Epigenetics and non-coding RNAs

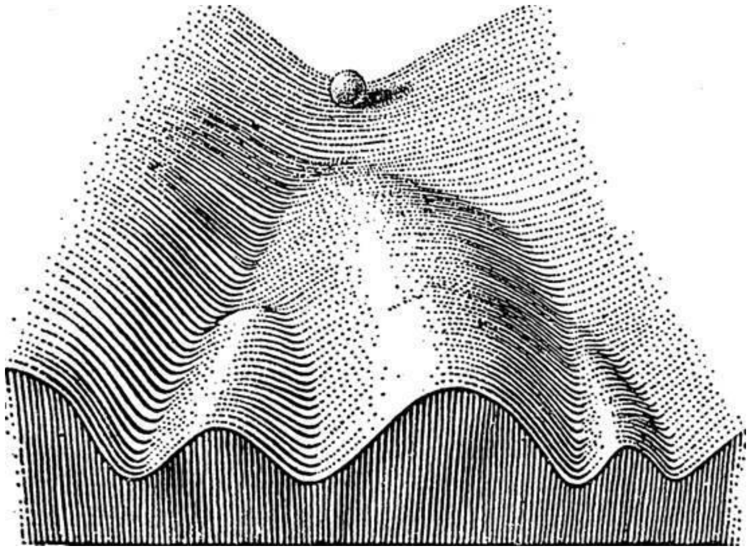
Feb 22 2016

Lecture Plan

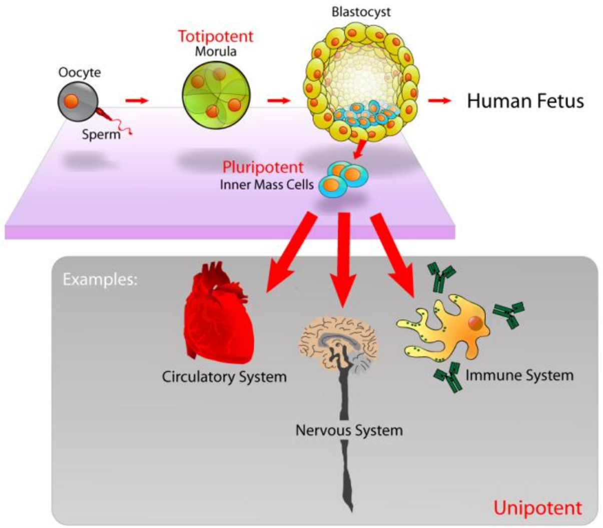
- Today: Epigenetics and lincRNAs
- Wednesday: statistical methods for RNAseq/ChIPseq
- Next Monday: microRNAs

Epigenetics

- Beyond/on-top-of genetic
  - Study of (stable/heritable) non-genetic differences in traits
- **Chromatin**: DNA and all the proteins bound it
- Chromatin structure dictates the transcriptional potential of a cell
- The structure is heritable: across cell divisions and sometimes trans-generationally

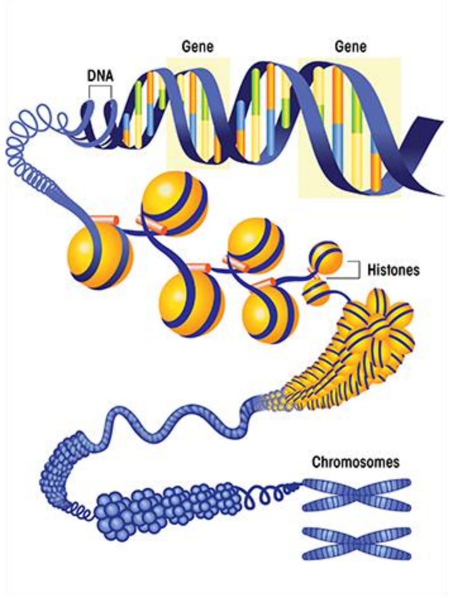


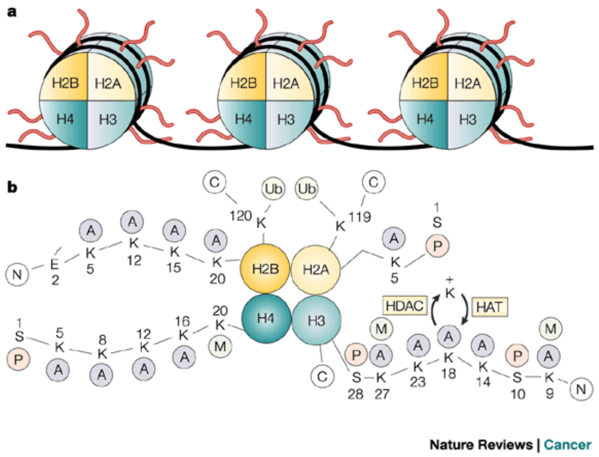
Waddington epigenetic landscape



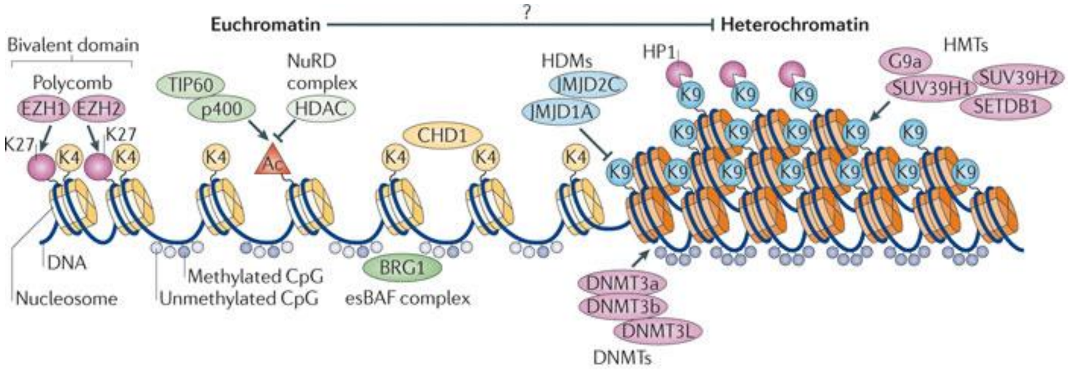
Terms and Definitions

- **Histones**: proteins found in eukaryotic cell nuclei that package and order the DNA into structural units called **nucleosomes**
- **Euchromatin:** open accessible chromatin, transcribed regions
- **Heterochromatin:** closed chromatin: inaccessible to transcriptional machinery
- **Epigenetic marks:** covalent alterations to histones or DNA
  - Generally copied at cell division
  - Determine chromosome structure recruitment of genes that execute and regulate gene expression
- **Epigenomics:** the study of the genome wide epigenetic state





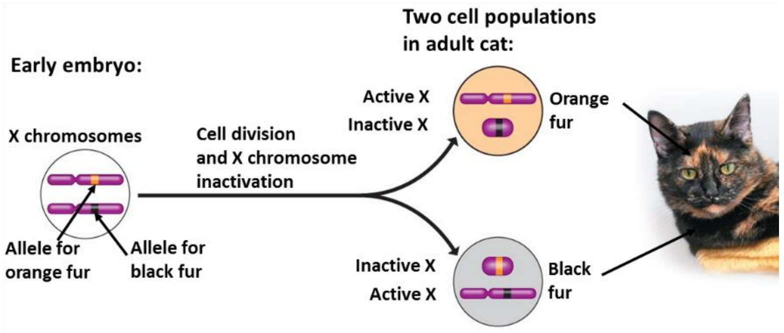
Nature Reviews | Cancer

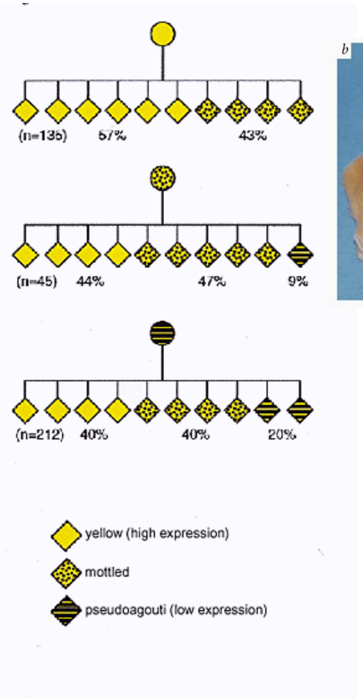


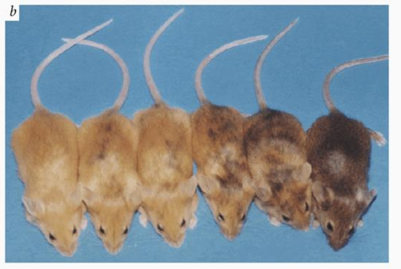
Nature Reviews | Molecular Cell Biology

Examples of epigenetic processes

- X chromosome inactivation
  - In females each cell randomly inactivate 1 X chromosome and this pattern is inherited through cell divisions
  - Mediated by a non-coding RNA Xist
- Coat coloration in mice
  - Isogenic mice have different coat patterns that are partially heritable
- Paternal/maternal imprinting
  - Some genes are only expressed from a single parental chromosome in a cell–type specific way
  - Loss of a part of chromosome 15
  - Maternal: Angelman syndrome
    - Sever learning problems, seizures, movement problems, unusually cheerful disposition
  - Paternal: Prader-Willi syndrome
    - Learning difficulties, short stature, compulsive eating

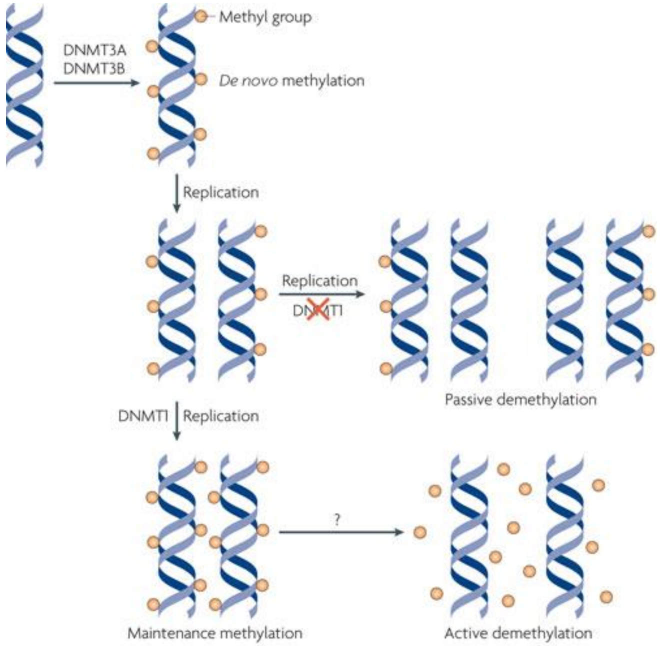






Epigenetic Marks – DNA Methylation

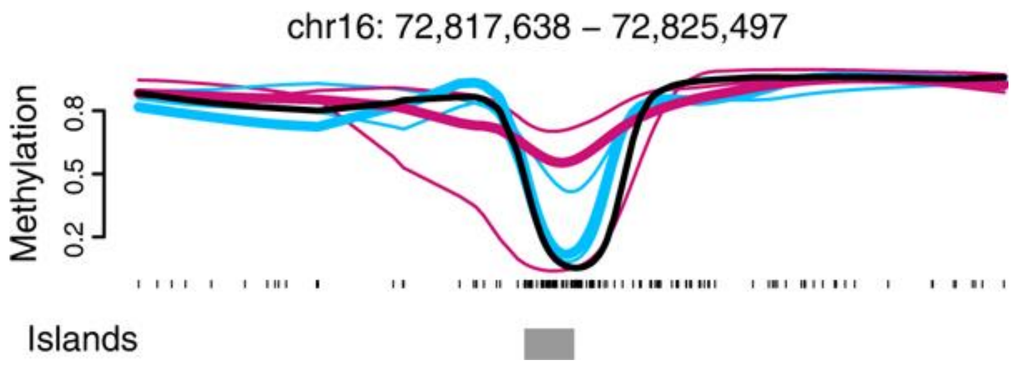
- Methylation (mammalian genome)
    - Mostly on CpG dinucleotides
    - Most are methylated in the mammalian genome (70-80%)
    - CG dinucleotides are relatively rare because of higher mutation rate
    - CpG islands are stretches of DNA relatively rich in CpGs
        - Working definition: had a GC content greater than 55%, and an observed-to-expected CpG ratio of 65%
        - Typically associated with promoters of genes
        - 70% of genes have canonical CpG islands
- Methylation is repressive: closed chromatin
- Methylations are relatively stable
    - Actively copied during replications (failure to do so results in passive loss)
    - Active demethylase activity is typically low

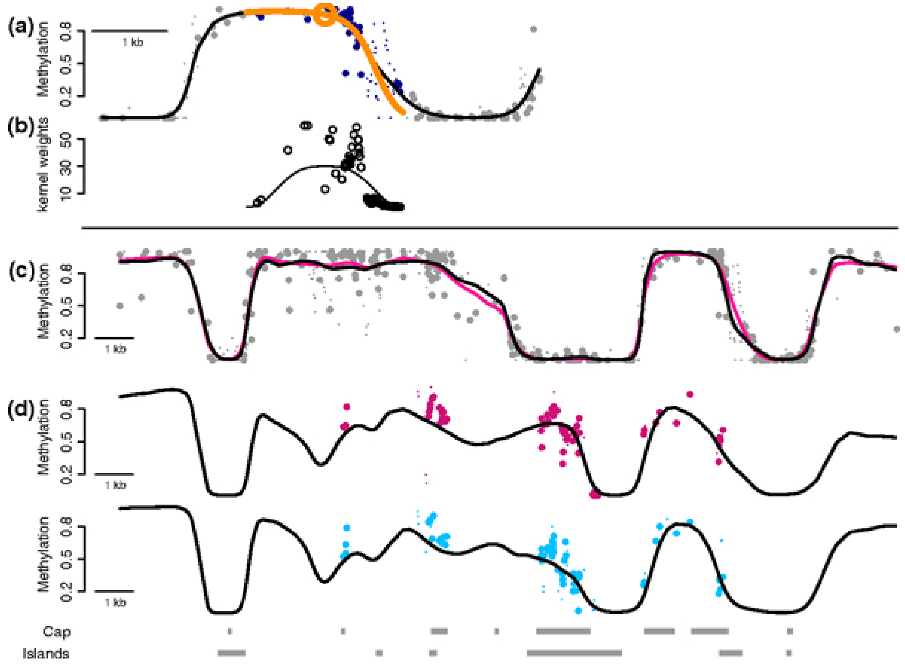


Nature Reviews | Molecular Cell Biology

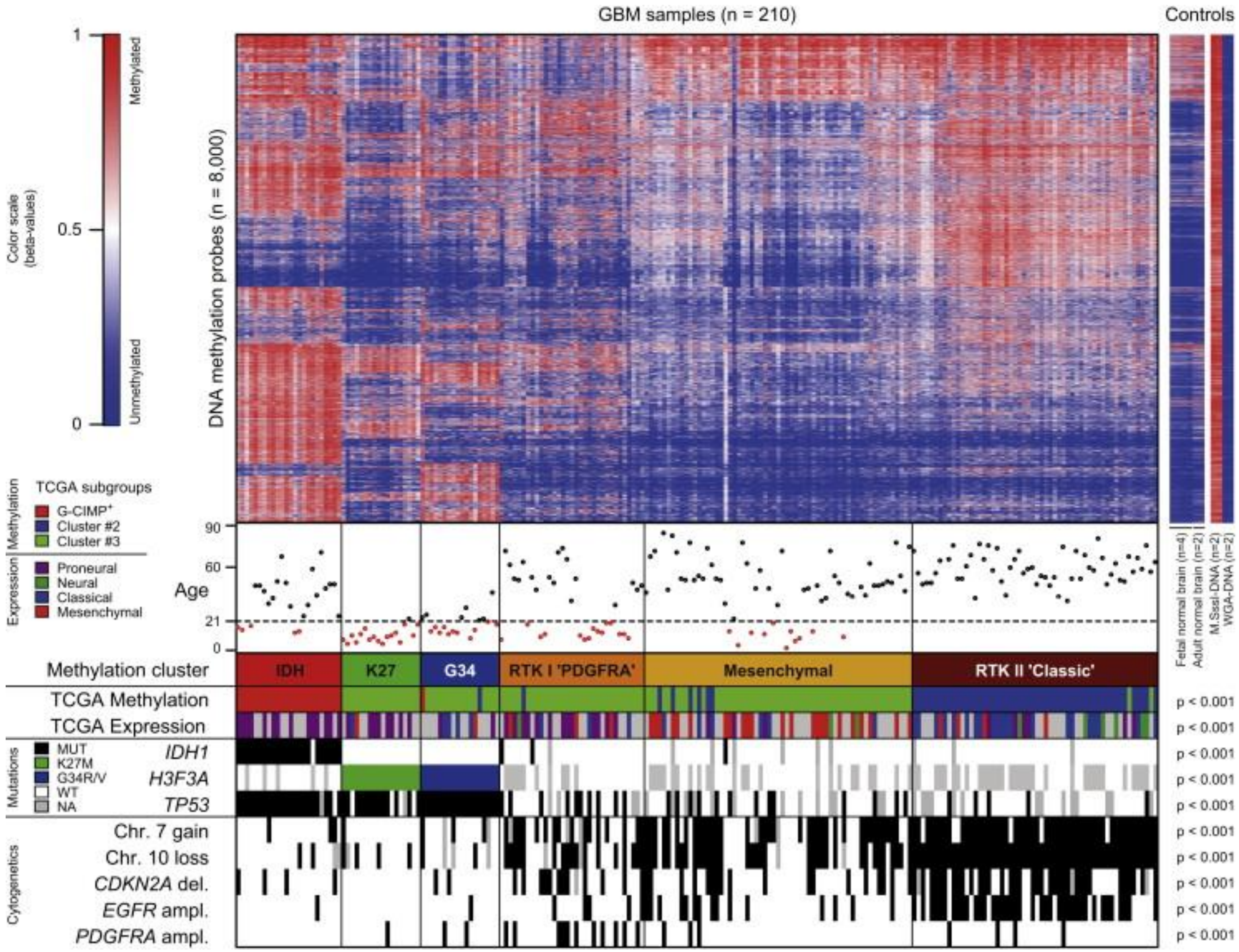
Methylation techniques and analysis

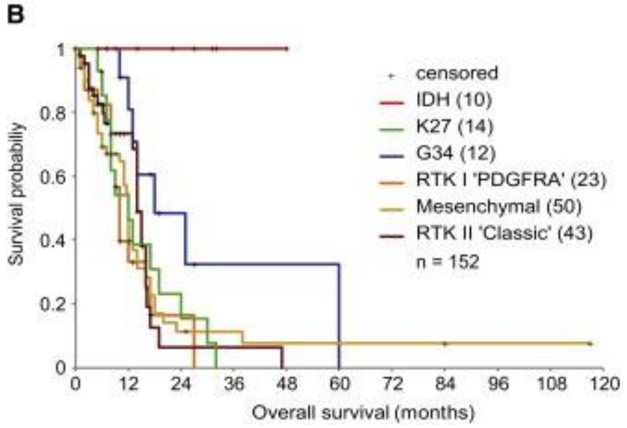
- Bisulfite conversion: chemical procedure turns unmethylated Cs into Us
- Relative abundance of Cs and Us at a single locus can be determined by hybridization or sequencing
- Only one genome per cell!
  - Single cell values are {0, 0.5, 1}
  - Assay values [0,1] represent population averages
- Computational challenges
  - Spatial dependence
  - Biological variability
  - Hypothesis testing: many CpG loci
- Popular software: bSmooth

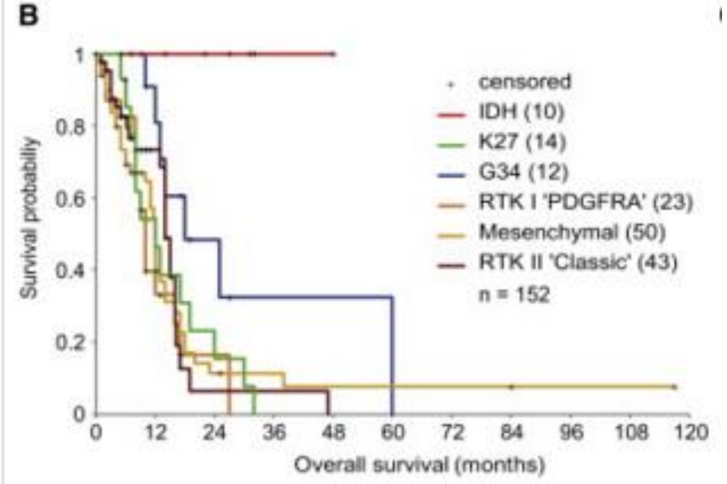




Methylation classification of cancers

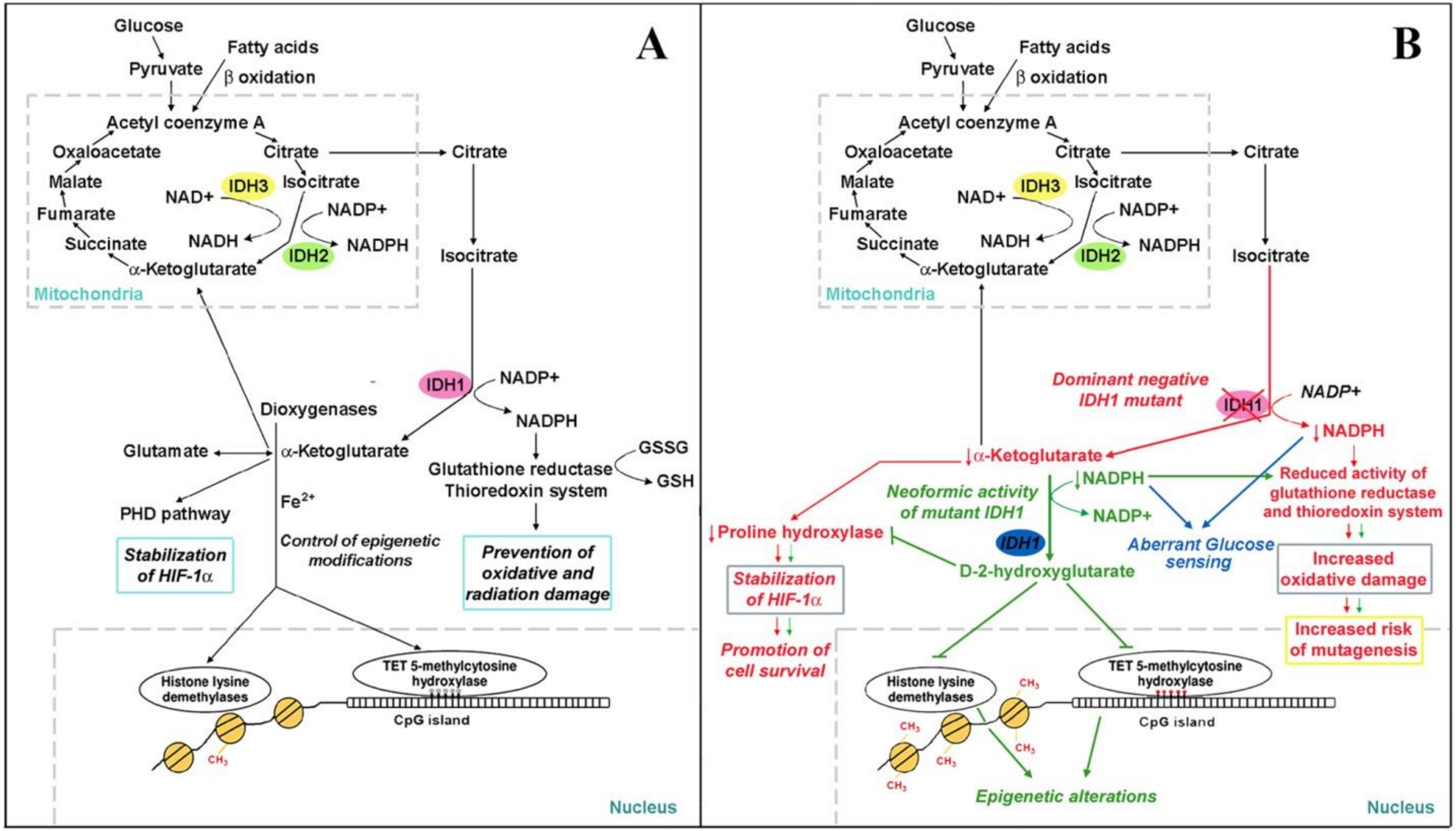






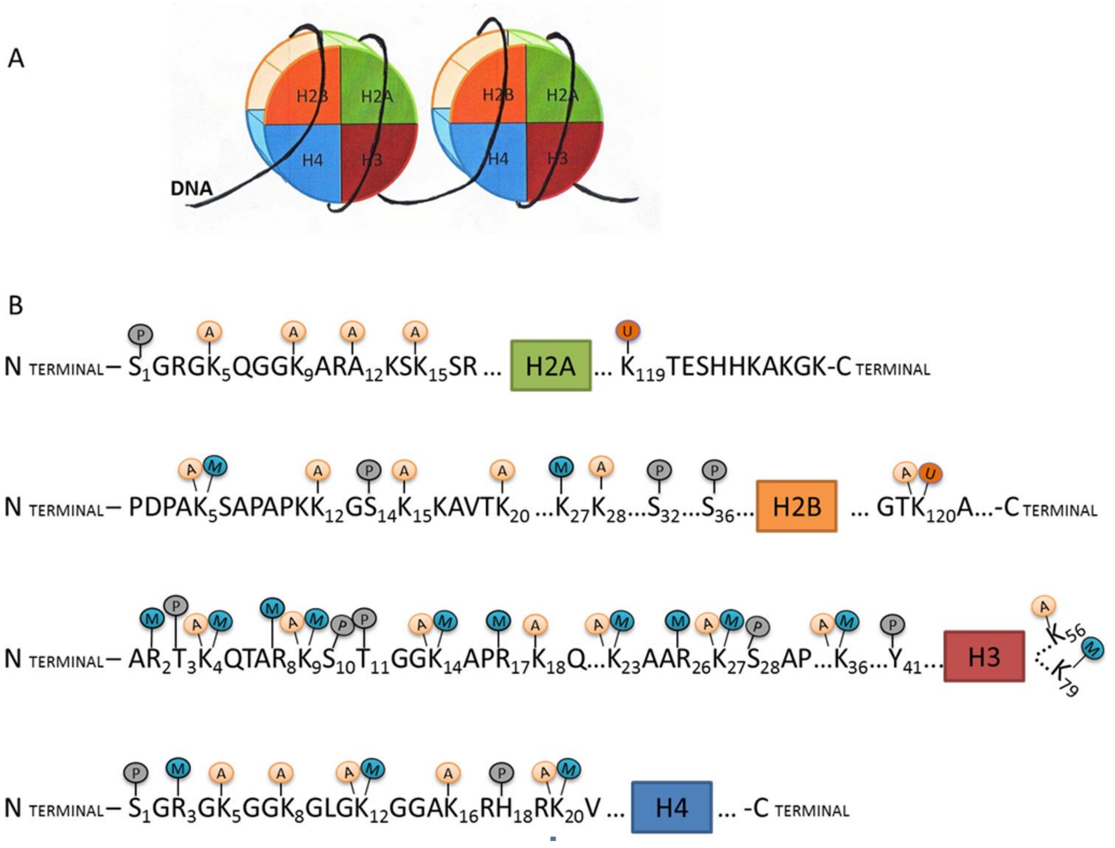
Strum et al. Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma

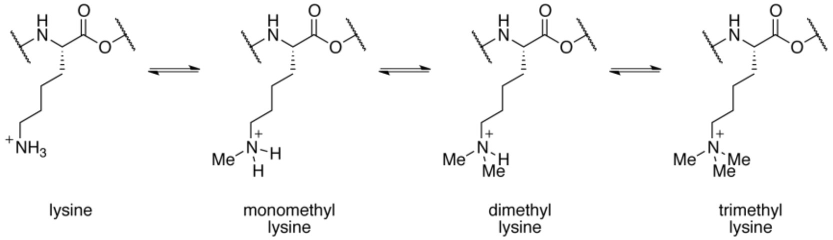
Mechanism—epigenetic and metabolism are linked



Histone modification

- Many different modifications possible
  - Acetylation
  - Methylation
  - Phosphorylation
  - Histone splice variants affect possible modification
- Consequences?
  - DNA is acidic and histones are basic
  - Histone marks change the DNA histone binding affinity creating open chromatin region—basic chemistry
  - Interact with gene transcription machinery—complex
  - Combinatorial effects?

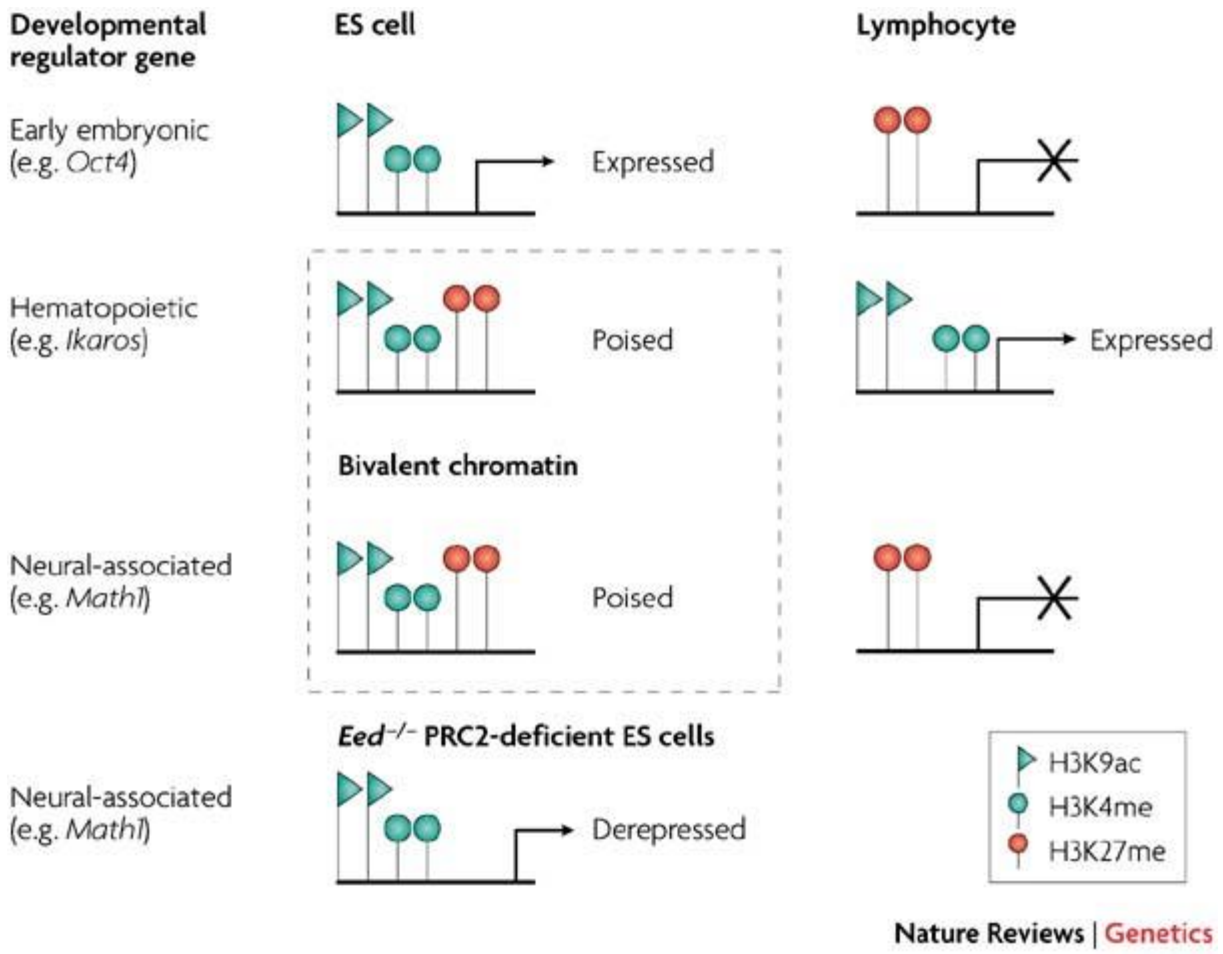

A
H2B
H2A
H4
H3
DNA
B
N TERMINAL– S1GRGK5QGGK9ARA12KSK15SR ... H2A ... K119TESHHKAKGK-C TERMINAL
N TERMINAL– PDPAK5SAPAPKK12GS14K15KAVTK20 ...K27K28...S32...S36... H2B ... GTK120A...-C TERMINAL
N TERMINAL– AR2T3K4QTAR8K9S10T11GGK14APR17K18Q...K23AAR26K27S28AP...K36...Y41... H3 K56 K79
N TERMINAL– S1GR3GK5GGK8GLGK12GGAK16RH18RK20V ... H4 ... -C TERMINAL
P
A
M
U



Histone modifications

Histone modification or variant	Signal characteristics	Putative functions
H2A.Z	Peak	Histone protein variant (H2A.Z) associated with regulatory elements with dynamic chromatin
H3K4me1	Peak/region	Mark of regulatory elements associated with enhancers and other distal elements, but also enriched downstream of transcription starts
H3K4me2	Peak	Mark of regulatory elements associated with promoters and enhancers
H3K4me3	Peak	Mark of regulatory elements primarily associated with promoters/transcription starts
H3K9ac	Peak	Mark of active regulatory elements with preference for promoters
H3K9me1	Region	Preference for the 5′ end of genes
H3K9me3	Peak/region	Repressive mark associated with constitutive heterochromatin and repetitive elements
H3K27ac	Peak	Mark of active regulatory elements; may distinguish active enhancers and promoters from their inactive counterparts
H3K27me3	Region	Repressive mark established by polycomb complex activity associated with repressive domains and silent developmental genes
H3K36me3	Region	Elongation mark associated with transcribed portions of genes, with preference for 3′ regions after intron 1
H3K79me2	Region	Transcription-associated mark, with preference for 5′ end of genes
H4K20me1	Region	Preference for 5′ end of genes

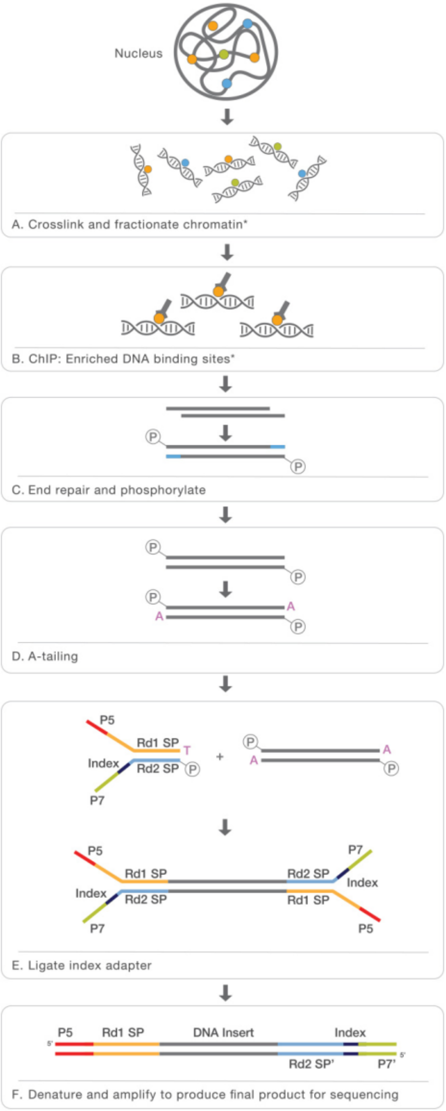
Bivalent chromatin in stem cells


Developmental regulator gene
ES cell
Lymphocyte
Early embryonic (e.g. *Oct4*)
Expressed
Hematopoietic (e.g. *Ikaros*)
Poised
Expressed
Bivalent chromatin
Neural-associated (e.g. *Math1*)
Poised
Eed$^{-/-}$ PRC2-deficient ES cells
Neural-associated (e.g. *Math1*)
Derepressed
H3K9ac
H3K4me
H3K27me

Nature Reviews | Genetics

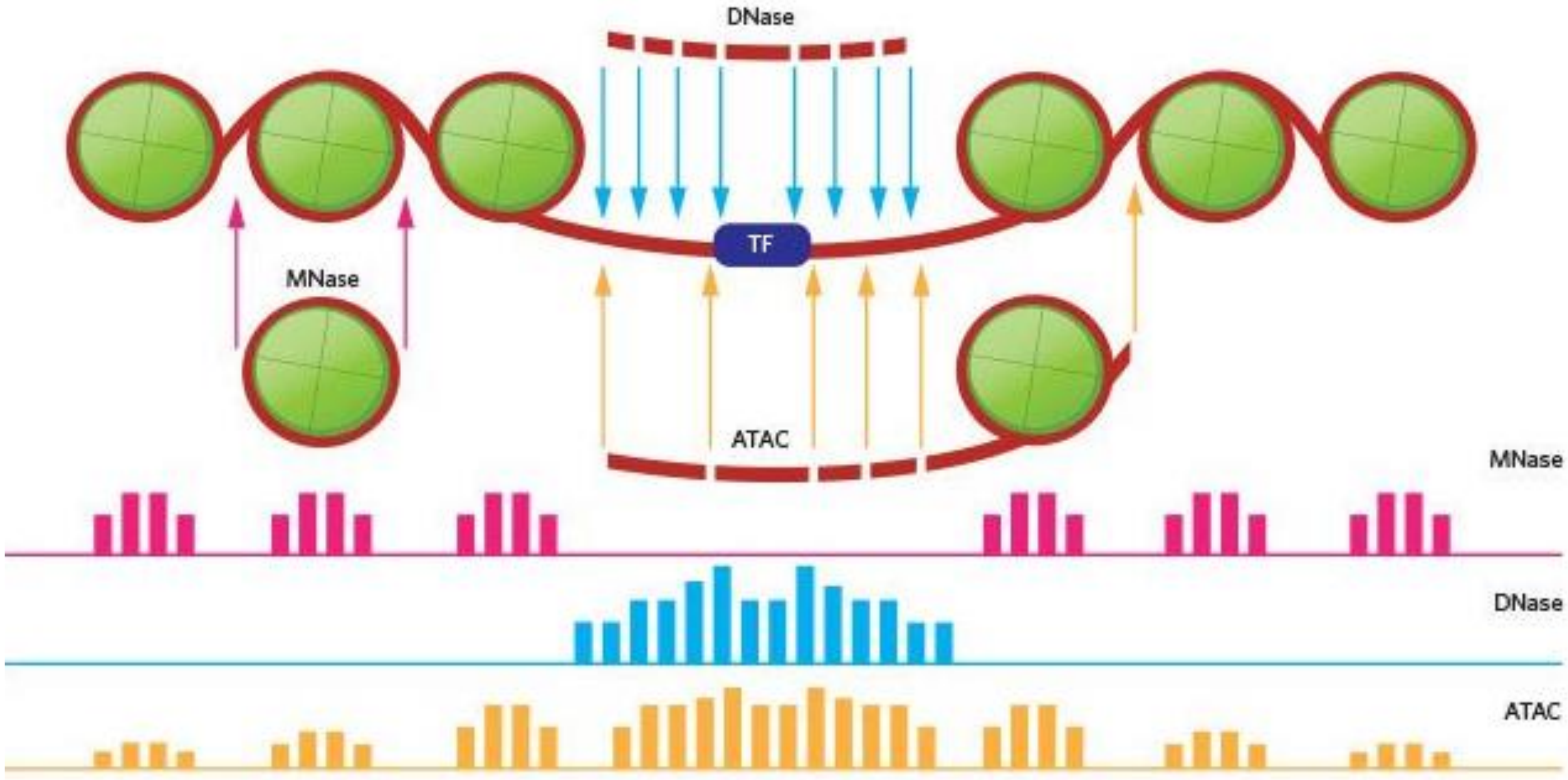
ChIPseq

- **ChIP**seq—chromatin immuno-precipitation
  - Grab the part of chromatin we are interested in
  - Nucleosome mark or other chromatin proteins –ex: Transcription factors
- ChIP**seq—how to quantify**
  - By sequencing-most recent method highly accurate and unbiased
  - ChIPchip-microarray like hybridization method– need genomics probes
  - ChIP-PCR-small scale validation method

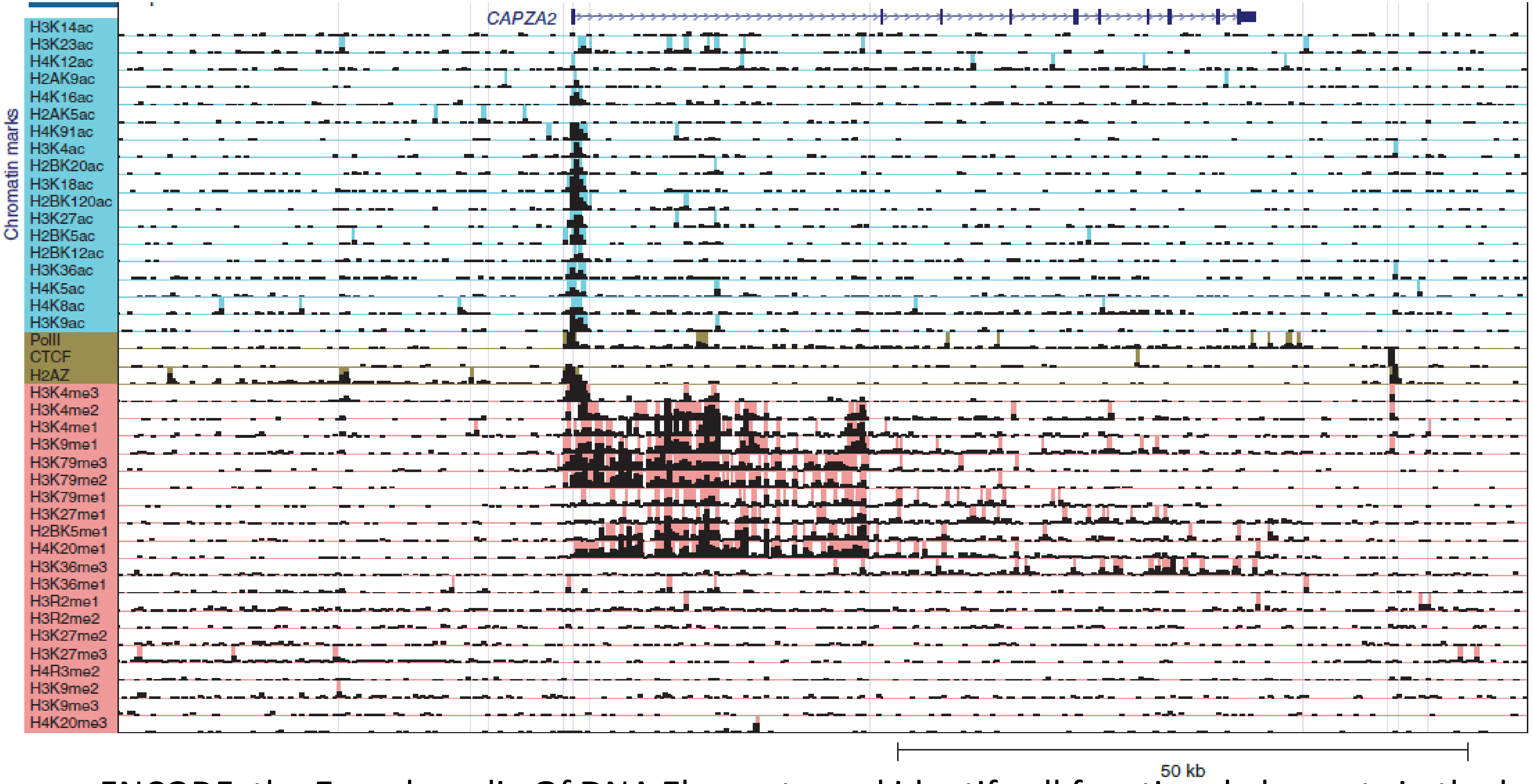

Nucleus
A. Crosslink and fractionate chromatin*
B. ChIP: Enriched DNA binding sites*
P
C. End repair and phosphorylate
P
A
D. A-tailing
P5
Rd1 SP
T
Index
Rd2 SP
P7
+
Rd2 SP
Rd1 SP
E. Ligate index adapter
P5
Rd1 SP
DNA Insert
Index
Rd2 SP'
P7'
F. Denature and amplify to produce final product for sequencing

Open chromatin assays

- Looking at which parts of the chromatin are accessible
- DNAseq—digest all unbound DNA
- ATACseq—transposon inserted near nucleosome boundaries



What the data looks like



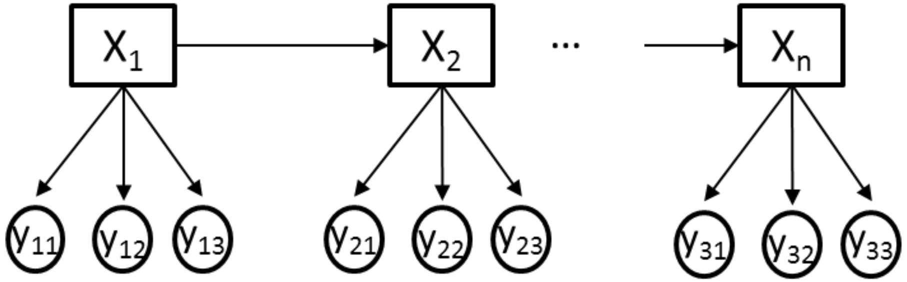
- ENCODE, the **Enc**yclopedia **O**f **D**NA **E**lements goal identify all functional elements in the human genome sequence
- 38 histone marks: 2^38 states?
- What is the combinatorial complexity?

Discovery and characterization of chromatin states for systematic annotation of the human genome
Jason Ernst & Manolis Kellis, Nature Biotech 2010

- **chromoHMM**
- Multivariate Hidden Markov Model
- The number of states is unknown!
- Many random initializations with variable number of states-fit parameters by EM
- Best model has 79 states-using **BIC**
- Use best model to initialize smaller models—nested initialization
- Final model has 51 states

X_t : hidden state variables

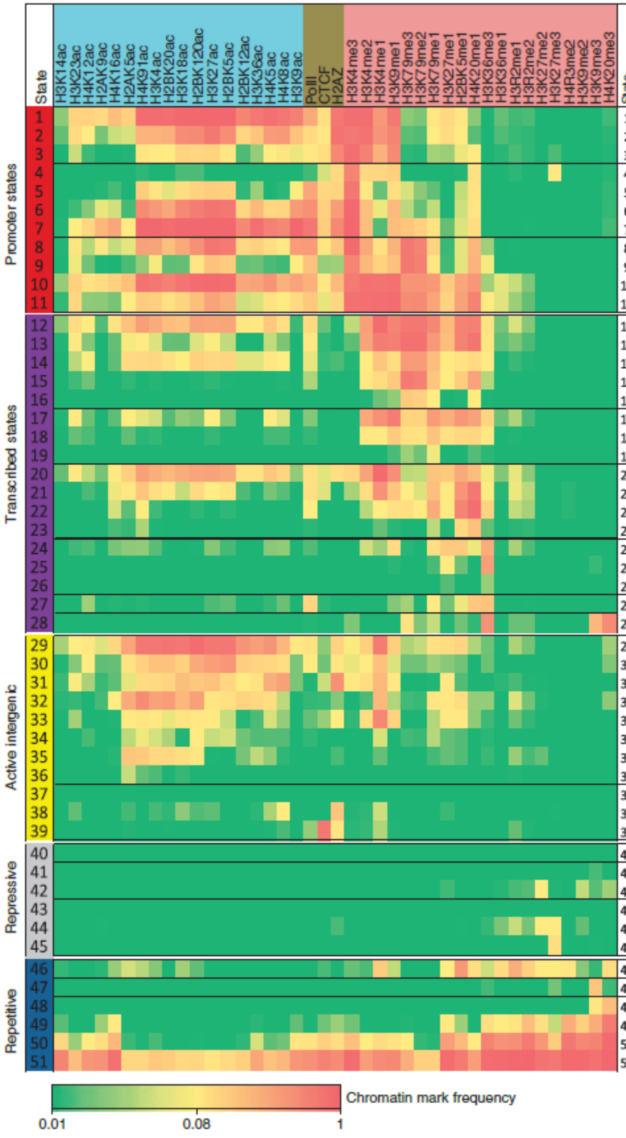
y_{ti} : ith observed variable @ t



Note about model selection

- Want the best model that doesn't overfit the data
- Bayesian information criterion: BIC = −2 × ln(likelihood) + **ln(N) × k**
  - k = number of parameters estimated
  - N = number of observations
- Akaike information criterion: AIC = −2 × ln(likelihood) + **2 × k** -- less conservative: will produce larger models
- BIC: attempt to finds the "true" model
- AIC: "true" model is infinitely complex so find the best predictive model
  - Asymptotically equivalent to minimizing parameter CVs
  - Similar to cross-validation

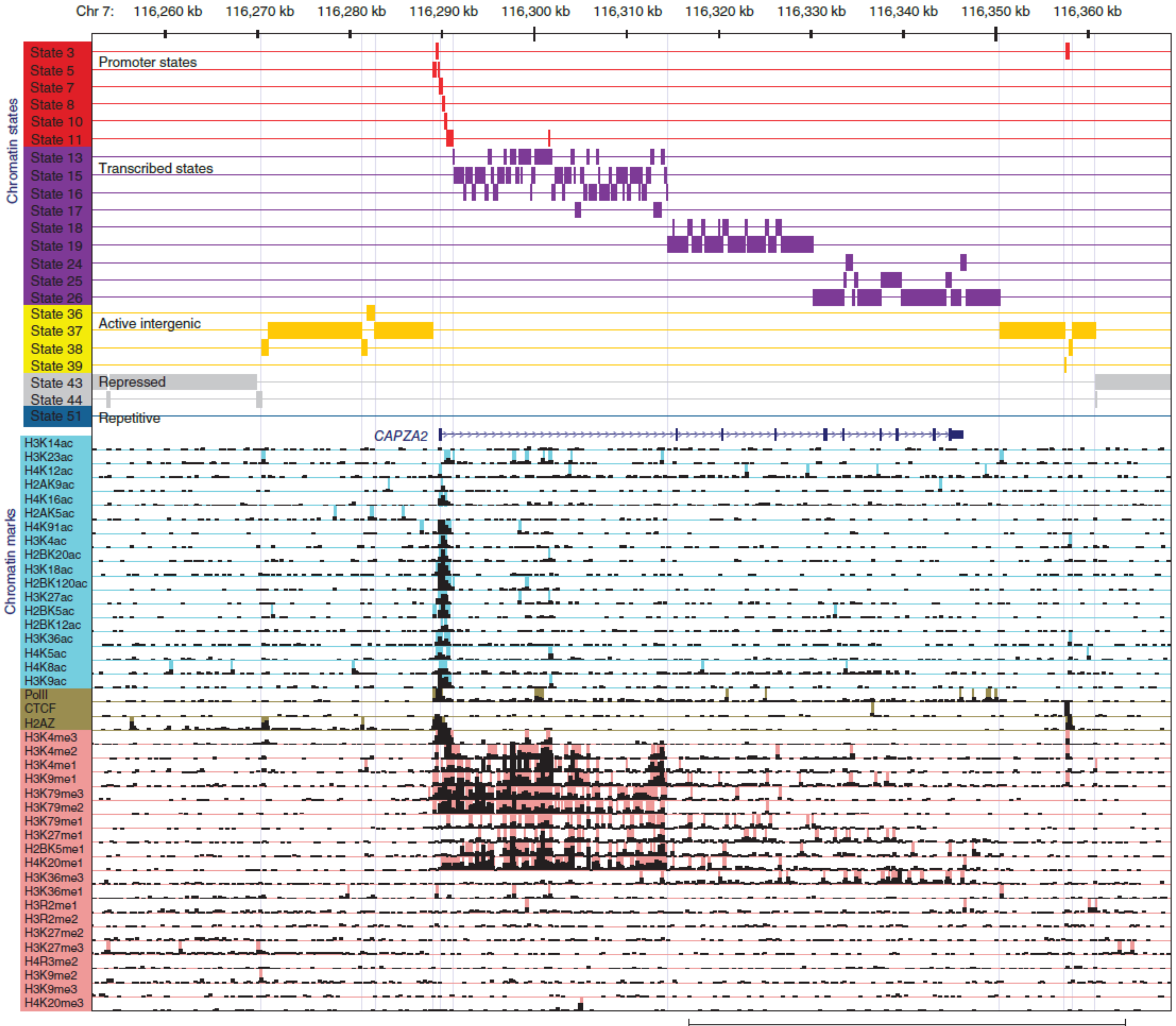
51 Epigenetic states



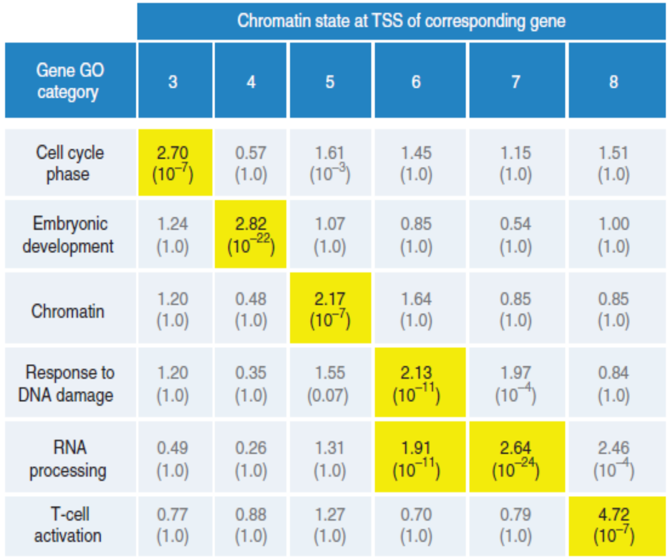
State	Percent of genome	% +-2kb TSS	Percent of TSS	xF TSS exact	% RefSeq gene	Expression level	xF ZNF gene	xF 5′ UTR	xF All exons	xF Spliced exons	xF 3′ UTR	xF TES	xF Conserved	xF DNaseI	xF TF binding	xF CpG island	% GC	% Lamina	% Repeat	
1	0.11	51	2.8	26	45	0.19	2.9	3.6	4.2	1.5	1.5	2.3	2.3	38	6.3	29	53	13	30	Promoter upstream high expr; potential enh looping
2	0.14	41	2.2	16	44	0.04	2.5	2.9	3.2	1.7	1.6	2.3	2.3	34	5.8	18	50	17	36	Promoter upstream med expr; potential enh looping
3	0.16	52	4.6	28	49	-0.15	2.6	3.7	4.9	2.2	1.7	2.5	2.8	32	7.1	37	53	16	35	Promoter upstream low expr; potential enh looping
4	0.19	57	8.5	44	52	-0.53	1.1	4.5	7.6	2.5	2.0	2.5	3.9	19	5.7	69	61	23	33	Repressed promoter
5	0.11	74	14	125	63	0.36	2.9	7.6	11	1.9	1.8	3.4	4.4	36	9.4	76	64	9.7	49	TSS low-med expr; most GC rich
6	0.11	78	12	109	61	0.41	4.1	7.7	12	1.8	1.9	3.6	4.1	39	10	78	61	9.5	27	TSS med expr
7	0.08	89	9.0	115	70	0.66	5.3	9.5	14	1.2	1.6	1.9	4.3	39	11	86	61	8.7	16	TSS high expr
8	0.07	72	3.0	40	87	0.96	1.8	8.1	7.6	1.9	1.4	1.6	3.7	35	8.6	44	54	10	26	Transcribed promoter; highest expr, TSS for active genes
9	0.10	41	1.0	10	89	1.04	1.3	6.0	3.6	2.2	2.6	3.1	2.8	20	3.3	11	42	13	39	Transcribed promoter; highest expr, downstream
10	0.08	48	0.6	7.3	87	0.75	2.5	5.7	4.0	2.4	1.3	1.3	2.4	33	6.7	17	53	7.1	29	Transcribed promoter; high expr, near TSS
11	0.09	54	0.4	4.3	92	0.85	2.4	5.4	4.0	3.2	2.4	3.4	2.6	25	5.8	10	50	7.2	34	Transcribed promoter; high expr, downstream
12	0.14	8.3	0.1	0.7	93	0.82	0.7	5.2	1.6	1.6	1.1	0.5	1.5	9.1	3.1	0.5	51	7.9	40	Transcribed 5′ proximal, higher expr, open chr, TF binding
13	0.39	6.3	0.2	0.5	95	0.77	1.8	5.8	1.2	1.3	1.0	0.5	1.3	3.3	1.9	0.7	46	6.0	52	Transcribed 5′ proximal, higher expr, open chr
14	0.22	9.2	0.1	0.7	82	0.70	1.4	4.4	1.4	1.5	1.6	1.2	1.5	6.2	1.7	0.2	42	12	55	Transcribed 5′ proximal, high expr, open chr
15	0.78	3.4	0.2	0.3	95	0.67	2.7	5.9	0.9	0.9	0.9	0.4	1.4	1.4	1.0	0.4	40	6.6	64	Transcribed 5′ proximal, high expr
16	0.90	3.7	0.1	0.2	92	0.53	3.2	5.4	0.5	0.5	1.0	0.3	1.0	0.9	0.7	0.1	41	8.5	79	Transcribed 5′ proximal, med expr; Alu repeats
17	0.25	6.6	0.1	0.5	94	0.49	3.2	3.8	1.6	1.6	1.4	0.8	1.3	2.9	2.0	0.1	46	10	52	Transcribed less 5′ proximal, med expr, open chr
18	0.79	2.7	0.2	0.3	93	0.43	2.8	3.5	1.3	1.4	1.1	0.7	1.3	1.1	1.0	0.1	42	10	60	Transcribed less 5′ proximal, med expr
19	2.02	2.1	0.4	0.2	89	0.32	2.6	3.5	1.0	1.0	1.2	0.8	1.1	0.5	0.6	0.1	40	11	73	Transcribed less 5′ proximal, lower expr; Alu repeats
20	0.14	21	0.4	2.6	70	0.39	2.4	2.9	3.5	3.5	3.0	3.4	1.9	20	5.2	4.1	52	9.2	39	Candidate strong enhancer in transcribed regions
21	0.21	28	0.7	3.5	83	0.51	2.8	2.7	7.8	8.1	4.8	8.4	2.4	8.5	4.9	9.1	55	3.9	36	Spliced exons/GC rich; open chr, TF binding
22	0.42	4.4	0.4	1.0	92	0.54	3.3	1.8	8.2	9.2	3.5	4.5	2.2	0.9	1.2	6.4	55	3.5	34	Spliced exons/GC rich
23	0.66	10	1.3	2.0	81	0.29	2.7	1.8	6.1	6.5	4.6	7.5	1.7	1.8	1.4	7.7	54	3.2	51	Spliced exons/GC rich; Alu repeats
24	0.76	1.3	0.3	0.4	90	0.40	2.9	0.8	5.1	5.7	3.3	2.9	1.8	0.8	1.0	0.3	46	13	48	Transcribed 5′ distal; exons
25	1.34	0.6	0.3	0.2	91	0.32	3.6	0.4	6.4	7.3	4.8	5.0	2.0	0.4	0.6	0.1	42	19	52	Transcribed further 5′ distal; exons
26	5.28	1.0	1.0	0.2	86	0.20	2.1	0.7	2.7	3.1	3.1	3.2	1.4	0.4	0.5	0.2	40	17	66	Transcribed 5′ distal; Alu repeats
27	0.39	3.2	0.4	1.0	88	0.92	1.8	0.5	8.6	9.7	8.9	12	2.2	1.5	0.9	0.8	44	6.8	48	End of transcription; exons; high expr
28	0.17	4.2	0.2	1.4	77	0.09	112	3.2	5.4	5.9	5.0	5.9	0.8	0.2	0.1	0.3	41	23	67	ZNF genes; KAP-1 repressed state
29	0.14	11	0.2	1.3	44	0.18	0.8	1.4	1.7	1.8	2.0	1.7	1.5	22	4.7	0.9	49	15	44	Cand strong distal enh; higher open chr; higher target expr
30	0.18	15	0.3	1.9	46	0.33	1.2	1.6	1.7	1.7	2.0	2.8	1.7	18	3.5	1.3	43	19	54	Cand strong distal enh; high open chr; higher target expr
31	0.17	15	0.4	2.2	32	-0.27	0.6	1.1	1.2	1.1	1.1	1.2	1.7	16	4.0	1.3	46	24	47	Intergenic H2AZ with open chr/TF binding. Cand. distal enh
32	0.24	8	0.3	1.1	42	-0.05	0.6	1.2	2.1	2.2	1.8	1.9	1.3	5.5	2.7	0.8	51	15	48	Candidate weak distal enhancer
33	0.34	14	0.6	1.9	51	0.02	1.3	1.8	1.9	1.9	2.0	2.6	1.6	11	3.2	1.6	48	15	54	Candidate distal enhancer
34	0.70	13	0.7	1.0	52	0.20	1.3	1.8	1.7	1.8	2.2	2.9	1.3	4.8	2.1	0.9	44	16	65	Proximal to active enhancers; Alu repeats
35	0.75	5.3	0.8	1.1	34	-0.26	0.5	0.8	1.7	1.8	1.5	1.7	1.3	1.7	2.2	0.4	49	18	52	Active intergenic regions not enhancer specific
36	3.61	4.3	2.7	0.8	33	-0.32	0.5	0.9	1.3	1.3	1.2	1.5	1.1	0.8	1.5	0.5	46	20	66	Active intergenic further from enhancers; Alu repeats
37	11.0	2.6	5.4	0.5	38	-0.38	0.6	1.0	1.0	1.0	1.1	1.1	1.0	0.5	0.9	0.5	42	26	71	Non-repressive intergenic domains; Alu repeats
38	0.84	8.4	1.1	1.3	32	-0.49	0.4	1.0	1.0	0.9	0.8	0.9	1.5	3.6	2.2	0.6	43	31	56	H2AZ specific state
39	0.17	4.1	0.3	1.5	37	-0.32	0.4	0.8	1.5	1.6	1.5	2.2	2.5	19	2.1	2.1	44	26	45	CTCF island; candidate insulator
40	13.2	0.9	4.2	0.3	9.7	-0.41	0.0	0.3	0.3	0.2	0.5	0.4	0.3	0.1	0.1	0.3	40	50	33	Unmappable
41	23.3	0.3	2.6	0.1	22	-0.80	0.4	0.5	0.2	0.2	0.4	0.2	0.8	0.1	0.3	0.1	37	66	66	Heterochr; nuclear lamina; most AT rich
42	1.55	0.5	0.4	0.2	23	-0.77	0.4	0.6	0.2	0.2	0.4	0.2	0.9	0.1	0.5	0.2	39	68	63	Heterochr; nuclear lamina; ERVL repeats
43	22.3	1.3	8.5	0.4	33	-0.69	0.4	0.8	0.6	0.6	0.6	0.6	1.1	0.2	1.0	0.2	41	48	66	Heterochr; lower gene depletion
44	1.95	1.5	1.1	0.5	34	-0.67	0.5	0.8	0.8	0.8	0.7	0.6	1.1	0.3	1.2	0.3	44	47	58	Heterochr; ERVL repeats: lower gene/exon depletion
45	1.59	11	4.2	2.6	34	-0.67	0.5	1.4	1.2	1.0	1.0	1.4	1.7	0.9	1.6	4.9	45	42	54	Specific repression
46	0.09	4.2	0.0	0.6	42	-0.43	0.9	1.0	1.5	1.5	1.0	1.1	0.7	1.2	2.1	2.2	51	24	69	Simple repeats (CA)n, (TG)n
47	1.09	1.6	0.4	0.4	29	-0.53	0.8	0.8	0.4	0.4	0.7	0.8	0.7	0.1	0.5	0.3	38	46	79	L1/LTR repeats
48	0.67	2.7	0.5	0.8	17	-0.22	11	0.7	0.9	0.7	1.6	1.8	0.3	0.2	0.4	1.3	41	57	83	Satellite repeat
49	0.06	1.9	0.1	0.9	8.7	-0.43	4.2	0.3	0.6	0.4	1.6	0.8	0.3	1.8	0.6	1.3	40	61	83	Satellite repeat; moderate mapping bias
50	0.02	0.9	0.0	0.2	5.8	-0.03	1.2	0.1	0.2	0.2	0.4	0.0	0.2	6.6	0.5	1.0	41	53	88	Satellite repeat; high mapping bias
51	0.01	1.1	0.0	0.8	6.1	0.05	0.0	0.1	0.4	0.4	0.0	0	0.5	20	15	3.5	43	52	88	Satellite repeat/rRNA; extreme mapping bias
	100	2.7	100	0.1	36	-0.16	0.5	5.1	3.2	2.8	1.5	0.1	15	2.5	N/A	0.9	41	43	61	Genome total/average

- Promoter
- Transcribed
- Active intergenic
  - Enhancers insulators
- Repressed states
  - Mostly large scale general repression
  - 1 state correspond to specific repression: enriched for TSS

Example of state assignment

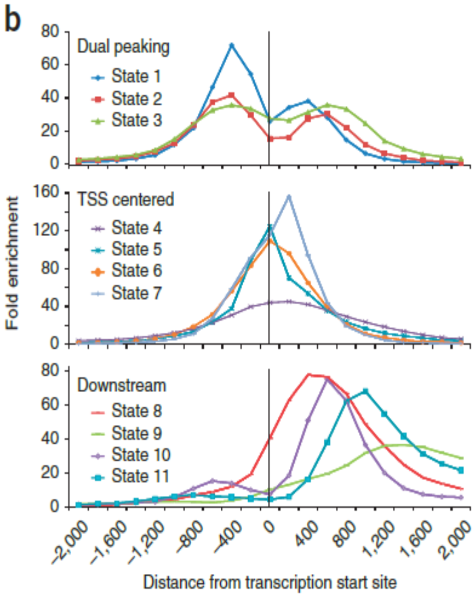


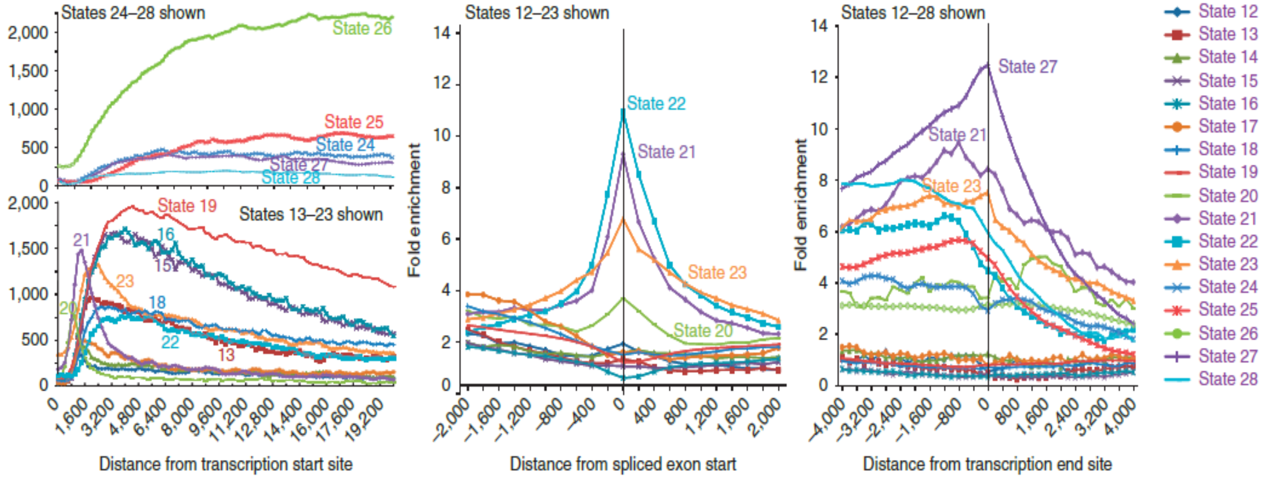
Comparisons with annotations



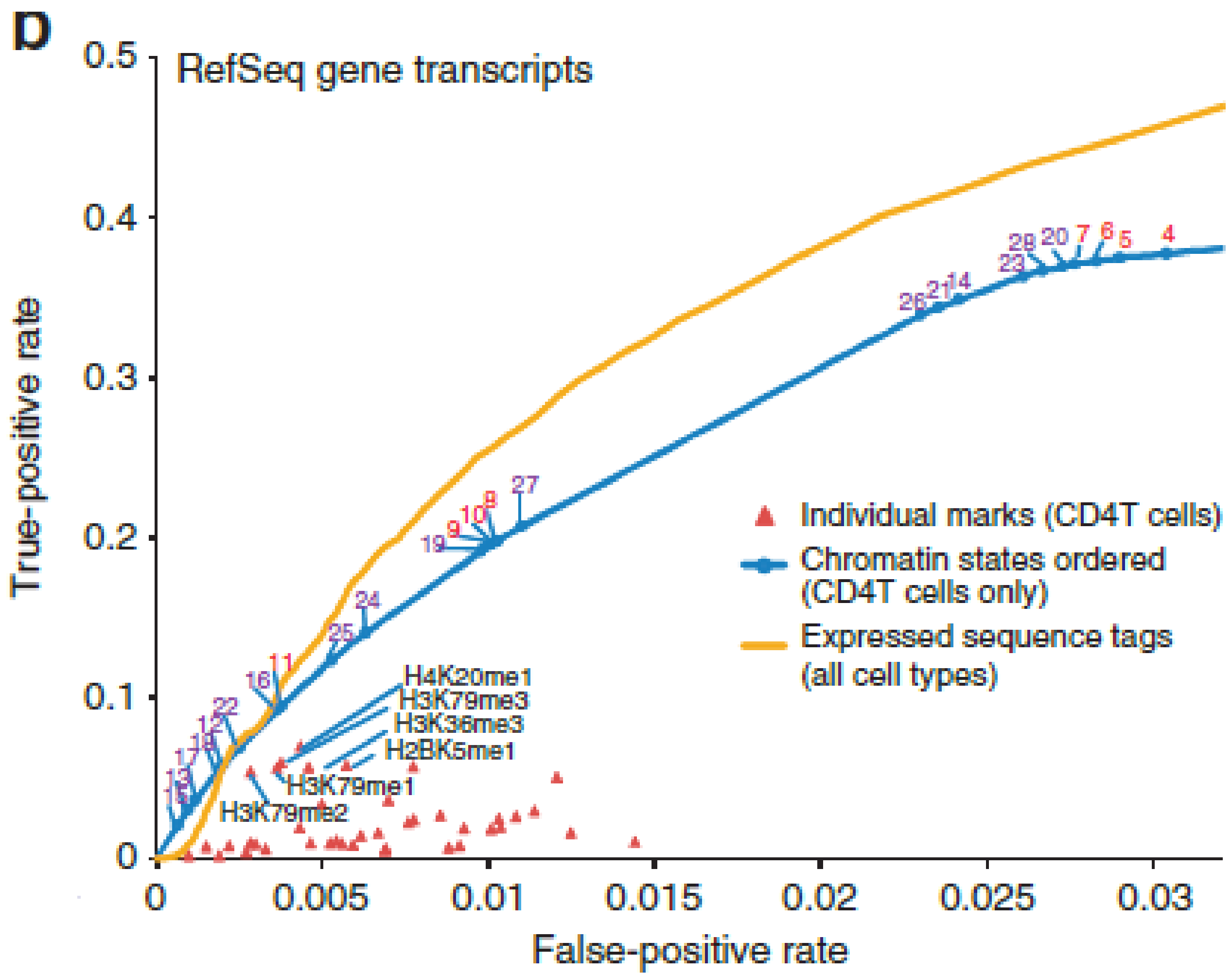
Gene GO category	Chromatin state at TSS of corresponding gene 3	4	5	6	7	8
Cell cycle phase	2.70 (10^{-7})	0.57 (1.0)	1.61 (10^{-3})	1.45 (1.0)	1.15 (1.0)	1.51 (1.0)
Embryonic development	1.24 (1.0)	2.82 (10^{-22})	1.07 (1.0)	0.85 (1.0)	0.54 (1.0)	1.00 (1.0)
Chromatin	1.20 (1.0)	0.48 (1.0)	2.17 (10^{-7})	1.64 (1.0)	0.85 (1.0)	0.85 (1.0)
Response to DNA damage	1.20 (1.0)	0.35 (1.0)	1.55 (0.07)	2.13 (10^{-11})	1.97 (10^{-4})	0.84 (1.0)
RNA processing	0.49 (1.0)	0.26 (1.0)	1.31 (1.0)	1.91 (10^{-11})	2.64 (10^{-24})	2.46 (10^{-4})
T-cell activation	0.77 (1.0)	0.88 (1.0)	1.27 (1.0)	0.70 (1.0)	0.79 (1.0)	4.72 (10^{-7})

b





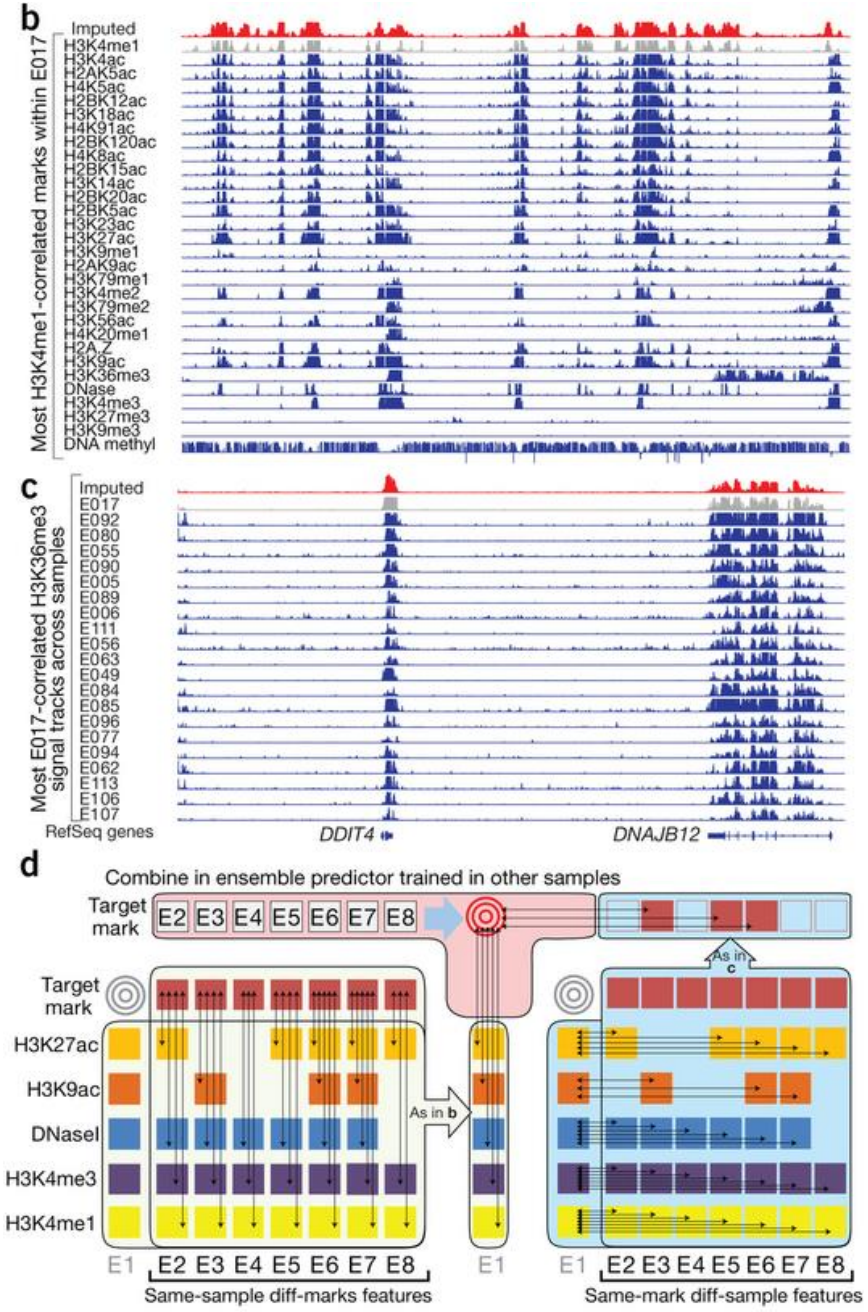
Many marks are correlated but combinatorial organization still matters



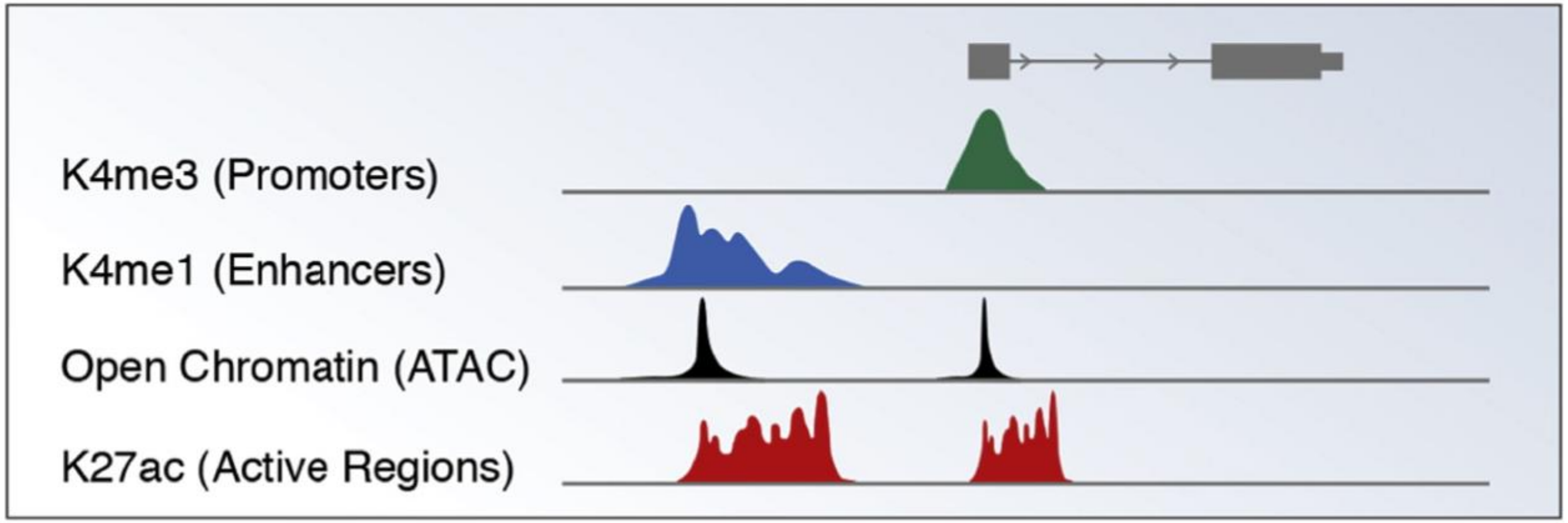
- We can predict novel transcripts from histone mark patterns
- Histone patterns are cell type specific but gene prediction is not!

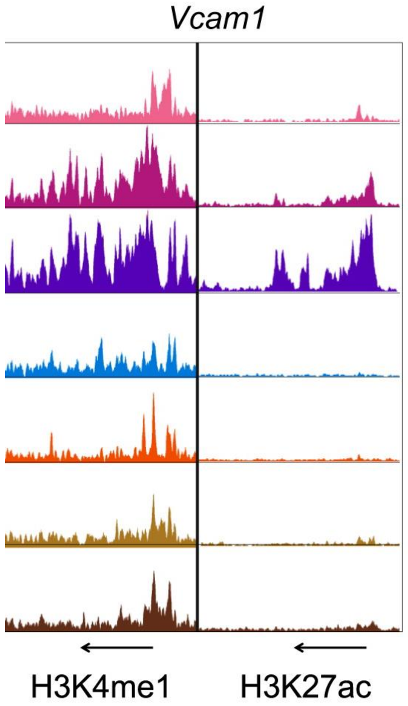
ChromImpute

- Chromatin marks are correlated with each other and also across samples
- We don’t have to measure everything
- Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Jason Ernst& Manolis Kellis *Nature Biotechnology*
- Prediction methodology—regression trees
- Coninuous analog of decision trees for predicting a continuous variable

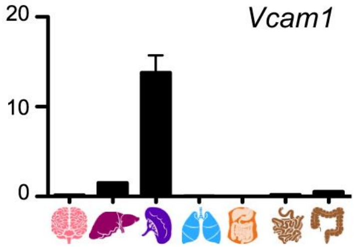


Commonly used chromatin measurements





- Since chromatin profiles are highly correlated we can measure just a few marks to get a comprehensive picture



LONG NON-CODING RNA

Non-coding RNAs

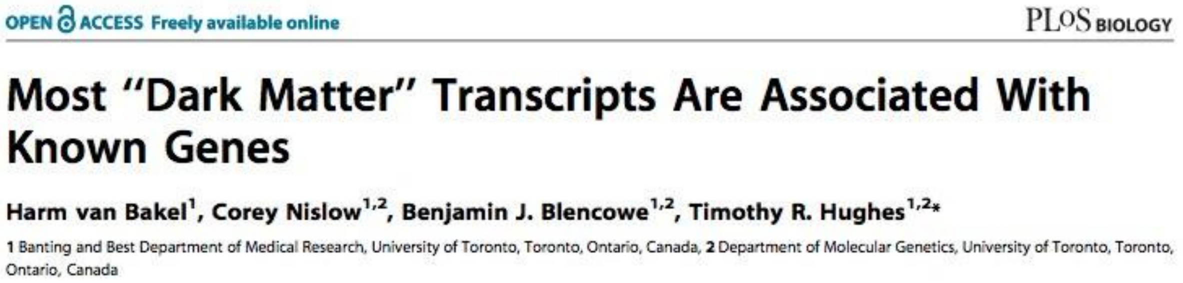
- A **non-coding RNA** (ncRNA) is a functional **RNA** molecule that is transcribed from DNA but not translated into proteins
- Infrastructure/house-keeping
  - Ribosome , spliceosome, transfer RNAs
  - snoRNAs (small nucleolar RNAs) guide chemical modifications to above
- Regulatory
  - **microRNAs-**pair with complementary sequence in the UTR of coding genes to induce gene down-regulation or silencing. Processed from longer genes into small final products **Small interfering RNAs (siRNAs)**
  - **Small interfering RNAs (siRNAs)**-can also refer to a synthetic product meant to coopt the miRNA mechanism
  - **Piwi-interacting RNAs (piRNAs)**-involved in silencing of transposable elements in the germ-line
  - **Long non-coding RNAs (lncRNAs)** lncRNAs are considered as non-protein coding transcripts >200 nt in length. The majority of non-coding RNAs belong to this group. Many RNAs in the group are treated by the cell as coding genes, they have exons, are spliced and have polyA tails

Table 1. Classes of non-coding RNA in mammals

NcRNA class	Characteristics
Established ncRNA classes	
Long (regulatory) non-coding RNAs (lncRNAs)	The broadest class, lncRNAs, encompass all non-protein-coding RNA species >~200 nt, including mRNA-like ncRNAs. Their functions include epigenetic regulation, acting as sequence-specific tethers for protein complexes and specifying subcellular compartments or localization
Small interfering RNAs (siRNAs)	Small RNAs ~21–22 nt long, produced by Dicer cleavage of complementary dsRNA duplexes. siRNAs form complexes with Argonaute proteins and are involved in gene regulation, transposon control and viral defence
microRNAs (miRNAs)	Small RNAs ~22 nt long, produced by Dicer cleavage of imperfect RNA hairpins encoded in long primary transcripts or short introns. They associate with Argonaute proteins and are primarily involved in post-transcriptional gene regulation
PIWI-interacting RNAs (piRNAs)	Dicer-independent small RNAs ~26–30 nt long, principally restricted to the germline and somatic cells bordering the germline. They associate with PIWI-clade Argonaute proteins and regulate transposon activity and chromatin state
Promoter-associated RNAs (PARs)	A general term encompassing a suite of long and short RNAs, including promoter-associated RNAs (PASRs) and transcription initiation RNAs (tiRNAs) that overlap promoters and TSSs. These transcripts may regulate gene expression
Small nucleolar RNAs (snoRNAs)	Traditionally viewed as guides of rRNA methylation and pseudouridylation. However, there is emerging evidence that they also have gene-regulatory roles
Other recently described classes	
X-inactivation RNAs (xiRNAs)	Dicer-dependent small RNAs processed from duplexes of two lncRNAs, Xist and Tsix, which are responsible for X-chromosome inactivation in placental mammals
Sno-derived RNAs (sdRNAs)	Small RNAs, some of which are Dicer-dependent, which are processed from small nucleolar RNAs (snoRNAs). Some sdRNAs have been shown to function as miRNA-like regulators of translation
microRNA-offset RNAs (moRNAs)	Small RNAs ~20 nt long, derived from the regions adjacent to pre-miRNAs. Their function is unknown
tRNA-derived RNAs	tRNAs can be processed into small RNA species by a conserved RNase (angiogenin). They are able to induce translational repression
MSY2-associated RNAs (MSY-RNAs)	MSY-RNAs are associated with the germ cell-specific DNA/RNA binding protein MSY2. Like piRNAs, they are largely restricted to the germline and are ~26–30 nt long. Their function is unknown
Telomere small RNAs (tel-sRNAs)	Dicer-independent ~24 nt RNAs principally derived from the G-rich strand of telomeric repeats. May have a role in telomere maintenance
Centrosome-associated RNAs (crasiRNAs)	A class of ~34–42 nt small RNAs, derived from centrosomes that show evidence of guiding local chromatin modifications

Pervasive transcription

- 1.2% of the mammalian genome codes for amino acids in proteins.
- evidence over the past decade has suggested that the vast majority of the genome is transcribed, well beyond the boundaries of known genes -- **pervasive transcription**
- Functionality has to be demonstrated via a phenotype

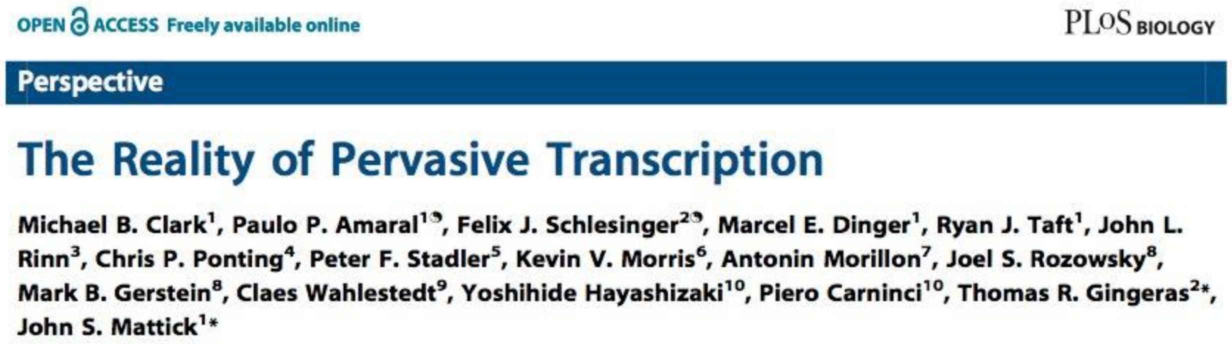


OPEN ACCESS Freely available online PLoS BIOLOGY

Most "Dark Matter" Transcripts Are Associated With Known Genes

Harm van Bakel[1], Corey Nislow[1,2], Benjamin J. Blencowe[1,2], Timothy R. Hughes[1,2]*

1 Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada, 2 Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

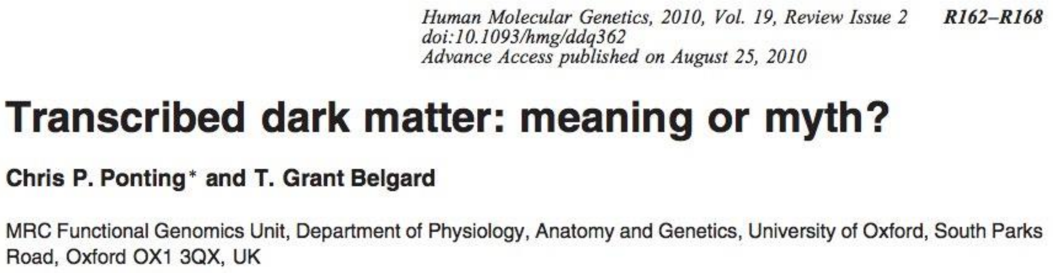


OPEN ACCESS Freely available online PLoS BIOLOGY

Perspective

The Reality of Pervasive Transcription

Michael B. Clark[1], Paulo P. Amaral[1,9], Felix J. Schlesinger[2,9], Marcel E. Dinger[1], Ryan J. Taft[1], John L. Rinn[3], Chris P. Ponting[4], Peter F. Stadler[5], Kevin V. Morris[6], Antonin Morillon[7], Joel S. Rozowsky[8], Mark B. Gerstein[8], Claes Wahlestedt[9], Yoshihide Hayashizaki[10], Piero Carninci[10], Thomas R. Gingeras[2]*, John S. Mattick[1]*



Human Molecular Genetics, 2010, Vol. 19, Review Issue 2 ***R162–R168***
doi:10.1093/hmg/ddq362
Advance Access published on August 25, 2010

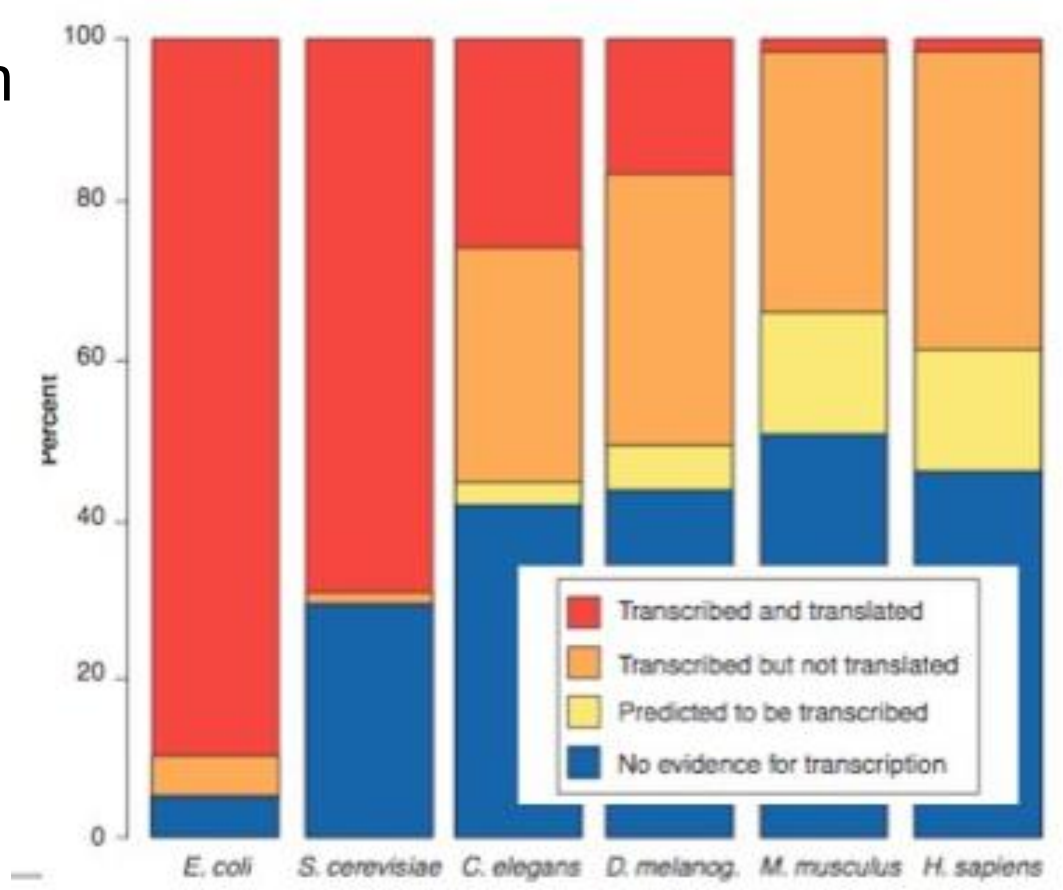
Transcribed dark matter: meaning or myth?

Chris P. Ponting* and T. Grant Belgard

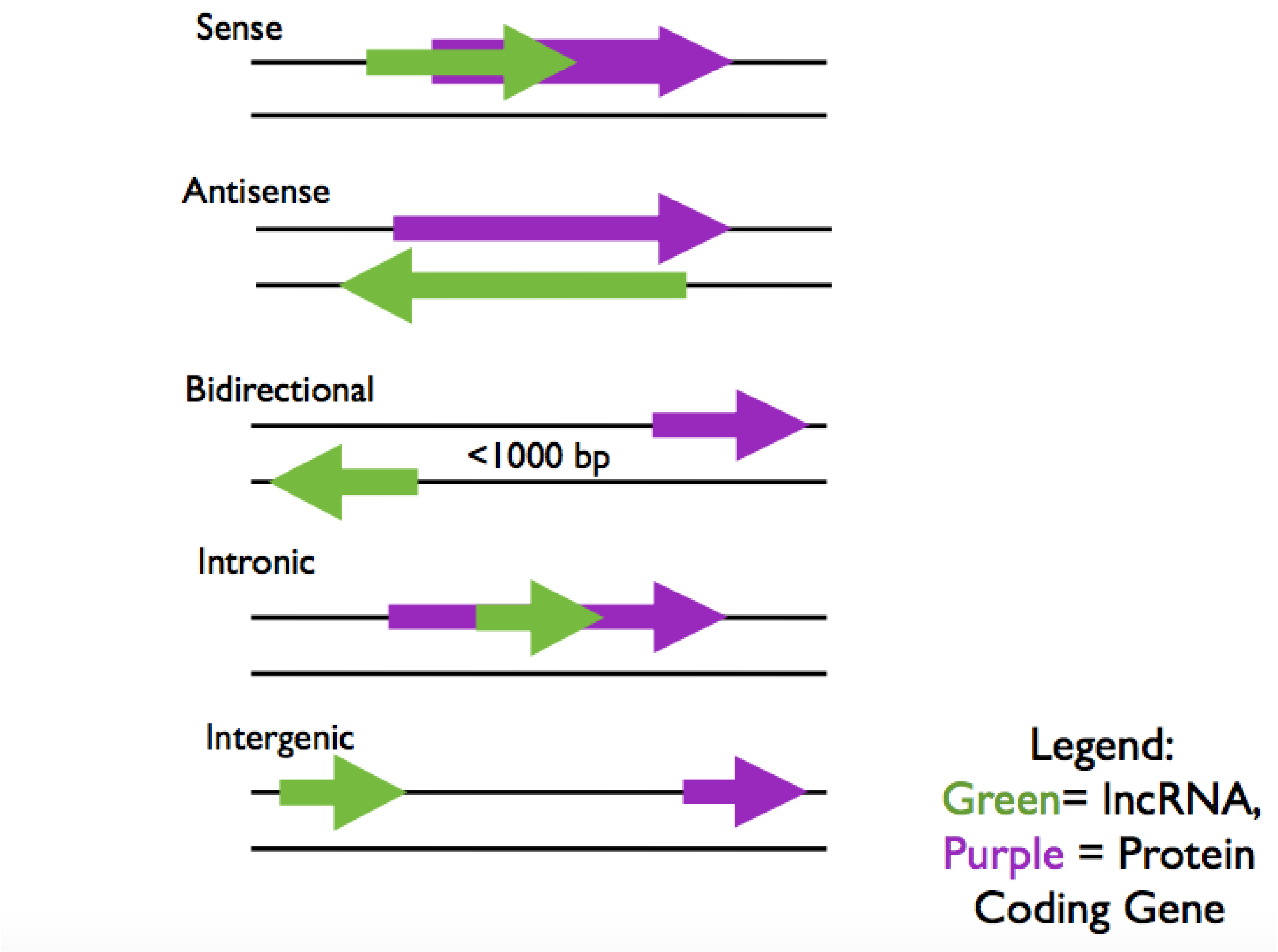
MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, UK

Long non-coding RNA

- 80% of the transcription in mammalian genomes is exclusively associated with long non-coding RNAs (lncRNAs)
- >2 (some >100) kb in length, spliced and could contain polyA signals
- No obvious open reading frame—can't get a long protein sequence without hitting a stop codon
- Mouse transcriptome (~180,000)
  - ~20,000 protein coding genes
  - ~160,000 lncRNAs

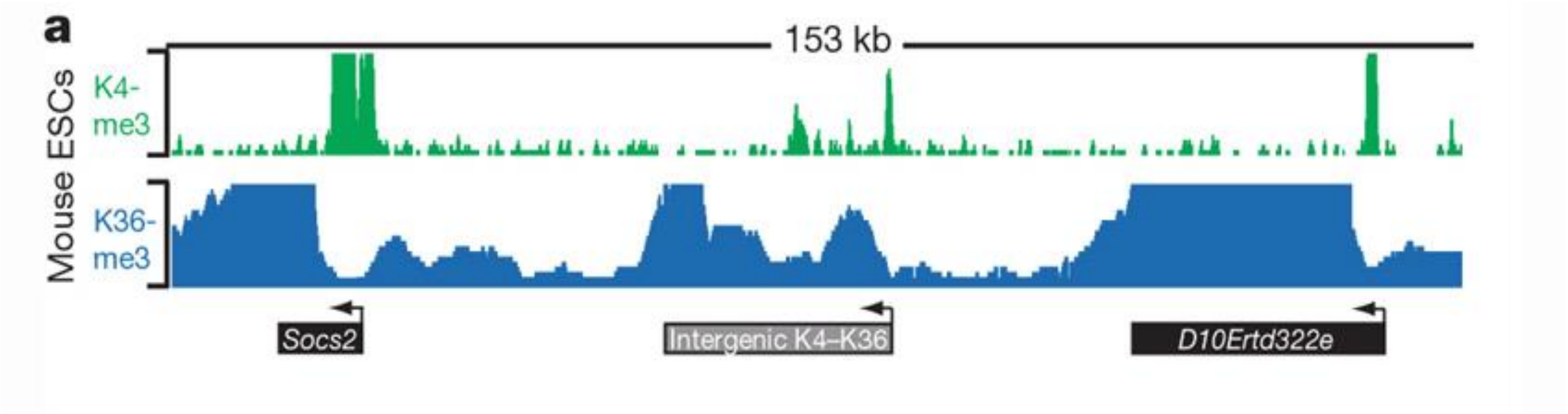


Coding gene proximity categorization

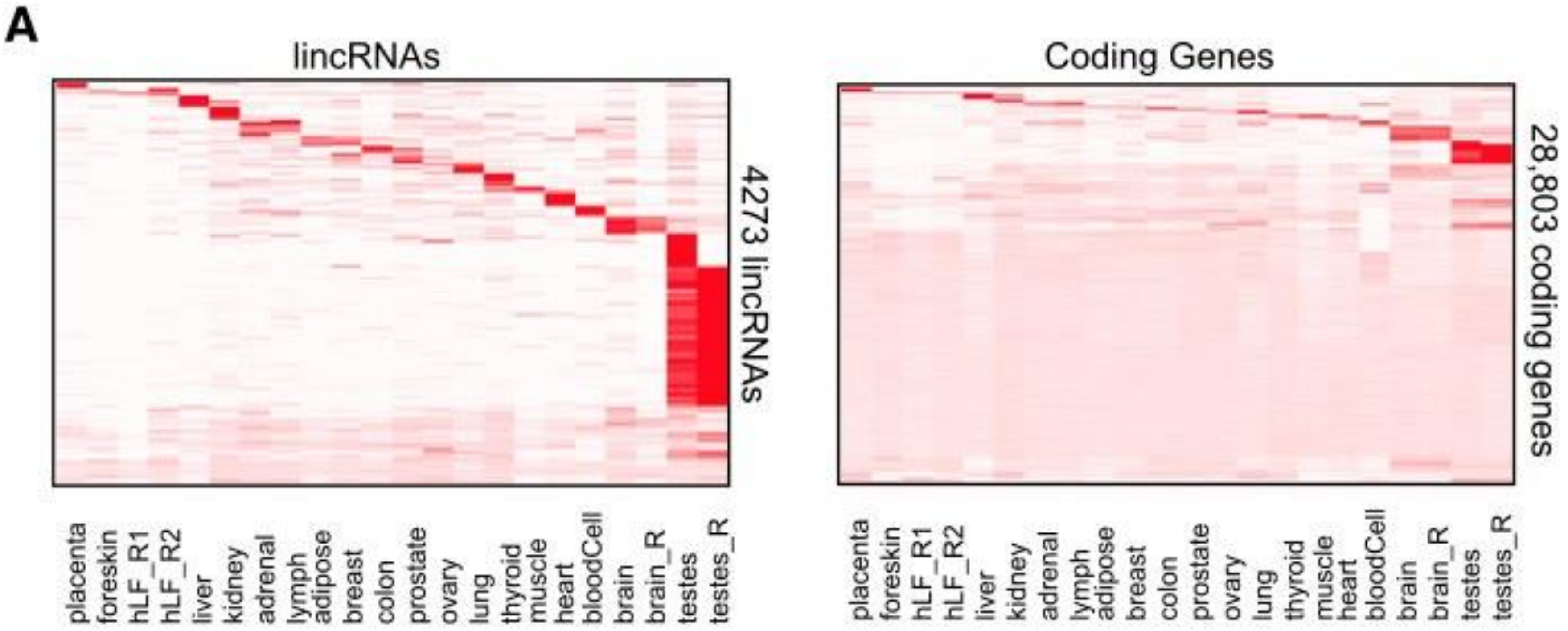


Non-coding RNAs have an epigenetic profile similar to coding genes

Guttman M, Amit I, Garber M, French C, Lin M, Feldser D, Huarte M, Cabili M, Carey BW, Cassady J, Jaenisch R, Mikkelsen T, Jacks T, Hacohen N, Bernstein BEB, Kellis M, Regev A, Rinn JL, Lander ES. (2009) Chromatin Structure Reveals Over a Thousand Highly Conserved, Large Non-coding RNAs in Mammals. Nature. 458(7235):223-7. PMCID:



Cell and tissue specific expression



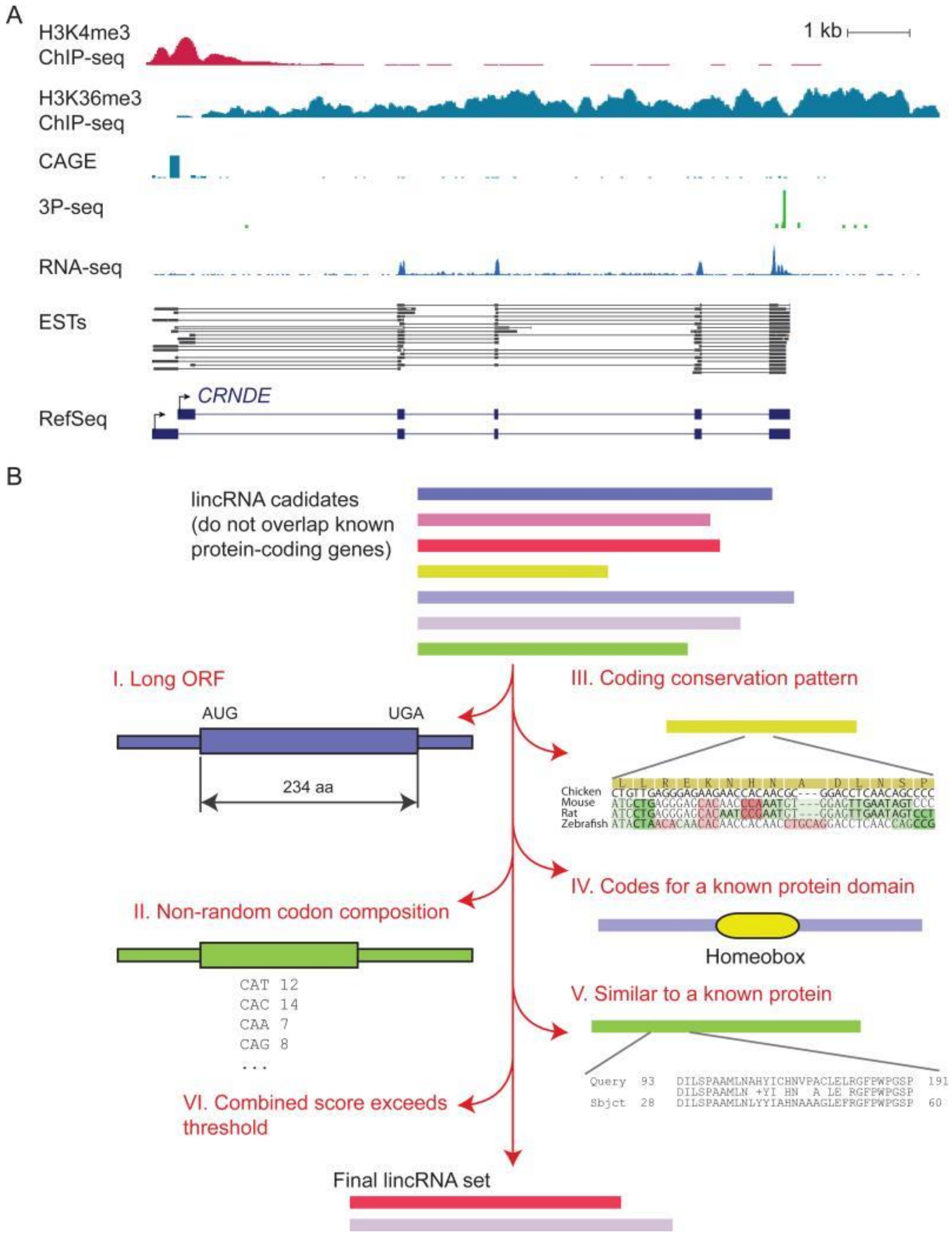
Cabili MN. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011 Sep 15;25(18):1915-27

Functionality controversy

- **Transcriptional Noise**
  - Low affinity binding of RNA polymerase to randomly generated promoter sequences.
  - More efficient to allow random transcripts than to downregulate nonspecific transcription.
  - LncRNAs are generally expressed at low levels
  - LncRNA sequences are not well conserved between species.
  - Sequencing with splicing/polyadenylation signals can occur by chance-regional chromatin state would direct tissue specific transcription
- **LncRNAs are Functional**
  - LncRNAs do not have the strict sequence conservation restraints that protein-coding genes do.
  - LncRNAs may be more plastic then protein coding genes and thus can evolve rapidly.
  - LncRNA promoter sequences are very well conserved.
- General consensus: some are functional and some are not but disagreement over relative frequency

lincRNA annotation

- Chromatin profiles
- ESTs-expressed sequence tag-short cDNA sequences—older technology used in gene discovery
- Cap analysis gene expression (CAGE) snapshot of the 5′ end of the messenger RNA



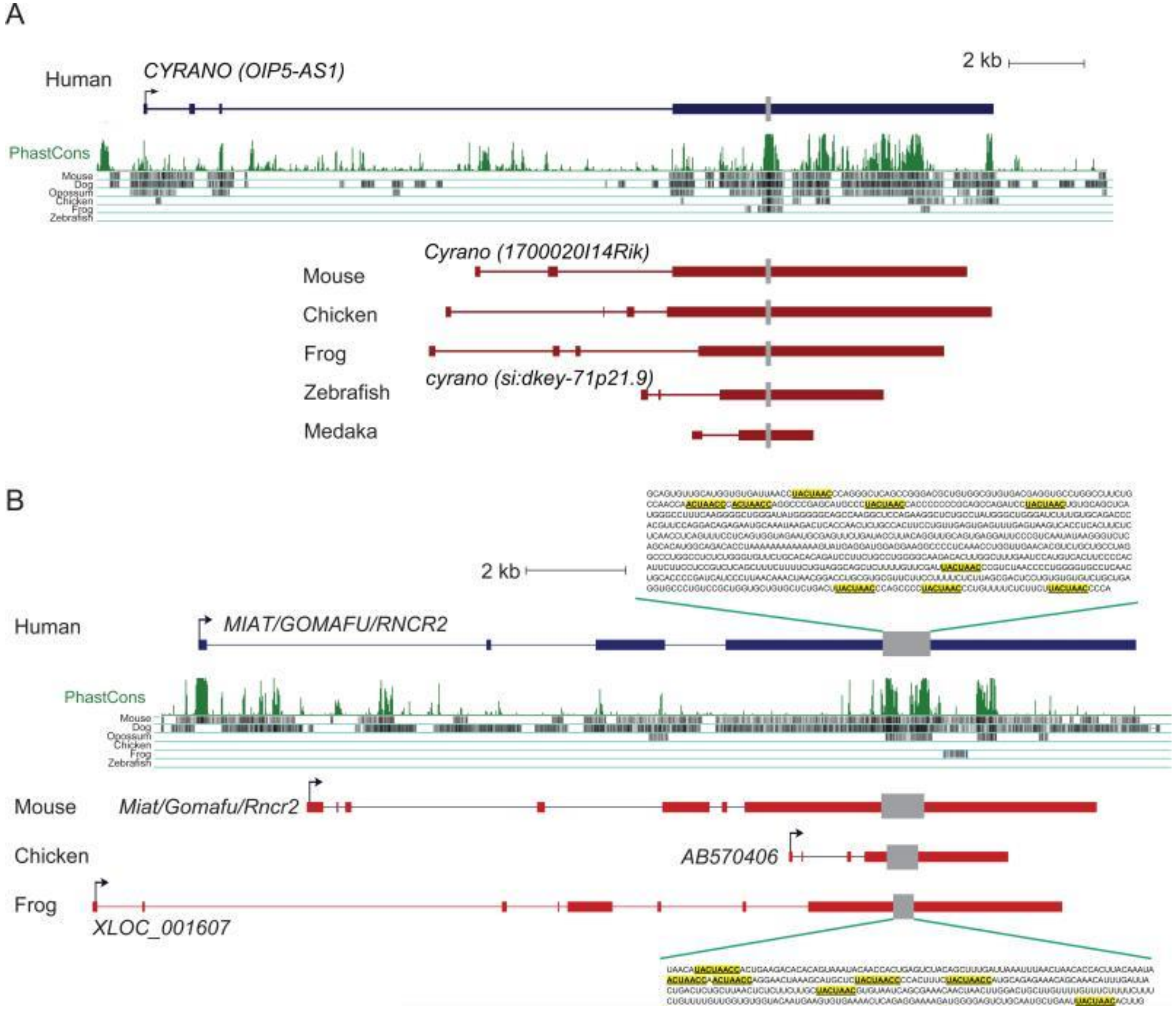
```
           L  L  R  E  K  N  H  N  A     D  L  N  S  P
Chicken    CTGTTGAGGGAGAAGAACCACAACGC---GGACCTCAACAGCCCC
Mouse      ATGCTGAGGGAGCACAACCCAAATGT---GGAGTTGAATAGTCCC
Rat        ATCCTGAGGGAGCACAATCCGAATGT---GGAGTTGAATAGTCCT
Zebrafish  ATACTAACACAACACAACCACAACCTGCAGGACCTCAACCAGCCG
```

```
CAT 12
CAC 14
CAA 7
CAG 8
...
```

```
Query  93   DILSPAAMLNAHYICHNVPACLELRGFPWPGSP  191
            DILSPAAMLN +YI HN  A LE RGFPWPGSP
Sbjct  28   DILSPAAMLNLYYIAHNAAAGLEFRGFPWPGSP  60
```

What about conservation?

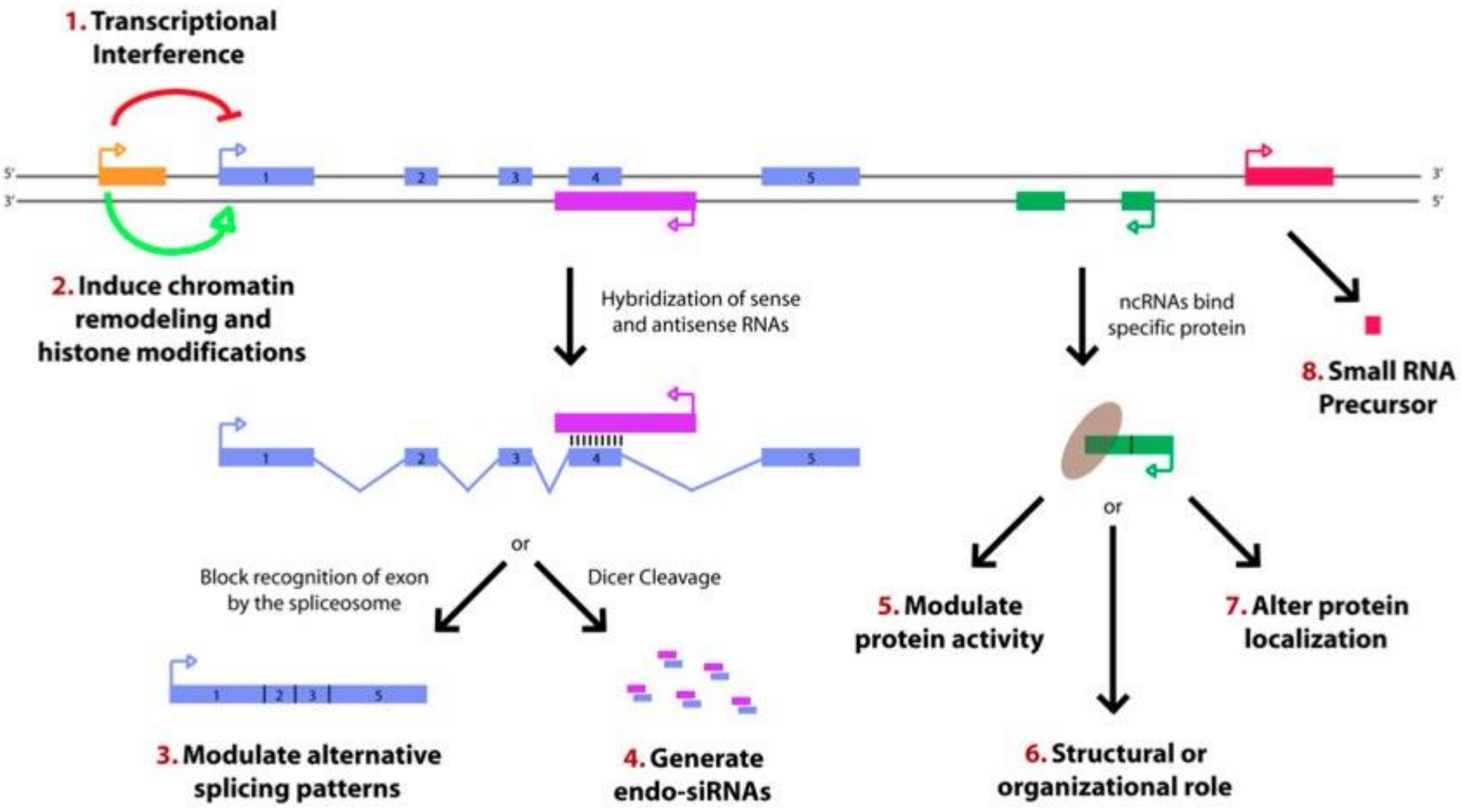
- Xist is a functional lincRNA with poor sequence conservation but significant exon structure conservation
- The *Cyrano* lincRNA is
  - conserved in vertebrates
  - required for proper morphogenesis and neurogenesis in zebrafish
- *Megamind*
  - conserved in vertebrates
  - required for proper brain development in zebrafish
  - No sequence level conservation
  - Cannot be identified via blastn alignment –but can be identified with HMM



lincRNA functional mechanisms

- Cis-acts on nearby gene only, depends on the site of transcription
  - Requires transcription only
  - Requires a processed transcript
- Trans-acts elsewhere in the genome
  - Does not depend on the site of transcription

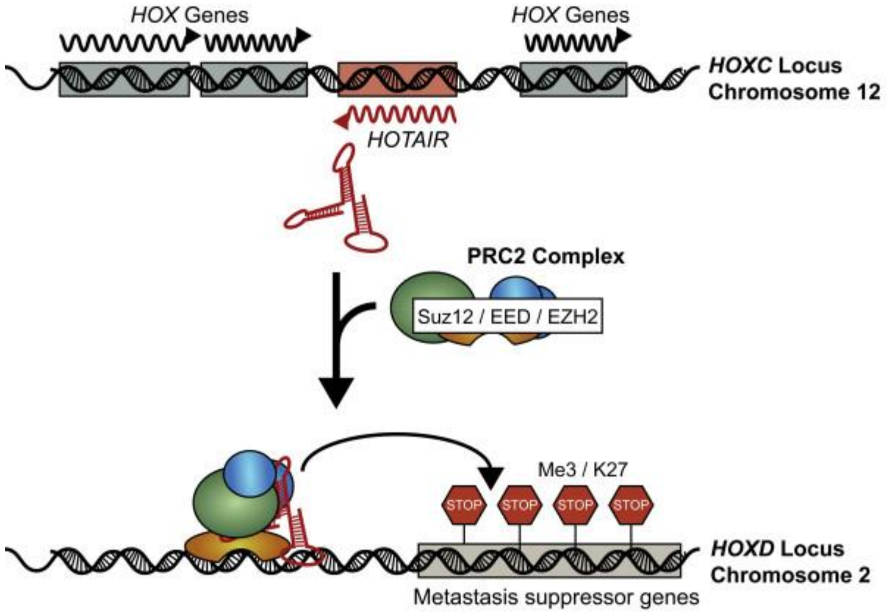
Potential functions of lncRNA


1. Transcriptional Interference
5'
3'
1
2
3
4
5
3'
5'
2. Induce chromatin remodeling and histone modifications
Hybridization of sense and antisense RNAs
ncRNAs bind specific protein
8. Small RNA Precursor
or
Block recognition of exon by the spliceosome
Dicer Cleavage
or
5. Modulate protein activity
7. Alter protein localization
3. Modulate alternative splicing patterns
4. Generate endo-siRNAs
6. Structural or organizational role

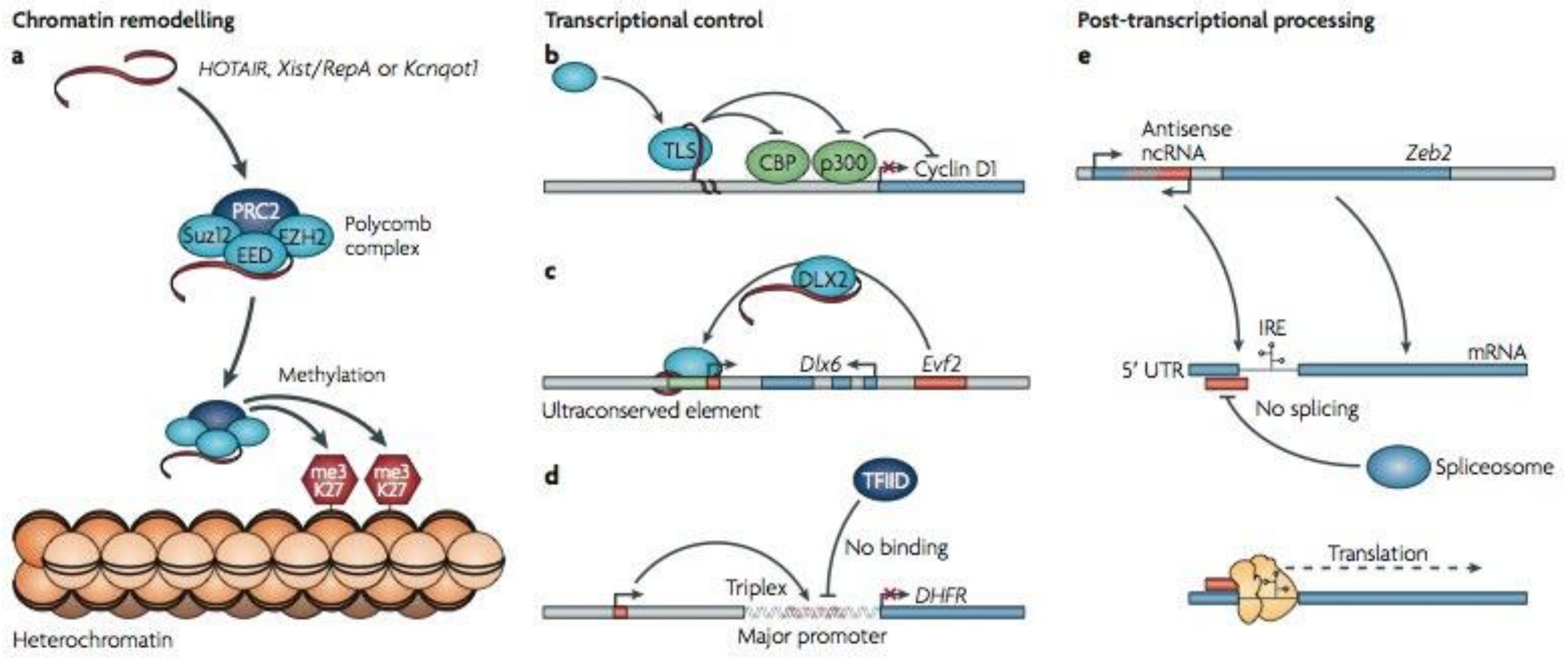
Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. 2009 23(13):1494-504.

Example: HOTAIR

- HOTAIR (for HOX transcript antisense RNA)is first example of an RNA expressed on one chromosome that has been found to influence transcription on another chromosome
- It is required for gene-silencing of the HOXD locus
- It is hypothesized to be important for epigenetic differentiation of skin over the surface of the body.
- HOTAIR was shown to contain distinct protein interaction domains that can associate with polycomb repressive complex 2 (PRC2) and the CoREST–LSD1 complex64, which together are required for correct function

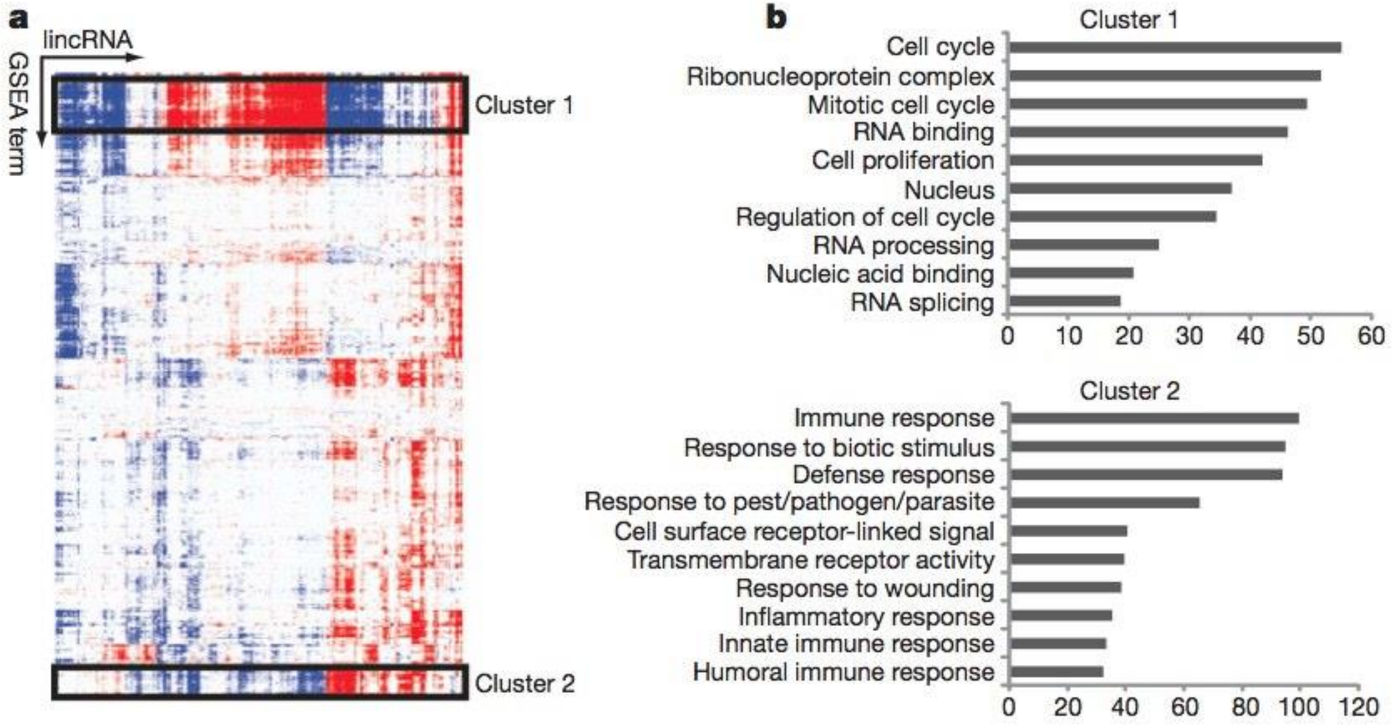


Known Examples



Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet. 2009 Mar;10(3):155-9

Functional assignment based on gene expression correlation

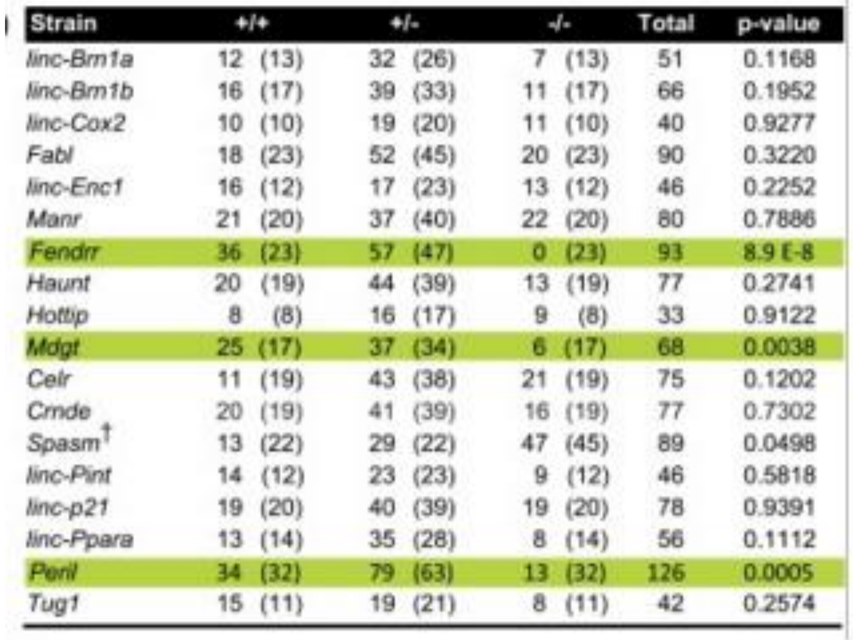


Guttman M. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009 458(7235):223-7

Knock-out studies

- Selected 18 lincRNAs beased on
  - Conservation
  - Chromatin features
  - Low protein-coding potential
    - Sequence based filter
    - Mass-spec based filter
- postnatal lethal phenotypes in three mutant strains (*Fendrr*, *Peril*, and *Mdgt*), the latter two exhibiting incomplete penetrance and growth defects in survivors
- growth defects for two additional mutant strains (*linc–Brn1b* and *linc–Pint*)

Multiple knockout mouse models reveal lincRNAs are required for life and brain development



Strain	+/+	+/-	-/-	Total	p-value
linc-Brn1a	12 (13)	32 (26)	7 (13)	51	0.1168
linc-Brn1b	16 (17)	39 (33)	11 (17)	66	0.1952
linc-Cox2	10 (10)	19 (20)	11 (10)	40	0.9277
Fabl	18 (23)	52 (45)	20 (23)	90	0.3220
linc-Enc1	16 (12)	17 (23)	13 (12)	46	0.2252
Manr	21 (20)	37 (40)	22 (20)	80	0.7886
Fendrr	36 (23)	57 (47)	0 (23)	93	8.9 E-8
Haunt	20 (19)	44 (39)	13 (19)	77	0.2741
Hottip	8 (8)	16 (17)	9 (8)	33	0.9122
Mdgt	25 (17)	37 (34)	6 (17)	68	0.0038
Celr	11 (19)	43 (38)	21 (19)	75	0.1202
Crnde	20 (19)	41 (39)	16 (19)	77	0.7302
Spasm†	13 (22)	29 (22)	47 (45)	89	0.0498
linc-Pint	14 (12)	23 (23)	9 (12)	46	0.5818
linc-p21	19 (20)	40 (39)	19 (20)	78	0.9391
linc-Ppara	13 (14)	35 (28)	8 (14)	56	0.1112
Peril	34 (32)	79 (63)	13 (32)	126	0.0005
Tug1	15 (11)	19 (21)	8 (11)	42	0.2574

lincRNA summary

- Many are not functional but some are
- Sequence conservation is poor but we can look for
    - Small conserved regions
    - Promoter conservation
    - Exon-intron structure
    - Synteny
    - Non-alignable conserved feature
- lincRNAs most likely come from different classes that differ
    - Functionality
    - Mechanism
    - Gene proximity
    - Conservation