (determined via the baySeq package) with those of differential expression (determined via the limma package).

$$B = \log(\frac{PP}{1 - PP})$$

- Rcade uses a fully Bayesian modelling approach. In particular, it uses log-odds values (a measure of probability), or B-values, in both its input and output. The log-odds value is related to the posterior probability (PP) of an event, as per the formula above.
- Priors need to be defined.
- A number of output files are generated by Rcade. Usually, the file of interest is "DEandChIP.csv", which contains a list of genes most likely to have both DE and ChIP signals ranked by their B-value.
- More on Rcade @ the practical!

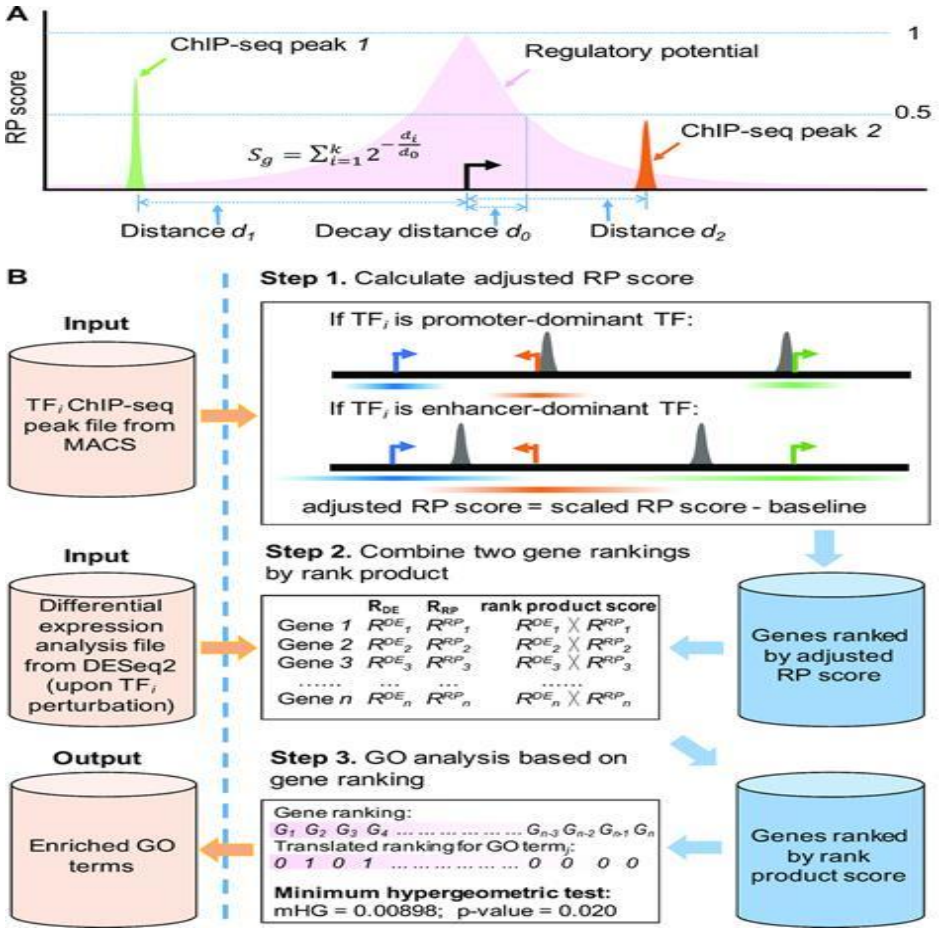
Beta

- Three main functionalities:
  - to predict whether a factor has activating or repressive function
  - to *infer* the factor's target genes
  - to identify the binding motif of the factor and its collaborators



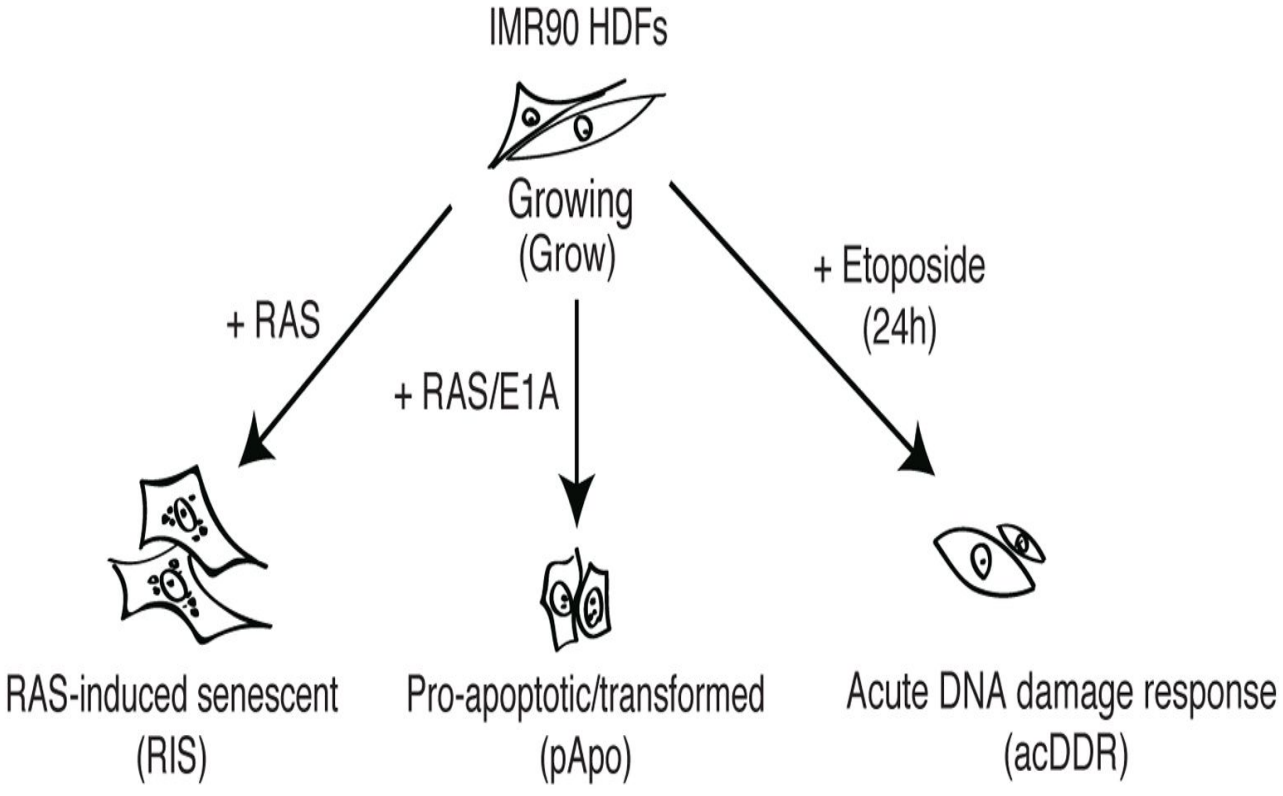
Wang, 2013 Nat Protoc. 2013

Cistrome GO

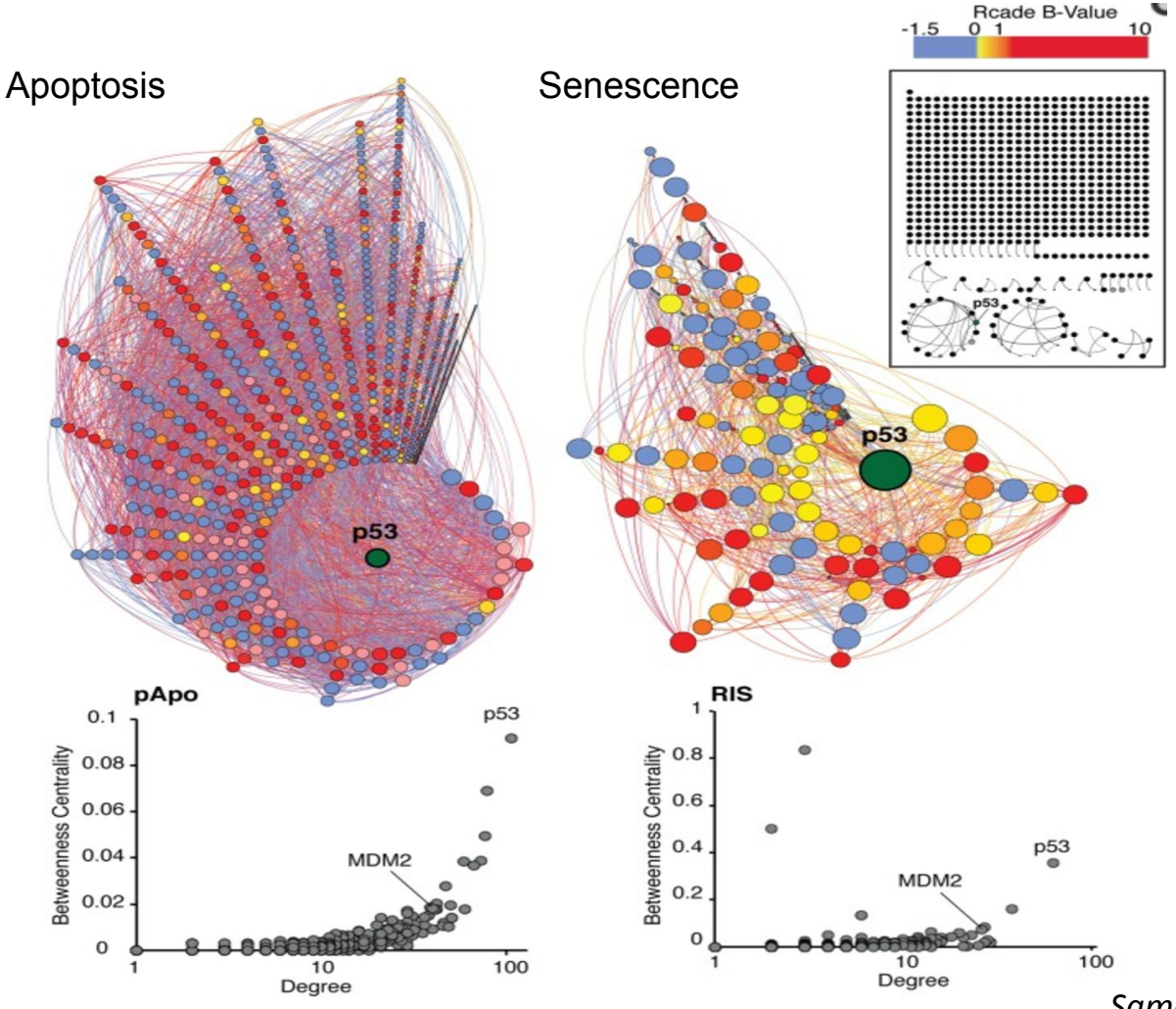


Li et al., Nuc Acid Res. 2019

TP53 direct targets



Identifying TP53 direct targets

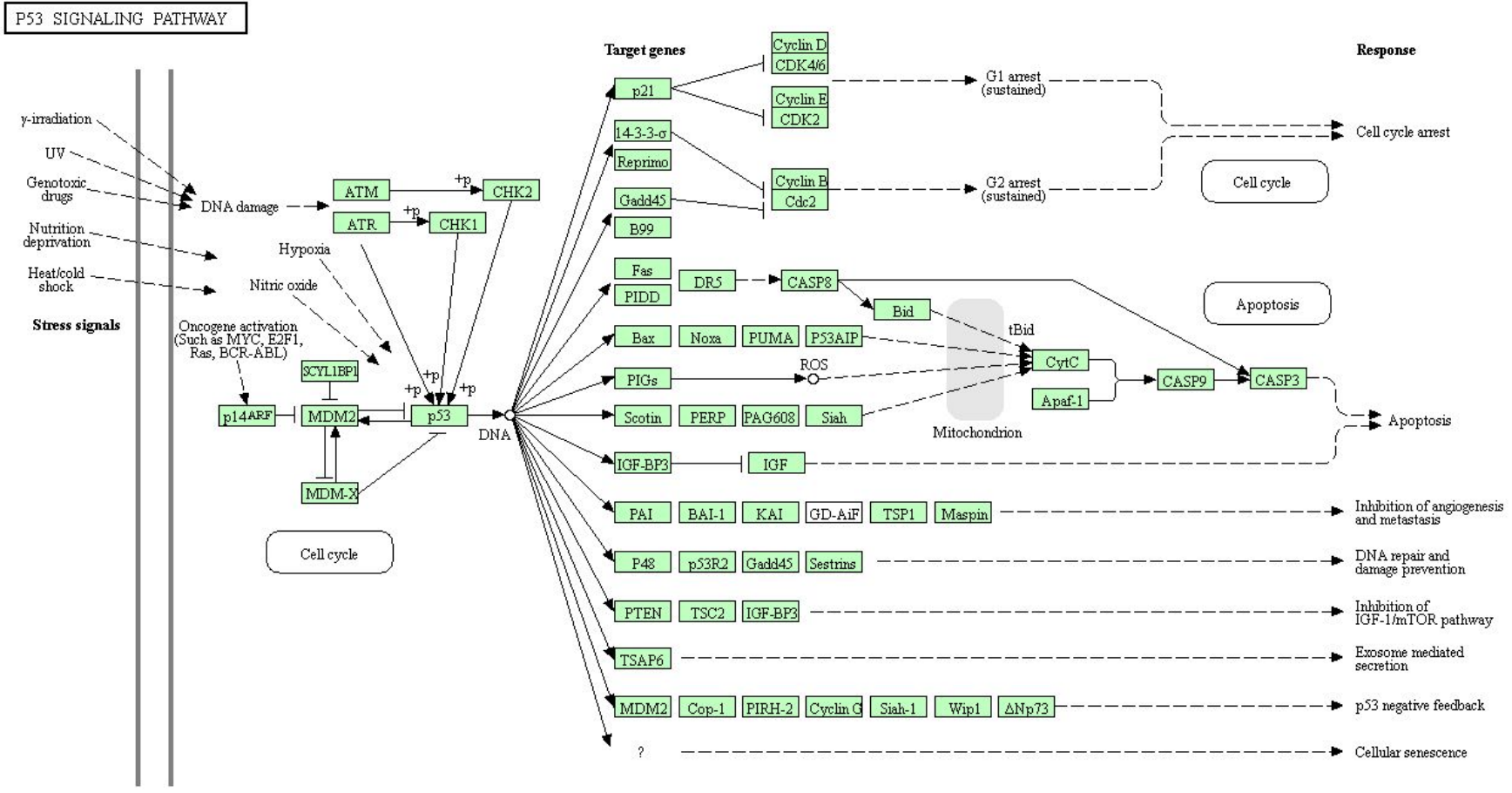


TF Direct Targets:

- TF binding
- Differential gene activation or inhibition
- Presence of a TF specific motif
- Literature evidence
- Enhancer - Promoter interactions

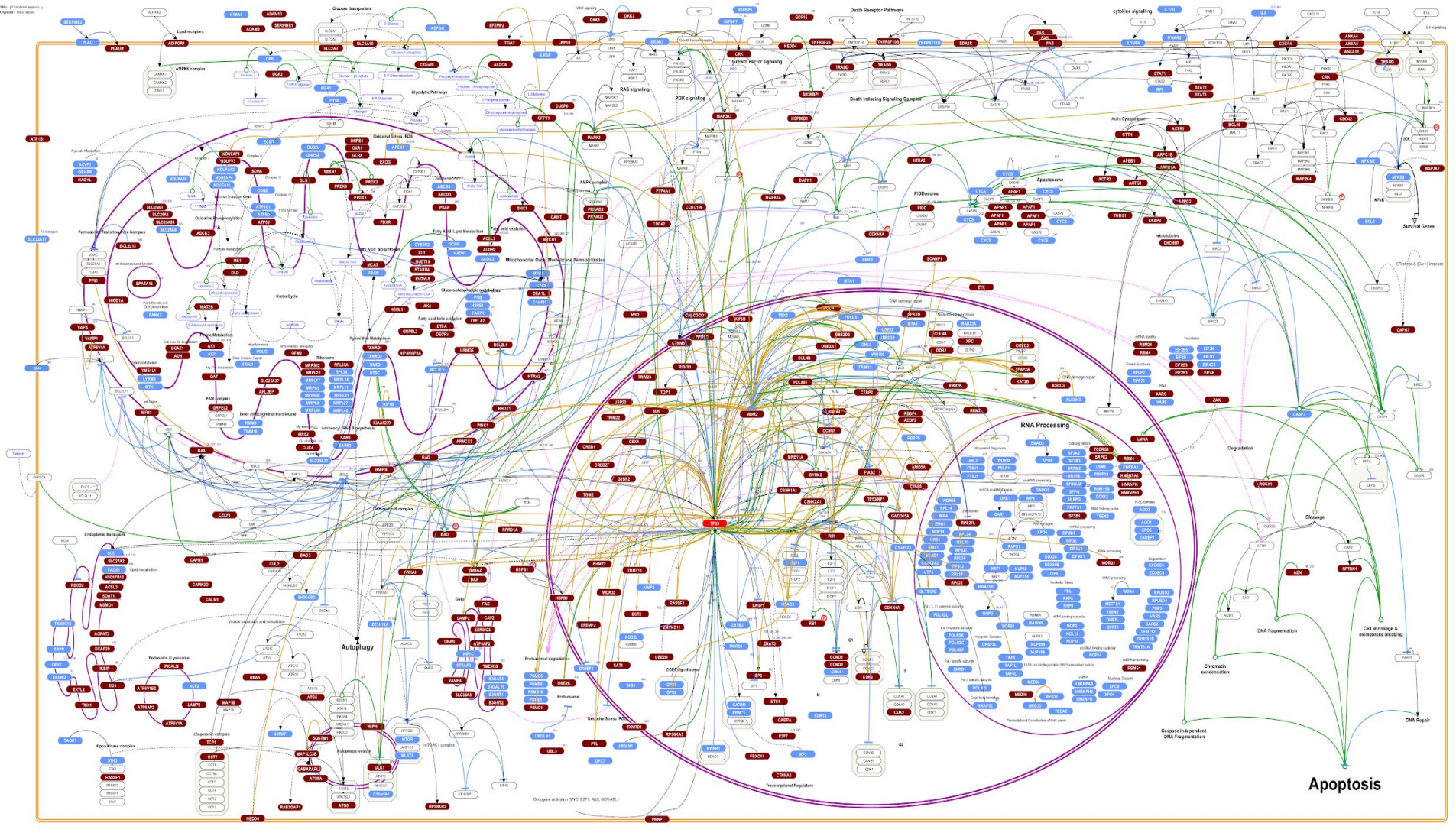
Samarajiwa et al., PLoS Genetics 2015

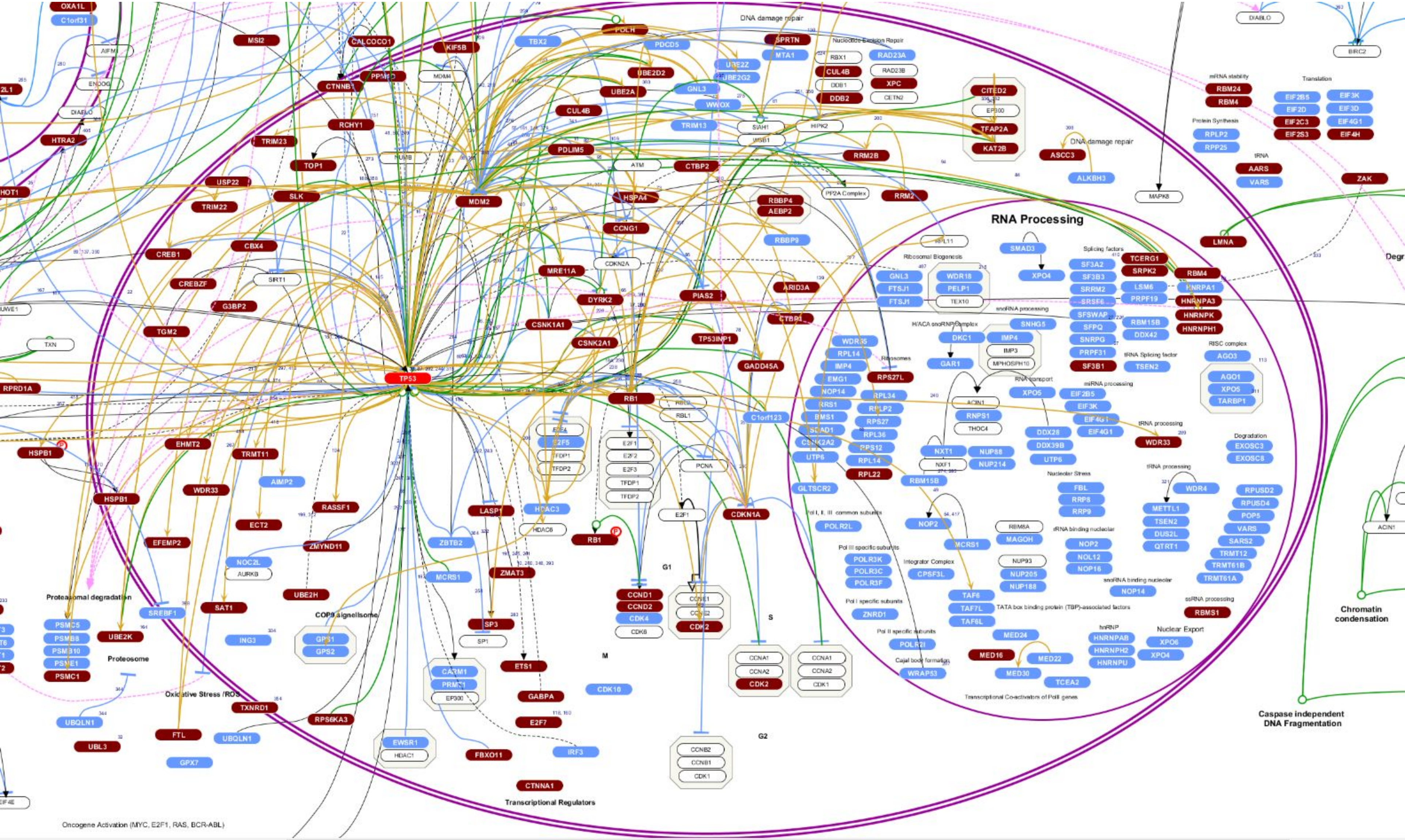
KEGG: p53 signalling pathway



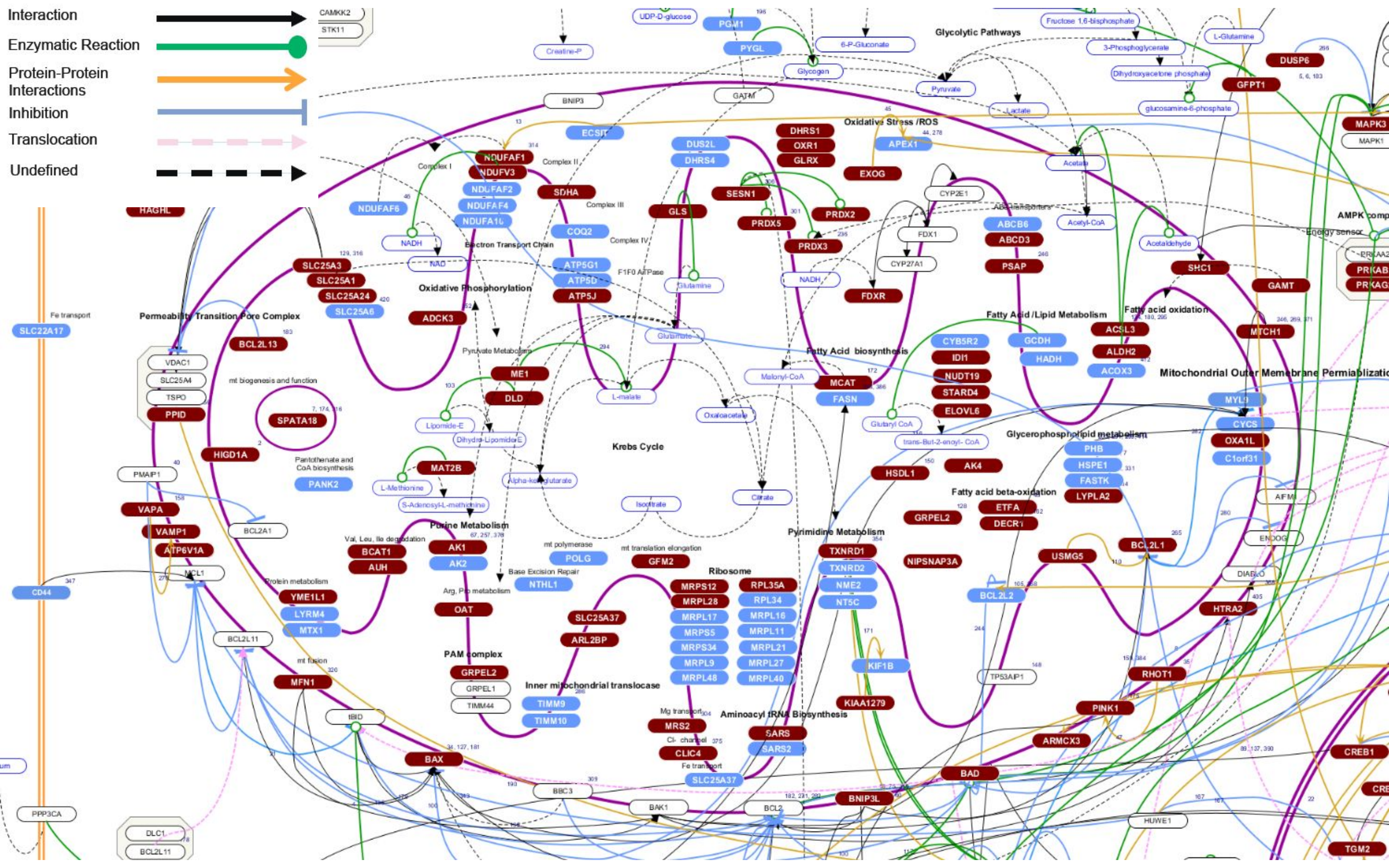
04115 6/21/16
(c) Kanehisa Laboratories

Molecular Cartography of TF regulation



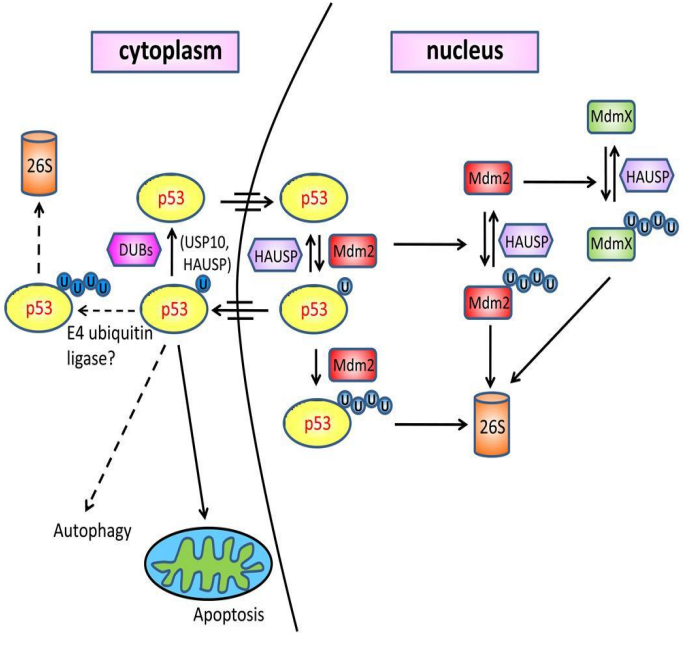


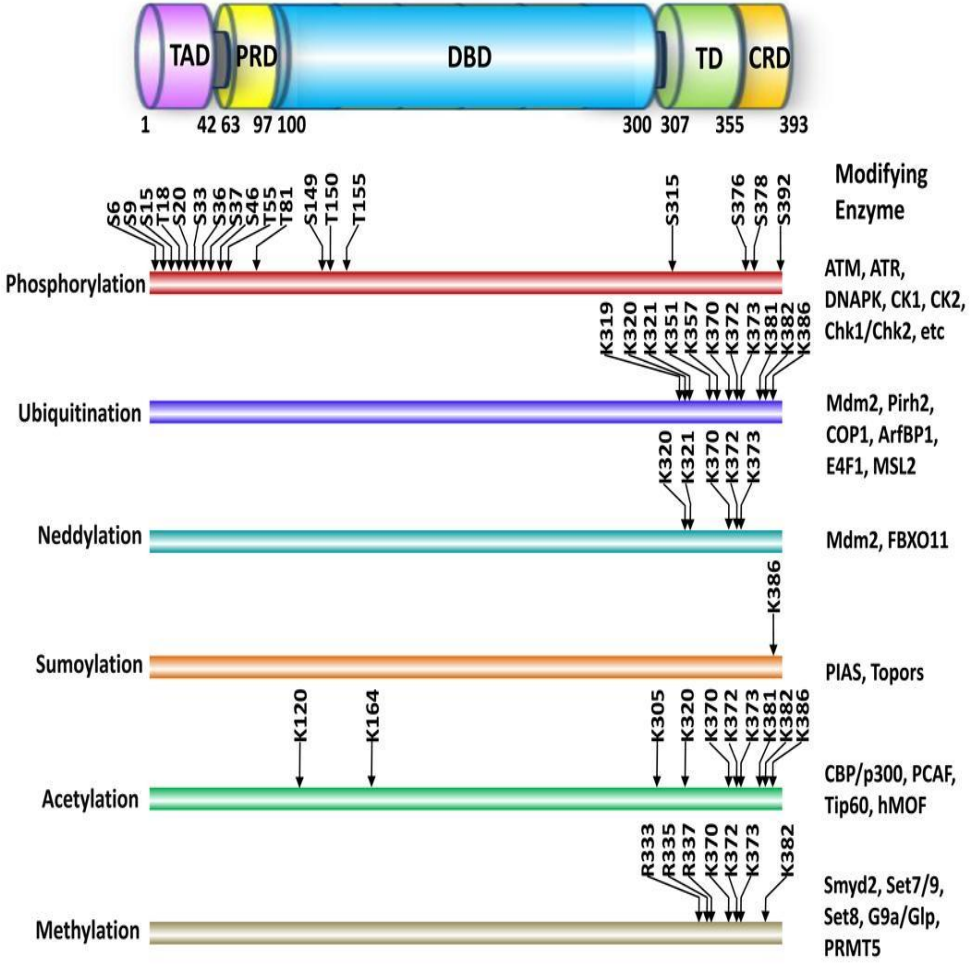
DNA damage repair
Nucleotide Excision Repair
mRNA stability
Translation
Protein Synthesis
tRNA
DNA damage repair
RNA Processing
Ribosomal Biogenesis
Splicing factors
snoRNA processing
H/ACA snoRNP Complex
RISC complex
tRNA Splicing factor
Ribosomes
RNA transport
miRNA processing
tRNA processing
Degradation
Nucleolar Stress
rRNA binding nucleolar
Pol I, II, III common subunits
Pol III specific subunits
Integrator Complex
snoRNA binding nucleolar
ssRNA processing
Pol I specific subunits
TATA box binding protein (TBP)-associated factors
hnRNP
Nuclear Export
Pol II specific subunits
Cajal body formation
Transcriptional Co-activators of PolII genes
Chromatin condensation
Caspase independent DNA Fragmentation
Proteasomal degradation
Proteosome
COP9 signellsome
Oxidative Stress /ROS
Transcriptional Regulators
Oncogene Activation (MYC, E2F1, RAS, BCR-ABL)
G1
S
G2
M
TP53
MDM2
RB1
OXA1L
C1orf31
MSI2
CALCOCO1
KIF5B
POLH
TBX2
PDCD5
SPRTN
MTA1
RBX1
CUL4B
DDB1
DDB2
RAD23A
RAD23B
XPC
CETN2
DIABLO
BIRC2
AIFM
ENDOG
CTNNB1
PPM1D
MDM4
UBE2D2
UBE2A
UBE2Z
UBE2G2
GNL3
WWOX
TRIM13
SIAH1
WSB1
HIPK2
CITED2
EP300
TFAP2A
KAT2B
RBM24
RBM4
EIF2B5
EIF2D
EIF2C3
EIF2S3
EIF3K
EIF3D
EIF4G1
EIF4H
RPLP2
RPP25
AARS
VARS
ZAK
HTRA2
TRIM23
RCHY1
CUL4B
NUMB
TOP1
PDLIM5
ATM
CTBP2
RRM2B
ASCC3
ALKHB3
MAPK8
HOT1
USP22
TRIM22
SLK
HSPA4
PP2A Complex
RRM2
RBBP4
AEBP2
RBBP9
CCNG1
CREB1
CBX4
CREBZF
SIRT1
CDKN2A
MRE11A
DYRK2
PIAS2
ARID3A
G3BP2
UWE1
TGM2
CSNK1A1
CSNK2A1
TP53INP1
CTBP1
GADD45A
TXN
RPRD1A
RBL2
RBL1
C1orf123
HSPB1
EHMT2
TRMT11
WDR33
AIMP2
RASSF1
ECT2
EFEMP2
ZMYND11
E2F4
E2F5
TFDP1
TFDP2
E2F1
E2F2
E2F3
PCNA
LASP1
HDAC3
HDAC6
ZBTB2
NOC2L
AURKB
MCRS1
ZMAT3
UBE2H
SAT1
SREBF1
PSMC5
PSMB8
PSMB10
PSME1
PSMC1
UBE2K
ING3
GPS1
GPS2
SP3
SP1
CDKN1A
CCND1
CCND2
CDK4
CDK6
CCNE1
CCNE2
CDK2
CCNA1
CCNA2
CDK1
CCNB2
CCNB1
CARM1
PRMT1
ETS1
GABPA
E2F7
CDK10
UBQLN1
TXNRD1
RPS6KA3
FTL
UBL3
GPX7
EWSR1
HDAC1
FBXO11
IRF3
CTNNA1
EIF4E
RPL11
SMAD3
XPO4
GNL3
FTSJ1
WDR18
PELP1
TEX10
SF3A2
SF3B3
SRRM2
SRSF6
SFSWAP
SFPQ
SNRPG
PRPF31
SF3B1
TCERG1
SRPK2
LSM6
PRPF19
RBM15B
DDX42
TSEN2
RBM4
HNRNPA1
HNRNPA3
HNRNPK
HNRNPH1
LMNA
SNHG5
DKC1
IMP4
IMP3
MPHOSPH10
GAR1
WDR55
RPL14
EMG1
NOP14
RRS1
BMS1
SDAD1
CSNK2A2
UTP6
RPS27L
RPL34
RPLP2
RPS27
RPL36
RPS12
RPL22
XPO5
ACIN1
RNPS1
THOC4
EIF2B5
EIF3K
EIF4G1
AGO3
AGO1
TARBP1
WDR33
EXOSC3
EXOSC8
NXT1
NXF1
NUP88
NUP214
DDX28
DDX39B
RBM15B
GLTSCR2
FBL
RRP8
RRP9
WDR4
METTL1
DUS2L
QTRT1
RPUSD2
RPUSD4
POP5
SARS2
TRMT12
TRMT61B
TRMT61A
NOP2
POLR2L
RBM8A
MAGOH
NOL12
NOP16
POLR3K
POLR3C
POLR3F
CPSF3L
NUP93
NUP205
NUP188
NOP14
RBMS1
TAF6
TAF7L
TAF6L
ZNRD1
MED24
HNRNPAB
HNRNPH2
HNRNPU
XPO6
POLR2I
MED16
MED22
MED30
TCEA2
WRAP53



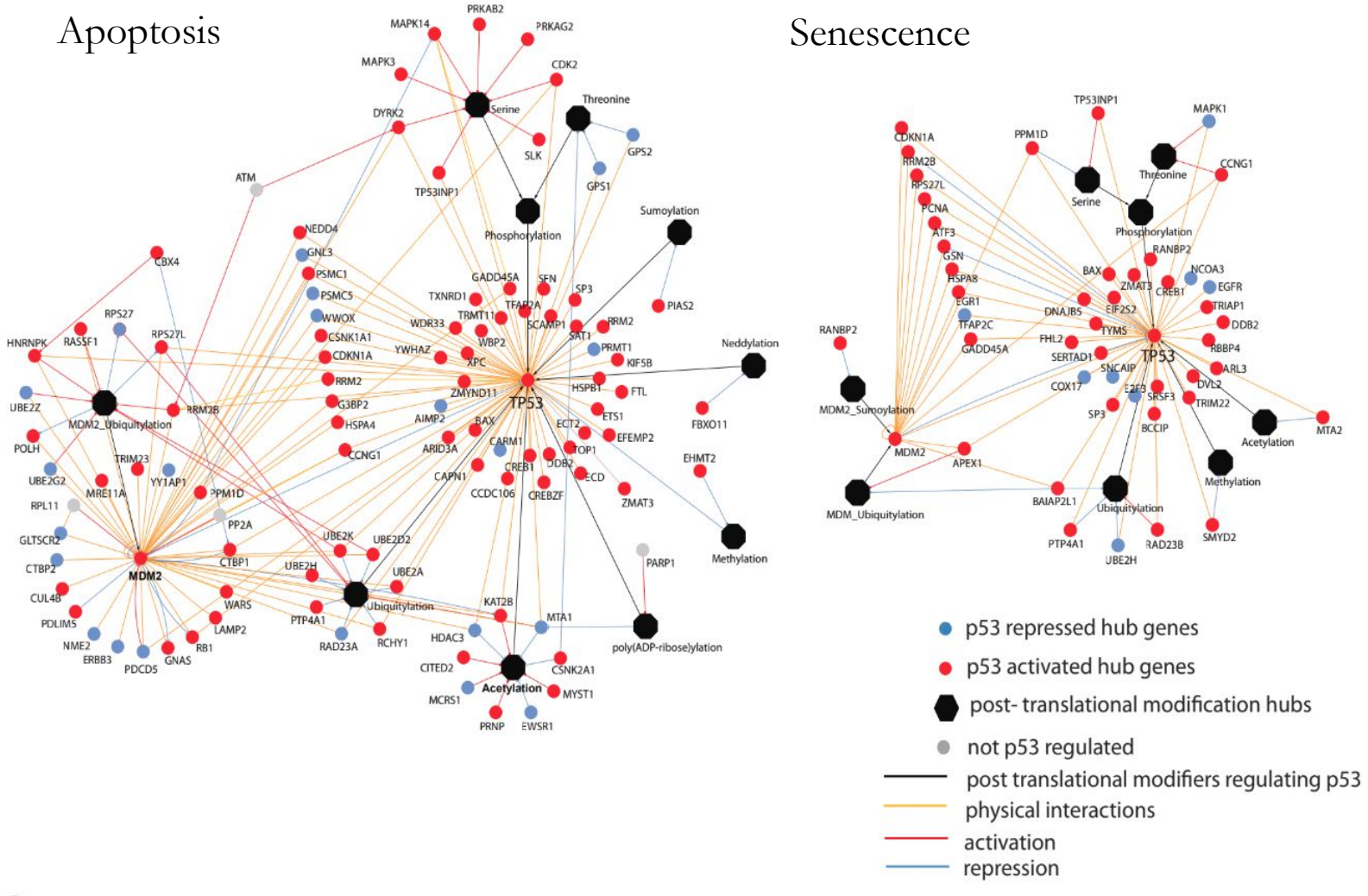
Interaction
Enzymatic Reaction
Protein-Protein Interactions
Inhibition
Translocation
Undefined
CAMKK2
STK11
Creatine-P
UDP-D-glucose
PGM1
PYGL
Glycogen
6-P-Gluconate
Glycolytic Pathways
Fructose 1,6-bisphosphate
3-Phosphoglycerate
L-Glutamine
Dihydroxyacetone phosphate
Pyruvate
Lactate
glucosamine-6-phosphate
DUSP6
GFPT1
MAPK3
MAPK1
BNIP3
GATM
ECSIT
DUS2L
DHRS4
DHRS1
OXR1
GLRX
Oxidative Stress /ROS
APEX1
EXOG
Acetate
NDUFAF1
NDUFV3
Complex I
Complex II
SDHA
NDUFAF2
NDUFAF4
NDUFA10
NDUFAF6
Complex III
COQ2
Complex IV
GLS
SESN1
PRDX5
PRDX2
PRDX3
CYP2E1
FDX1
CYP27A1
ABC-transporters
ABCB6
ABCD3
PSAP
Acetyl-CoA
Acetaldehyde
AMPK complex
Energy sensor
PRKAA2
PRKAB
PRKAG
NADH
NAD
Electron Transport Chain
SLC25A3
SLC25A1
SLC25A24
SLC25A6
ATP5G1
ATP5D
ATP5J
F1F0 ATPase
Glutamine
NADH
FDXR
SHC1
GAMT
Oxidative Phosphorylation
ADCK3
Fe transport
SLC22A17
Permeability Transition Pore Complex
BCL2L13
Glutamate
Fatty Acid /Lipid Metabolism
GCDH
HADH
Fatty acid oxidation
ACSL3
ALDH2
ACOX3
MTCH1
CYB5R2
IDI1
NUDT19
STARD4
ELOVL6
Fatty Acid biosynthesis
Malonyl-CoA
MCAT
FASN
Mitochondrial Outer Membrane Permeabilization
MYL9
CYCS
OXA1L
C1orf31
VDAC1
SLC25A4
TSPO
PPID
mt biogenesis and function
SPATA18
Pyruvate Metabolism
ME1
DLD
L-malate
Oxaloacetate
Glutaryl CoA
trans-But-2-enoyl-CoA
Glycerophospholipid metabolism
PHB
HSPE1
FASTK
LYPLA2
HIGD1A
Lipoamide-E
Dihydro-LipoamideE
Krebs Cycle
PMAIP1
Pantothenate and CoA biosynthesis
PANK2
MAT2B
L-Methionine
S-Adenosyl-L-methionine
Alpha-ketoglutarate
HSDL1
AK4
Fatty acid beta-oxidation
ETFA
DECR1
GRPEL2
AIFM1
ENDOG
VAPA
VAMP1
ATP6V1A
BCL2A1
Isocitrate
Citrate
Purine Metabolism
AK1
AK2
Val, Leu, Ile degradation
BCAT1
AUH
mt polymerase
POLG
mt translation elongation
GFM2
Pyrimidine Metabolism
TXNRD1
TXNRD2
NME2
NT5C
NIPSNAP3A
USMG5
BCL2L1
DIABLO
BCL2L2
HTRA2
CD44
MCL1
Protein metabolism
YME1L1
LYRM4
MTX1
Base Excision Repair
NTHL1
Arg, Pro metabolism
OAT
Ribosome
MRPS12
MRPL28
MRPL17
MRPS5
MRPS34
MRPL9
MRPL48
RPL35A
RPL34
MRPL16
MRPL11
MRPL21
MRPL27
MRPL40
SLC25A37
ARL2BP
BCL2L11
KIF1B
TP53AIP1
RHOT1
mt fusion
MFN1
PAM complex
GRPEL2
GRPEL1
TIMM44
Inner mitochondrial translocase
TIMM9
TIMM10
Mg transport
MRS2
Cl- channel
CLIC4
Fe transport
SLC25A37
Aminoacyl tRNA Biosynthesis
SARS
SARS2
KIAA1279
PINK1
ARMCX3
CREB1
tBID
BAX
BBC3
BAK1
BCL2
BNIP3L
BAD
HUWE1
TGM2
PPP3CA
DLC1
BCL2L11

Fine tuning regulation: post-translational modifications





The Self-Regulatory TP53 Network



Epigenetics and Epigenomics

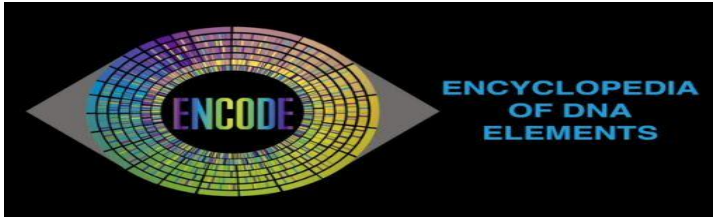
- **Epigenetics** encompasse processes that lead to heritable change in gene expression without changes to the DNA itself.
- DNA is packaged into chromatin. This nucleoprotein structure is highly dynamic and important for gene regulation. Chromatin states can vary between conditions, cells and tissue types and even within a single chromosome.
- The **Epigenome** refers to these chromatin states at a whole genome level. A multicellular organism has a single genome but many epigenomes.
    - Paradox: Although overall rates of cardiovascular disease increase with rising national prosperity, the least prosperous residents of a wealthy nation suffer the highest rates.

Developmental origin of health and disease

- The dutch famine ("Hongerwinter") 1944-45 in German occupied Netherlands towards the end of the WWII affected 4.5 million people and led to ~22000 deaths.
- *"People ate grass and tulip bulbs, and burned every scrap of furniture they could get their hands on, in a desperate effort to stay alive."*
- The Dutch Hunger Winter study, from which results were first published in 1976, provides an almost perfectly designed, although tragic, human experiment in the effects of intrauterine deprivation on subsequent adult health.
- Critical windows during development where epigenetic modification will affect adult health.
- Those exposed during **early gestation** experienced elevated rates of obesity, altered lipid profiles, and cardiovascular disease. In contrast, markers of reduced renal function were specific to those exposed in **mid-pregnancy**. Those who were exposed to the famine only during **late gestation** were born small and continued to be small throughout their lives, with lower rates of obesity as adults than in those born before and after the famine.

Schulz, PNAS 2010

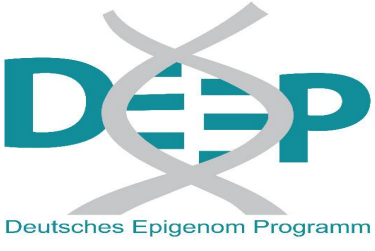
Large-scale epigenomic studies

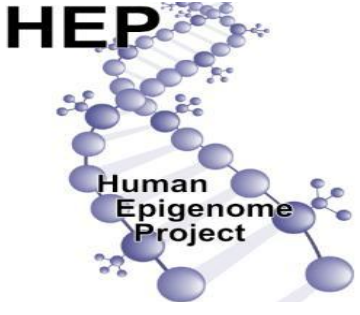


Histone and TF ChIP-seq, Transcriptomics, Hi-C

Epigenomes of 100 blood cell types

Stem cells, fetal tissues, adult tissues

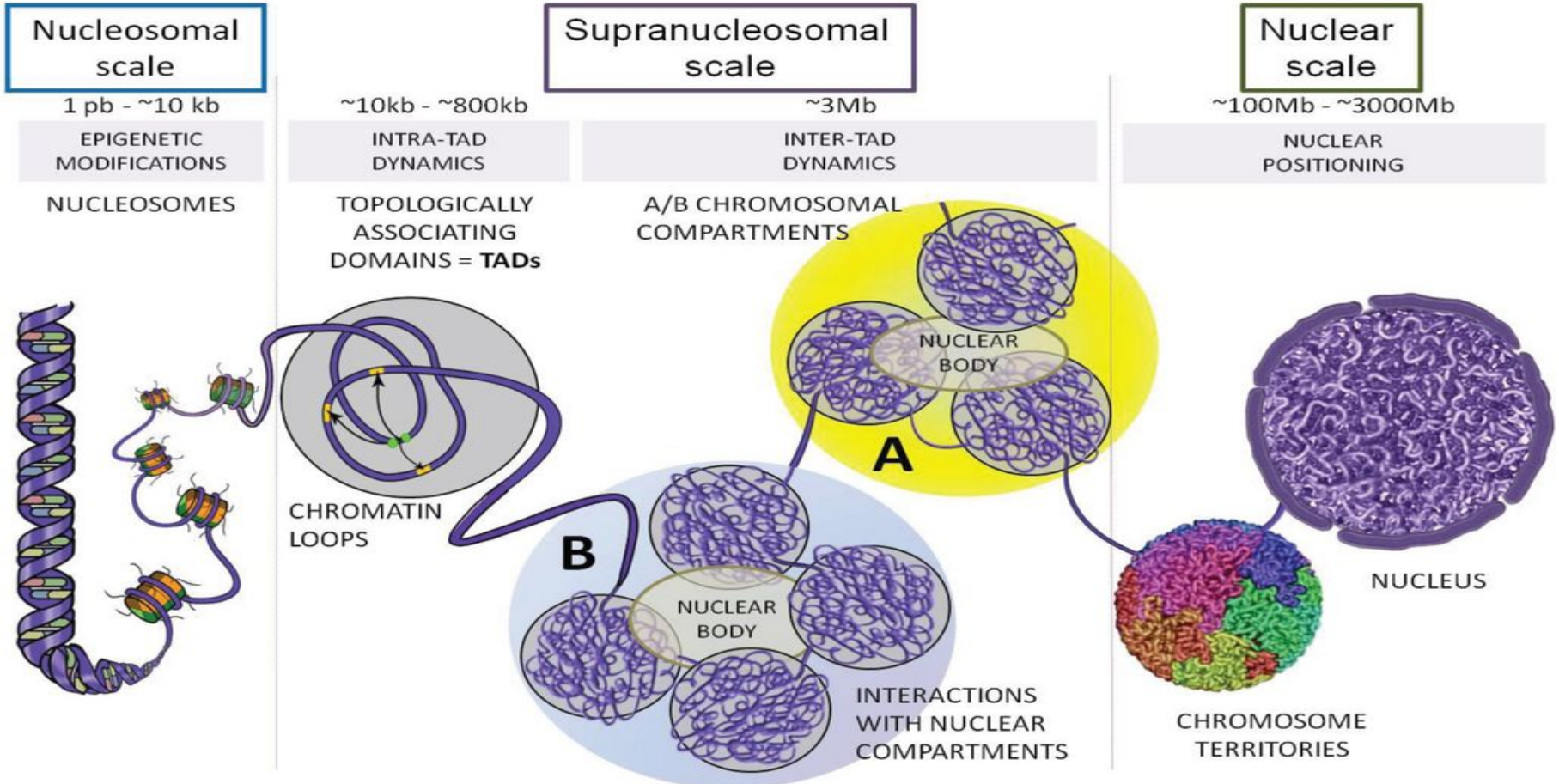




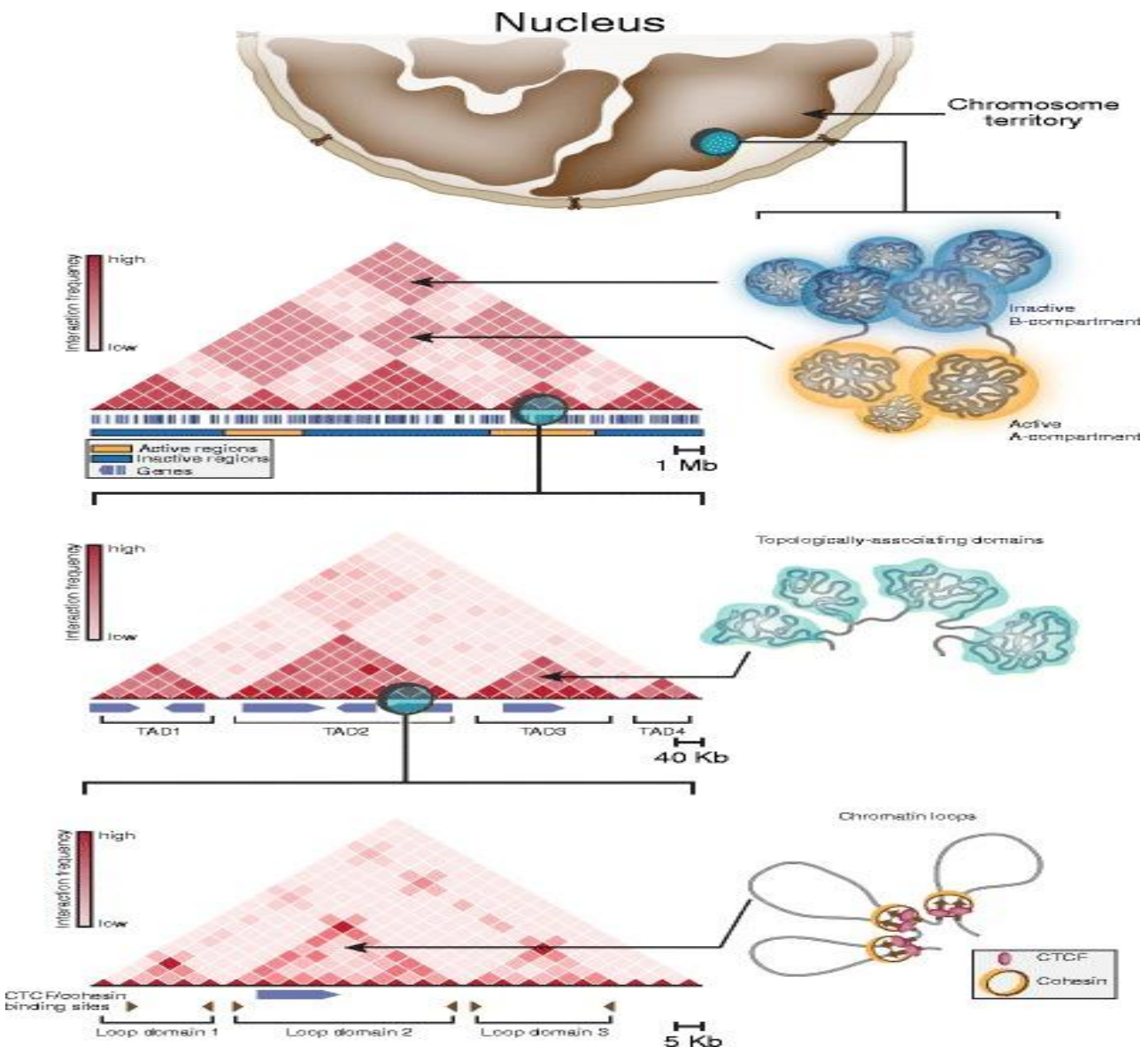
Methylomes

Various human, mouse tissues

Current model of chromatin organization



Ea et al., Genes 2015

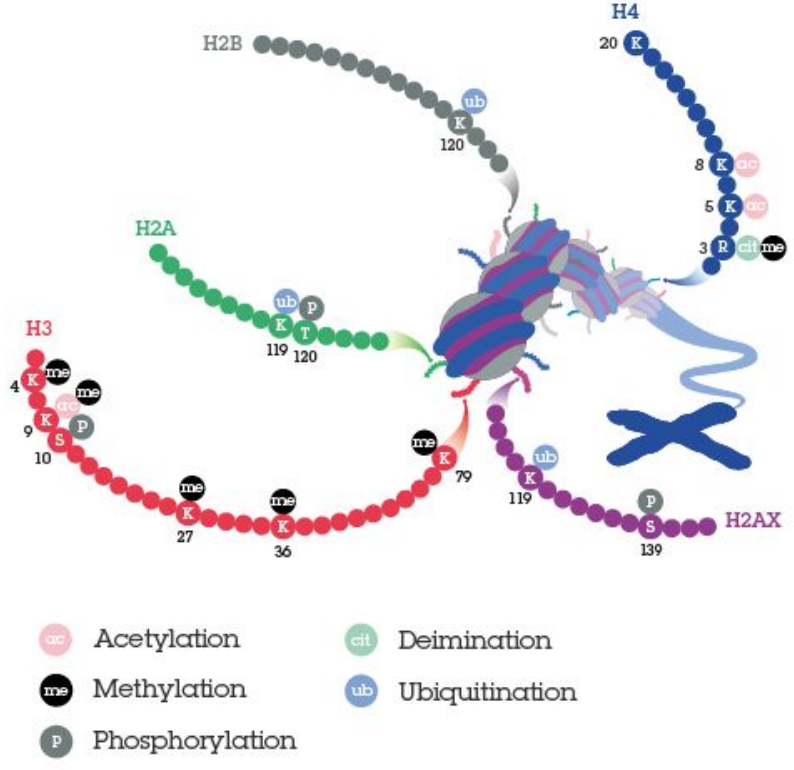

Nucleus
Chromosome territory
Interaction frequency
high
low
Inactive B-compartment
Active A-compartment
Active regions
Inactive regions
Genes
1 Mb
Topologically-associating domains
TAD1
TAD2
TAD3
TAD4
40 Kb
Chromatin loops
CTCF
Cohesin
CTCF/cohesin binding sites
Loop domain 1
Loop domain 2
Loop domain 3
5 Kb

Histone Modifications

- Nucleosomes consist of 2x H2A/H2B and 2x H3/H4 histones.
- 80 known covalent modifications

H3K4me3 →

H3	Histone 3
K	Residue is lysine, K
4	4th residue.
me3	Trimethylation



The most common histone modifications

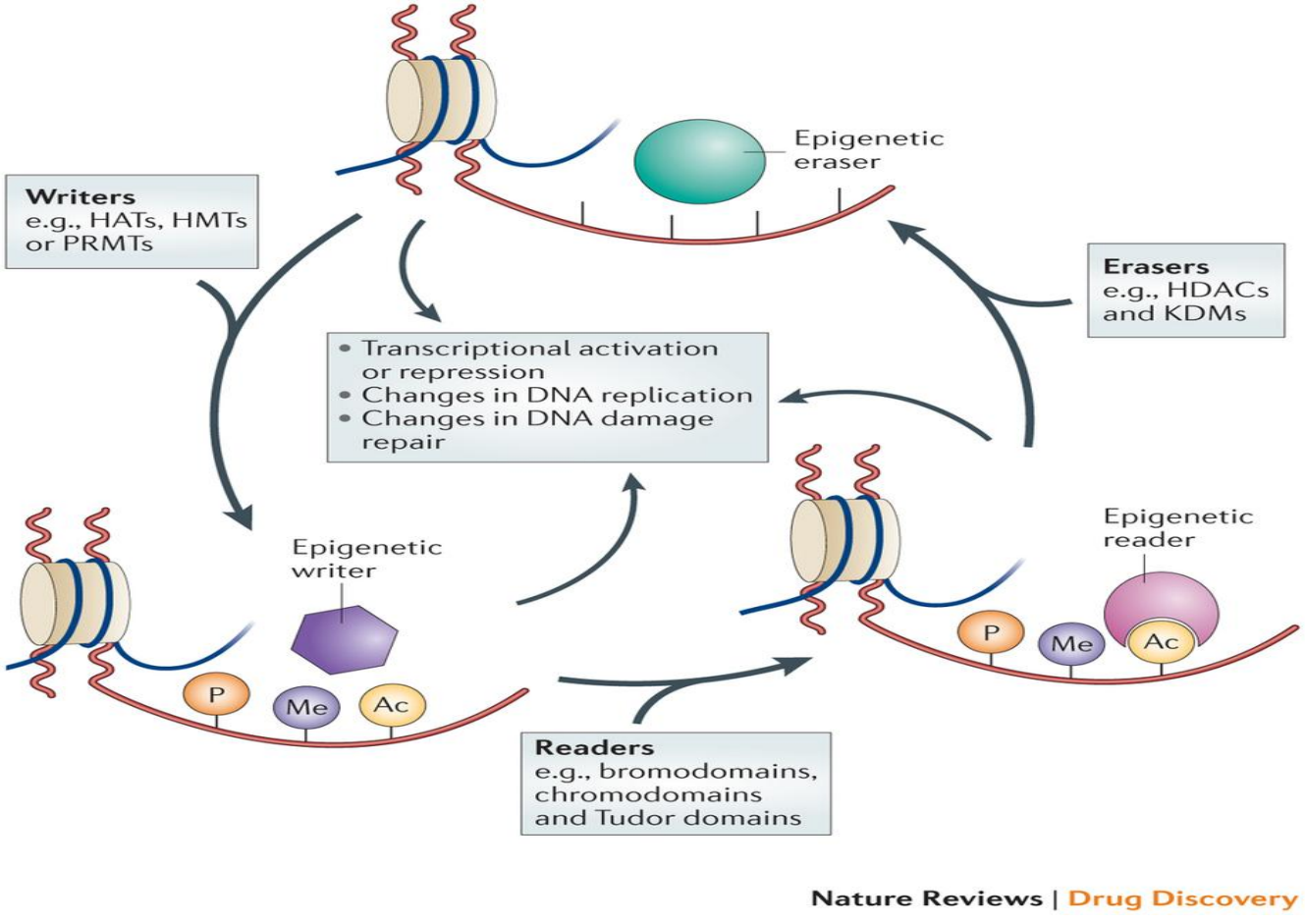
Histone Modifications

Some examples:

- H3K4me3 - active promoters
- High H3K4me1 and H3k27Ac, low H3K4me3 - active enhancers
- H3K27me3 -repression at promoters
- H3K9me3 - Heterochromatin (inactive, condensed chromatin)

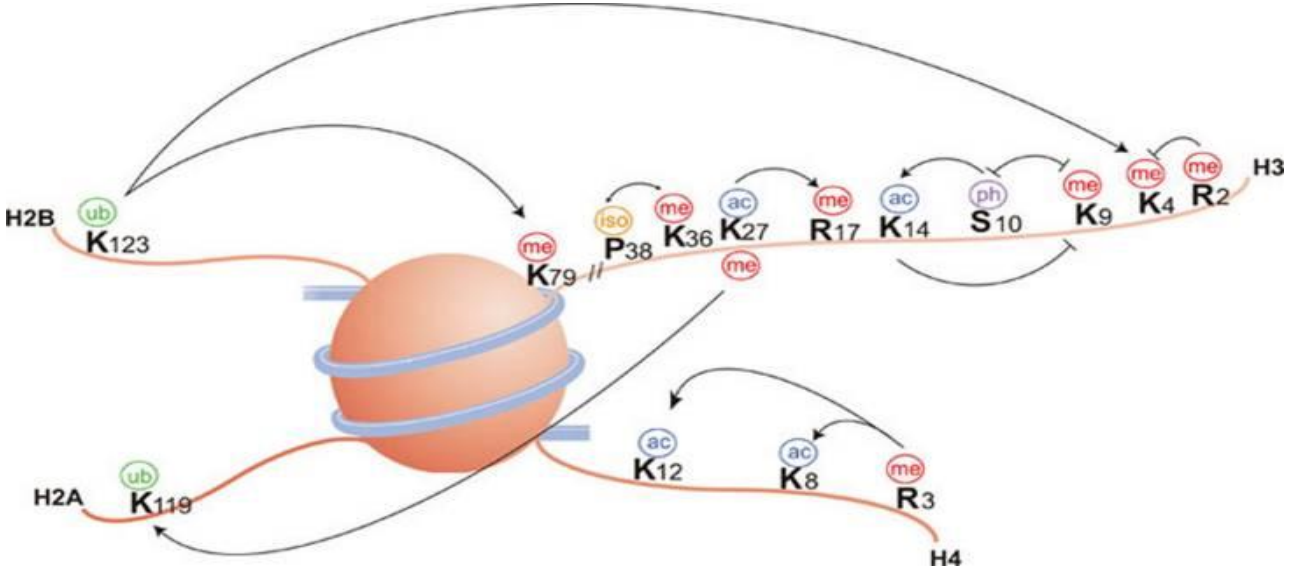
More information at:
http://epigenie.com/key-epigenetic-players/histone-proteins-and-modifications/

Epigenetic Readers, Writers and Erasers



Falkenberg and Jonstone, Nature Reviews Drug Discovery, 2014

Combinations of marks can have different effects



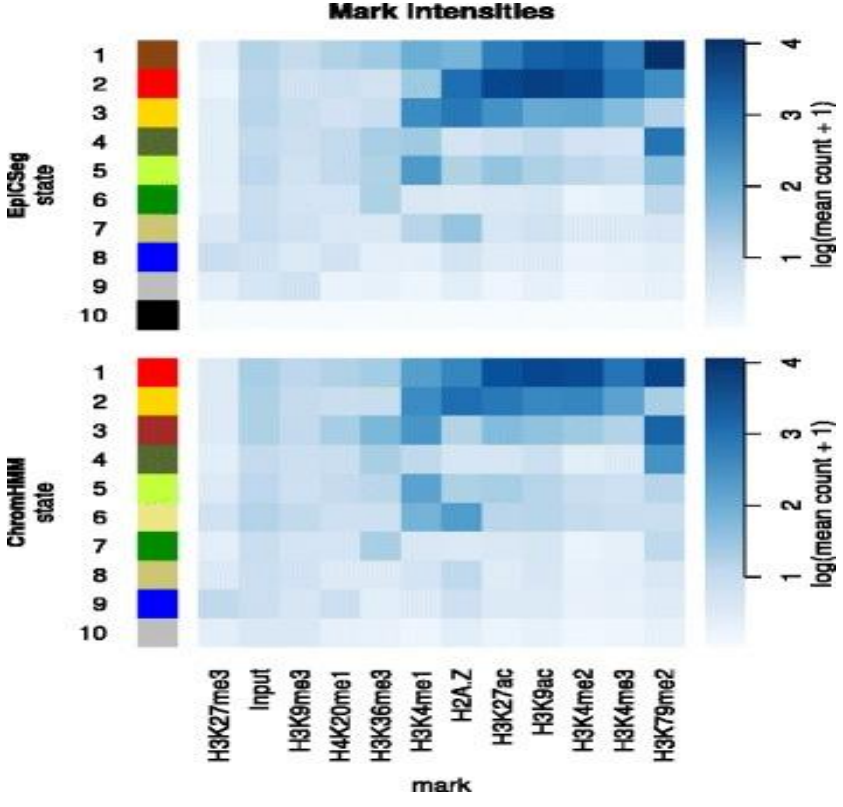
Bannister and Kouzarides (Cell Res. 2011)

To understand the entire code need to ChIP-seq each mark. This information has to be integrated and simplified.

Simplifying histone marks

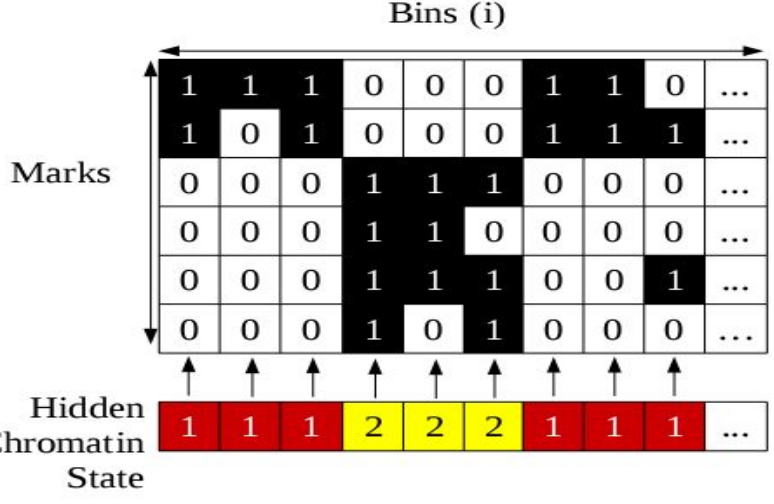
Unsupervised learning methods for *segmentation*;

- ChromHMM (Ernst et al., 2011)
- Segway (Hoffman et al., 2012)
- EpiCSeg (Mammana and Chung, 2015)
- GenoSTAN (Zacher et al., 2016)



Chromatin Segmentation Algorithms

- Genome divided into 200bp bins
- Adjust read position (shift 5’ of each read 5’->3’ by 0.5 the fragment length)
- Count reads in each bin for each mark and generate count matrix
- HMM with specified states is used to model the count matrices and derive segmentation



Chromatin Segmentation

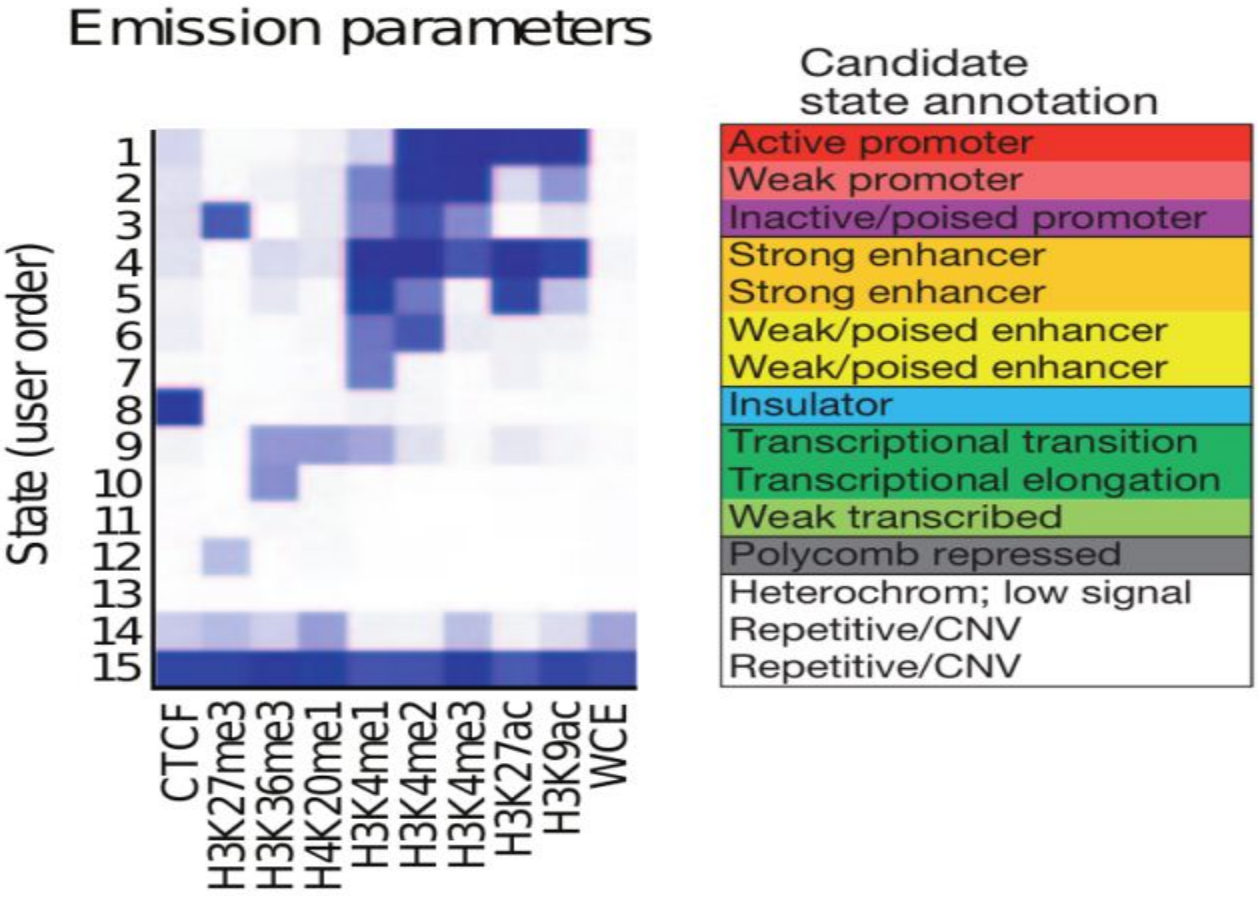
Advantages:

- Derived states, not vectors of chromatin marks -easier to determine genome wide properties.
- Can train on one set and apply to another.

Disadvantages:

- How many states?
- Histone states binary -lose information (except in EpiCSeq)
- Causality unknown

Chromatin Colours



Visualizing Chromatin Marks