

Lecture 13. Functional Genomics 2

Michael Schatz

March 16, 2017

JHU 600.649: Applied Comparative Genomics



Assignment 2

Due: Thursday March 16 @ 11:59pm

The screenshot shows a GitHub repository page for 'appliedgenomics/assignments'. The repository has 6 stars, 8 forks, and 0 issues. The README.md file contains the assignment details:

Assignment 2: Variant Analysis

Assignment Date: Tuesday, March 7, 2017
Due Date: Tuesday, March 14, 2017 @ 11:59pm

Assignment Overview

In this assignment, you will identify variants in a human genome and then analyze the properties for them. Make sure to show your work in your writeup! As before, any questions about the assignment should be posted to [Piazza](#).

Some of the tools you will need to use only run in a linux environment. If you do not have access to a linux machine, download and install a virtual machine following the directions here: <https://github.com/schatzlab/appliedgenomics/blob/master/assignments/virtualbox.md>

Question 1. Gene Annotation Preliminaries [10 pts]

Download the annotation of build 38 of the human genome from here: ftp://ftp.ensembl.org/pub/release-87/gtf/homo_sapiens/Homo_sapiens.GRCh38.87.gtf.gz

- Question 1a. How many GTF data lines are in this file? [Hint: The first few lines in the file beginning with "#" are so-called "header" lines describing things like the creation date, the genome version (more on that later in the course), etc. Header lines should not be counted as data lines.]
- Question 1b. How many annotated protein coding genes are on each chromosome of the human genome? [Hint: Protein coding genes will contain the following text: transcript_biotype "nonsense_mediated_decay"]
- Question 1c. What is the maximum, minimum, mean, and standard deviation of the span of protein coding genes?
- Question 1d. What is the maximum, minimum, mean, and standard deviation in the number of exons for protein coding genes? [Hint: you should separately consider each isoform]

Question 2. Genome Sequence Analysis [10 pts]

Download chromosome 22 from build 38 of the human genome from here: <http://hgdownload.cse.ucsc.edu/goldenPath/hg38/chromosomes/chr22.fa.gz>

- Question 2a. What is the length of chromosome 22? [Hint: You should include Ns in the length]
- Question 2b. How many Ns are in chromosome 22? What is the GC content? [Hint: You should exclude Ns when computing GC content]
- Question 2c. Restriction enzymes cleave DNA molecules at or near a specific sequence of bases. For example, the HindIII enzyme cuts at the "A" in either this motif: 5'-A/AGCTT-3' or its reverse complement, 3'-TTCTGA/A-5'. How many perfectly matching HindIII restriction site cut sites are there on chr22?
- Question 2d. How many HindIII cut sites are there on chr22, assuming that a mutant form of HindIII will tolerate a mismatch in the second position? Think about ways in which you could best test for all the possible DNA combinations. [Hint: There are many valid approaches]

Question 3. Small Variant Analysis [10 pts]

Download the read set from here: <https://github.com/schatzlab/appliedgenomics/blob/master/assignments/assignment2/sample.tgz>

For this question, you may find this tutorial helpful: <http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html>

- Question 3a. How many single nucleotide and indel variants does the sample have? [Hint: Align reads using `bwa mem`, identify variants using `freebayes`, filter using `vcffilter -f "QUAL > 20"`, and summarize using `vcfstats`]
- Question 3b. How many of the variants fall into genes? How many fall into exons? [Hint: `bedtools`]
- Question 3c. What is the transition/transversion ratio of the variants in protein coding genes? What is the ratio of variants in the exons? [Hint: try `bedtools` and `vcfstats`]
- Question 3d. Does the sample have any 'nonsense' or 'missense' mutations? [Hint: try the Variant Effect Predictor using the Gencode basic transcripts]

Question 4. Structural Variation Analysis [10 pts]

For this question, you should use the same reads and `bwa mem` alignments as question 3.

- Question 4a. Plot the copy number status of the sample across the chromosome divided into 10kb bins [Hint: your plot should show how many reads align to bases 1-10k, 10k-20k, 20k-30, etc]

Project Proposal

Due: Thursday March 30 @ 11:59pm

Inbox - michael.schatz@gmail.com appliedgenomics/projectpropo appliedgenomics/projectpropo Michael

GitHub, Inc. [US] https://github.com/schatzlab/appliedgenomics/blob/master/assignments/projects/projectproposal.md

JHUMail Daily schatzlab SL cshl jhu Media edit Remove NYT Cookies Other Bookmarks

This repository Search Pull requests Issues Gist

schatzlab / appliedgenomics

Unwatch 6 Unstar 9 Fork 0

Branch: master appliedgenomics / assignments / projects / projectproposal.md Find file Copy path

mschatz Update projectproposal.md 1310549 4 minutes ago

1 contributor

24 lines (16 sloc) 1.46 KB Raw Blame History

Project Proposal

Assignment Date: March 16, 2017
Due Date: Thursday, March 30, 2017 @ 11:59pm

Review the [Project Ideas](#) page

Form a team for your class project (no more than 3 people to a team) and email a PDF of your project proposal (1/2 to 1 page) to "jhuappliedgenomics at gmail dot com" by 11:59pm on March 30.

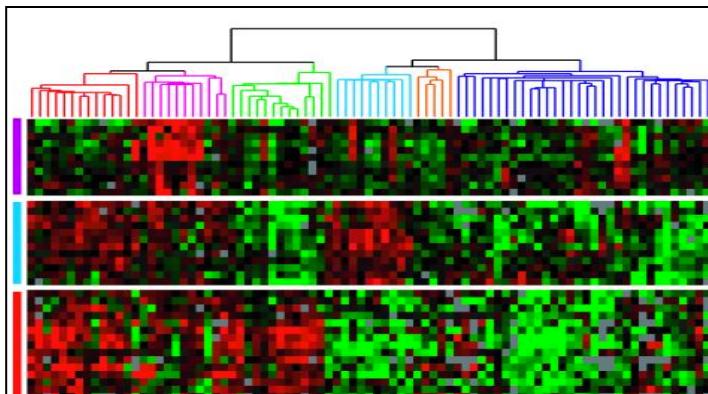
The proposal should have the following components:

- Name of your team
- List of team members and email addresses
- Short title for your proposal
- 1 paragraph description of what you hope to do and how you will do it
- References to relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)

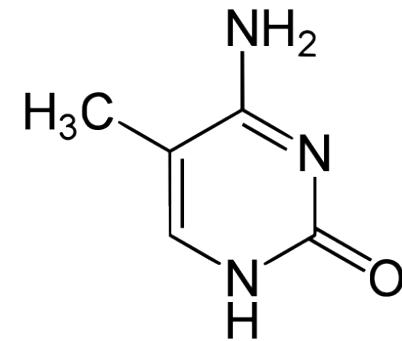
After submitting your proposal, we will schedule a time to discuss your proposal, especially to ensure you have access to the data that you need. The sooner that you submit your proposal, the sooner we can schedule the meeting. No late days can be used for this project.

*-seq in 4 short vignettes

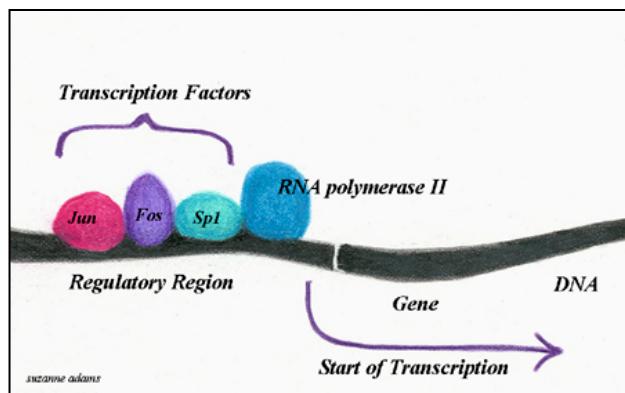
RNA-seq



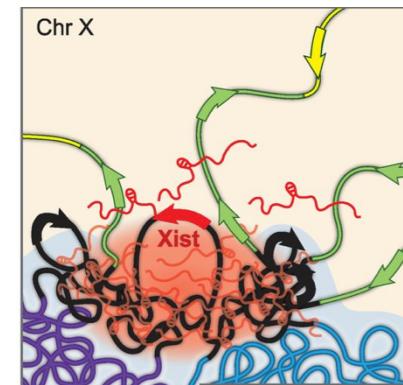
Methyl-seq



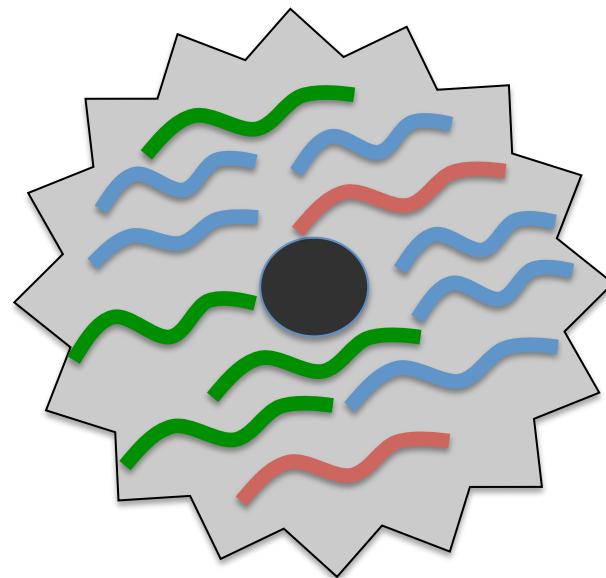
ChIP-seq



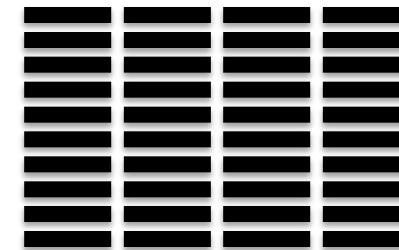
Hi-C



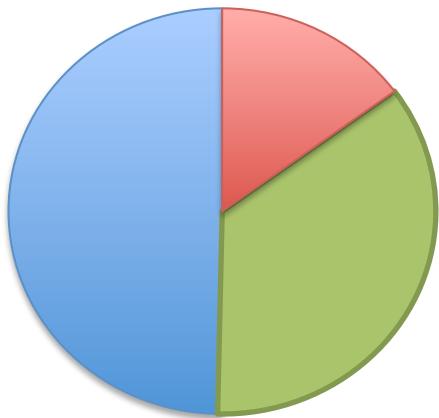
RNA-seq Overview



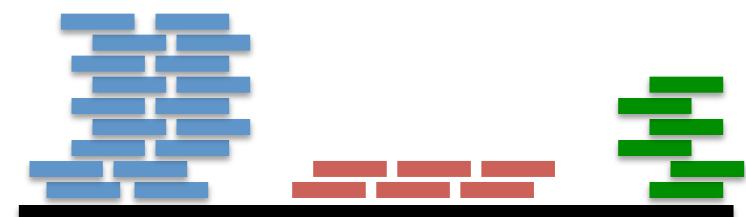
Sequencing



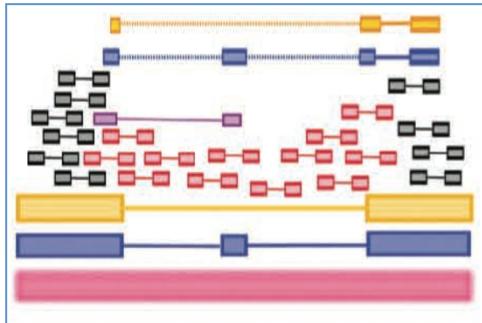
Mapping & Assembly



Quantification



RNA-seq Challenges

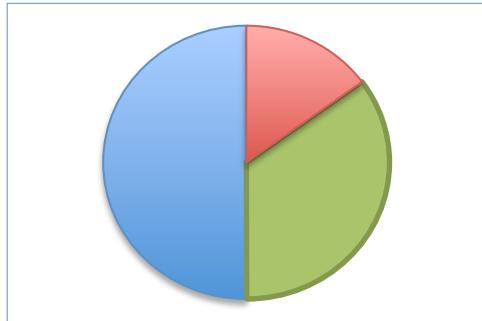


Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

TopHat: discovering spliced junctions with RNA-Seq.

Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111

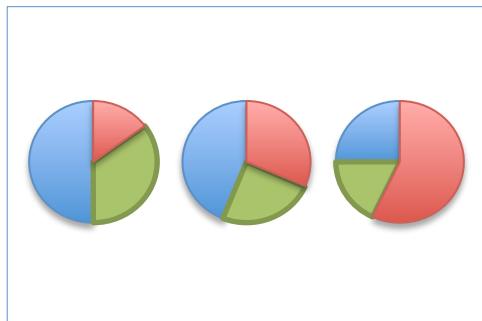


Challenge 2: Read Count != Transcript abundance

Solution: Infer underlying abundances (e.g. TPM)

Transcript assembly and quantification by RNA-seq

Trapnell et al (2010) *Nat. Biotech.* 25(5): 511-515



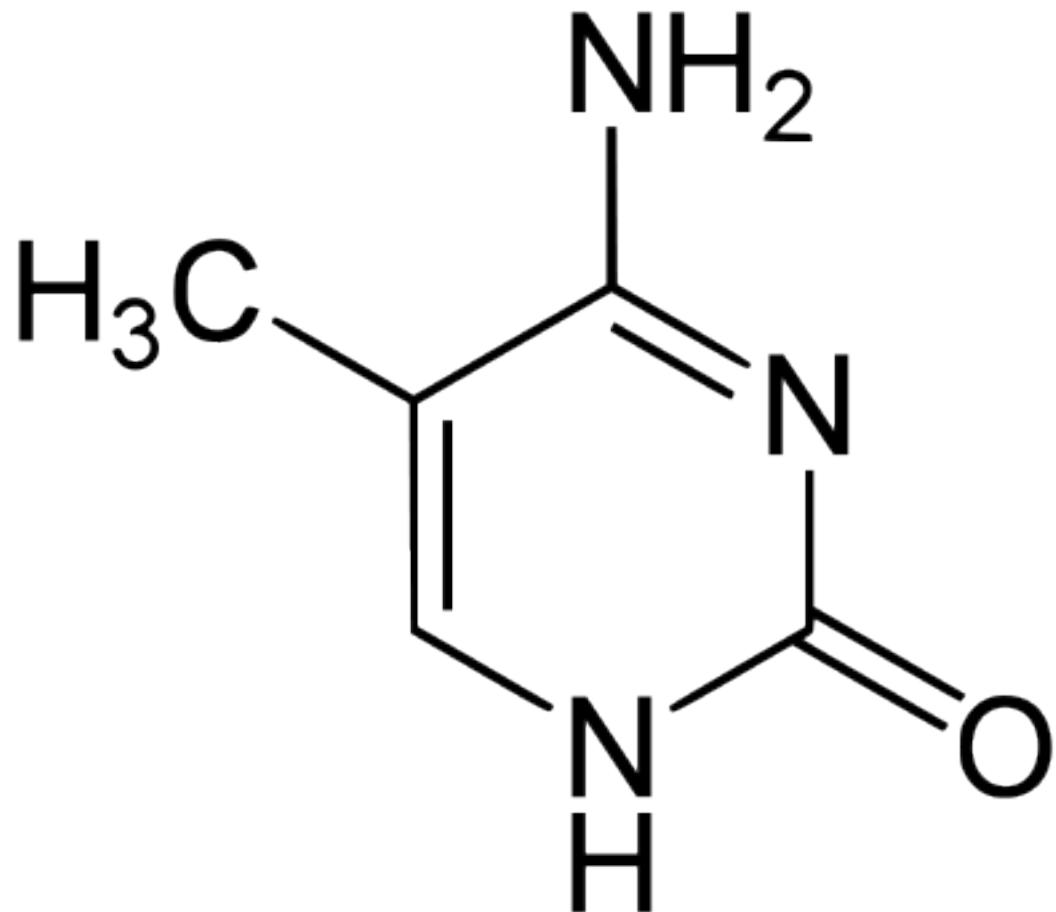
Challenge 3: Transcript abundances are stochastic

Solution: Replicates, replicates, and more replicates

RNA-seq differential expression studies: more sequence or more replication?

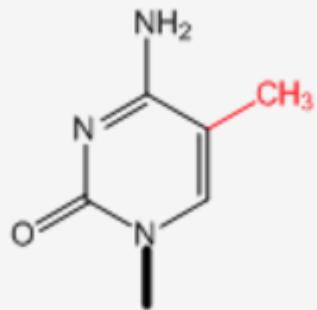
Liu et al (2013) *Bioinformatics*. doi:10.1093/bioinformatics/btt688

Methyl-seq

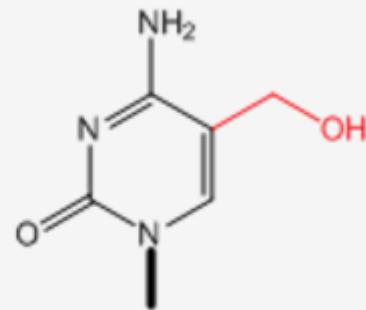


Finding the fifth base: Genome-wide sequencing of cytosine methylation
Lister and Ecker (2009) *Genome Research.* 19: 959-966

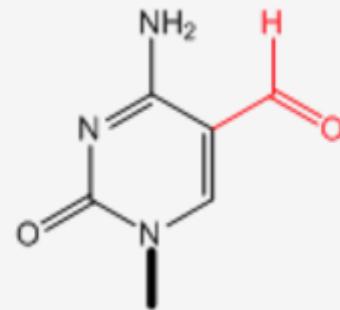
Epigenetic Modifications to DNA



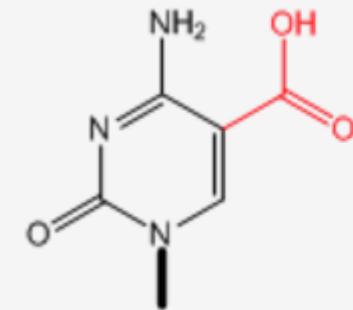
5-mC



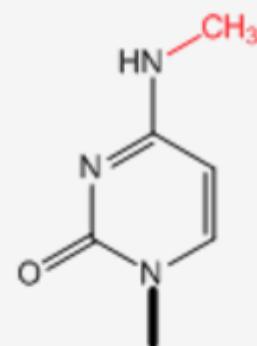
5-hmC



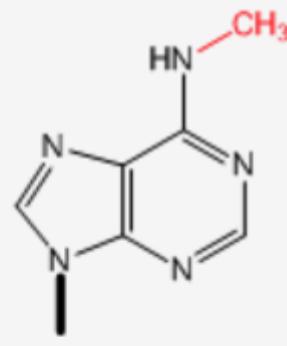
5-fC



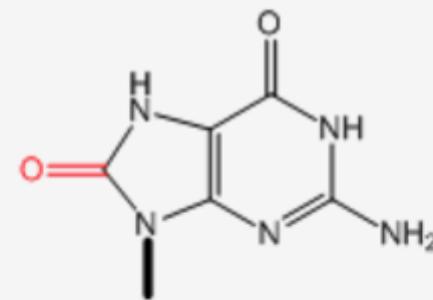
5-caC



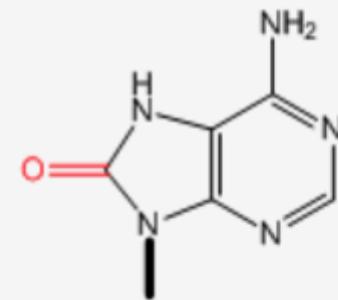
4-mC



6-mA



8-oxoG



8-oxoA

The Honey Bee Epigenomes: Differential Methylation of Brain DNA in Queens and Workers

Frank Lyko¹*, Sylvain Foret²*, Robert Kucharski³, Stephan Wolf⁴, Cassandra Falckenhayn¹, Ryszard Maleszka^{3*}

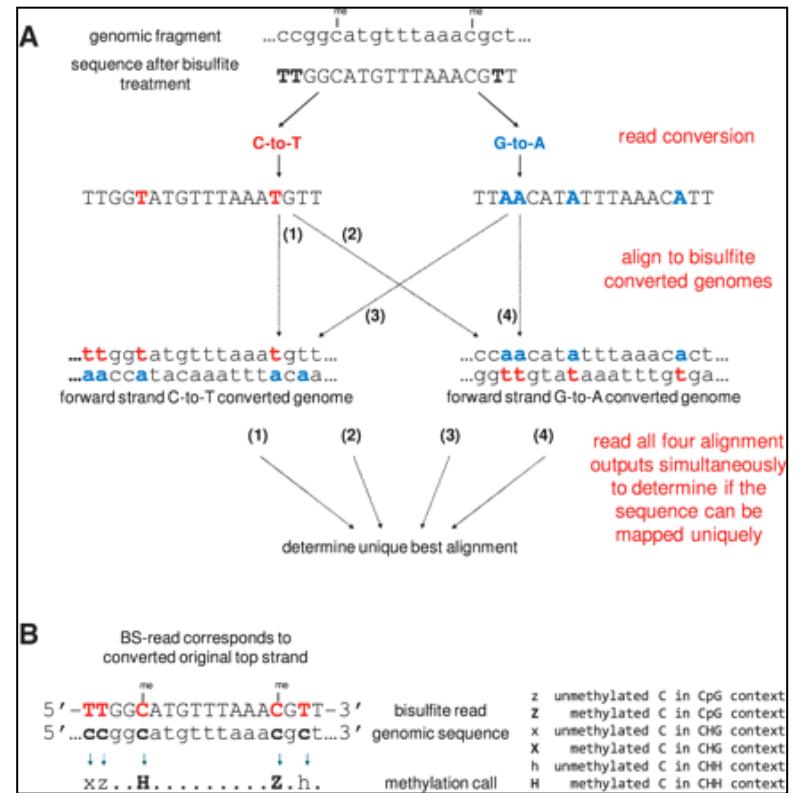
1 Division of Epigenetics, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany, **2** ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Australia, **3** Research School of Biology, the Australian National University, Canberra, Australia, **4** Genomics and Proteomics Core Facility, German Cancer Research Center, Heidelberg, Germany



Bisulfite Conversion

Treating DNA with sodium bisulfite will convert unmethylated C to T

- 5-MethylC will be protected and not change, so can look for differences when mapping
- Requires great care when analyzing reads, since the complementary strand will also be converted (G to A)
- Typically analyzed by mapping to a “reduced alphabet” where we assume all Cs are converted to Ts once on the forward strand and once on the reverse



Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications
Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.

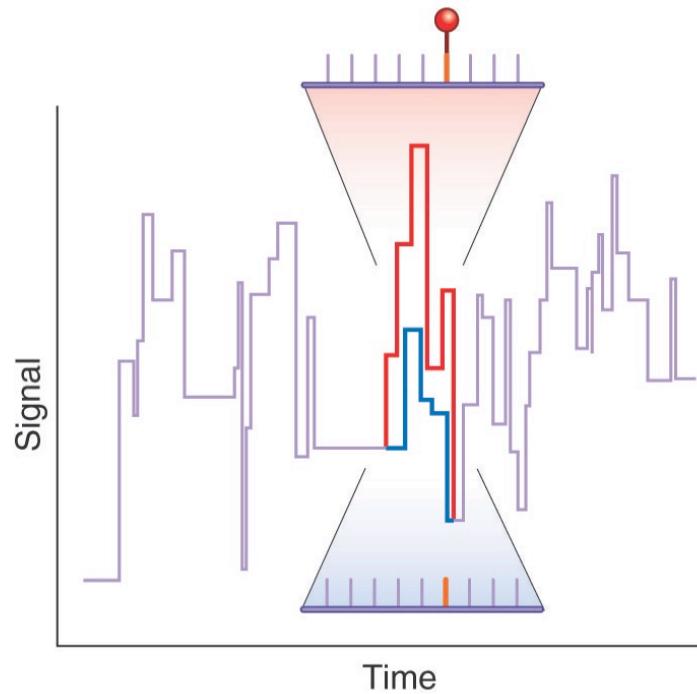
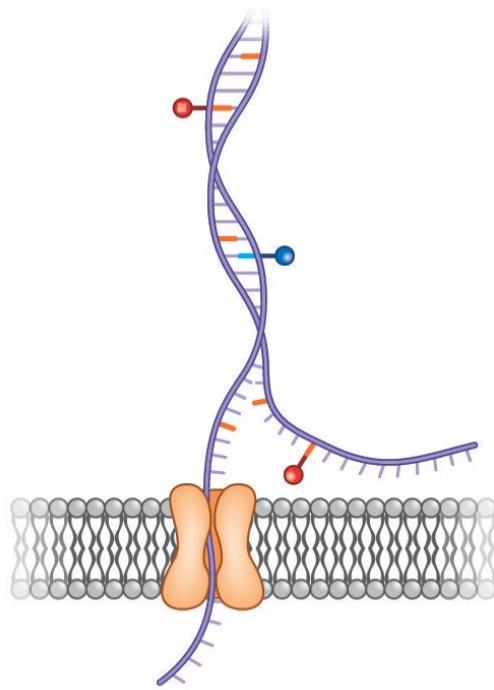
Bisulfite Conversion

T
W
•
•
•
•
•



Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications
Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.

Methylation Detection using Oxford Nanopore Sequencing

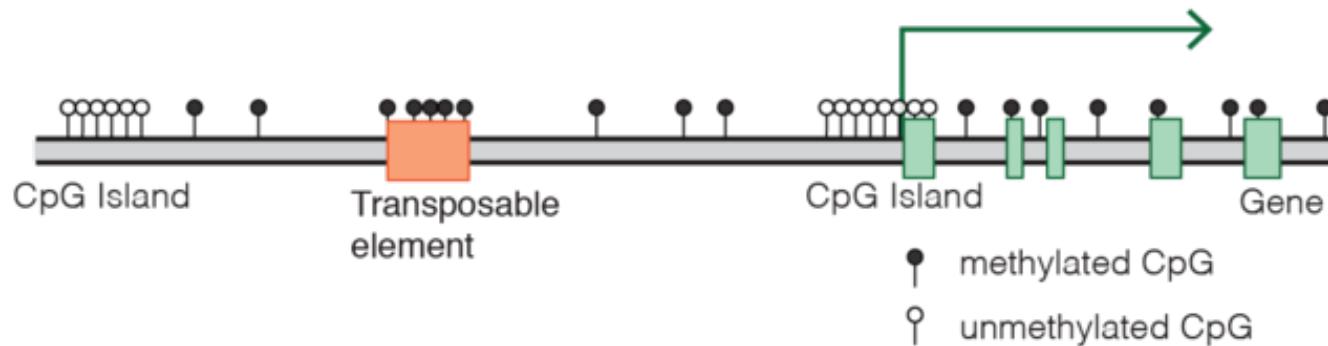


Detecting DNA cytosine methylation using nanopore sequencing
Simpson et al (2017) Nature Methods. doi:10.1038/nmeth.4184

Mapping DNA methylation with high-throughput nanopore sequencing
Rand et al (2017) Nature Methods. doi:10.1038/nmeth.4189

Methylation of CpG Islands

Typical mammalian DNA methylation landscape



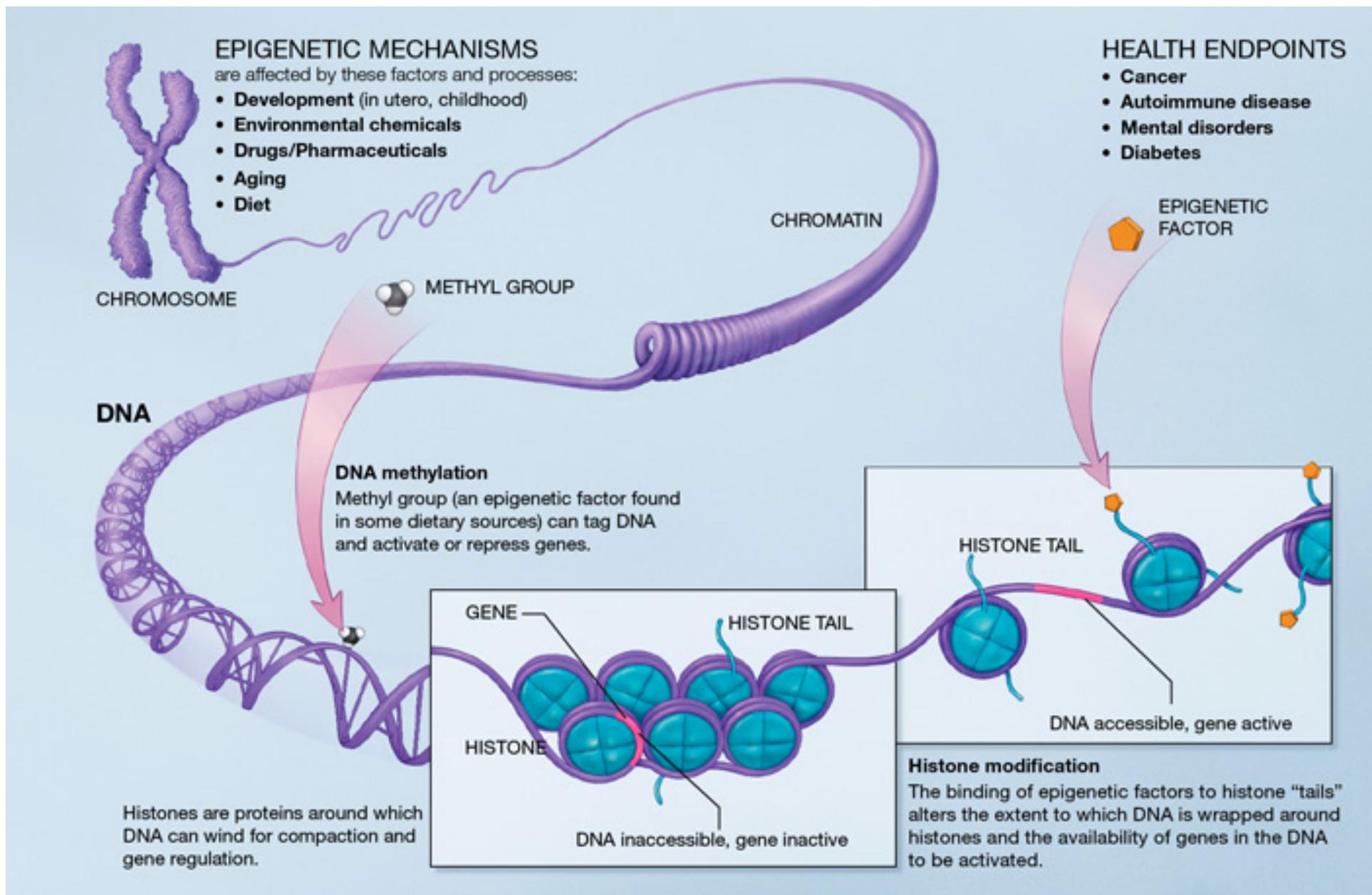
CpG islands are (usually) defined as regions with

- 1) a length greater than 200bp,
- 2) a G+C content greater than 50%,
- 3) a ratio of observed to expected CpG greater than 0.6

Methylation in promoter regions correlates negatively with gene expression.

- CpG-dense promoters of actively transcribed genes are never methylated
- In mouse and human, around 60-70% of genes have a CpG island in their promoter region and most of these CpG islands remain unmethylated independently of the transcriptional activity of the gene
- Methylation of DNA itself may physically impede the binding of transcriptional proteins to the gene
- Methylated DNA may be bound by proteins known as methyl-CpG-binding domain proteins (MBDs) that can modify histones, thereby forming compact, inactive chromatin, termed heterochromatin.

Methylation & Epigenetics



Methylation in Cancer

Michael

www.hopkinsmedicine.org/scical/ Other Bookmarks

Johns Hopkins Science Calendar

A listing of scientific events for the Johns Hopkins community

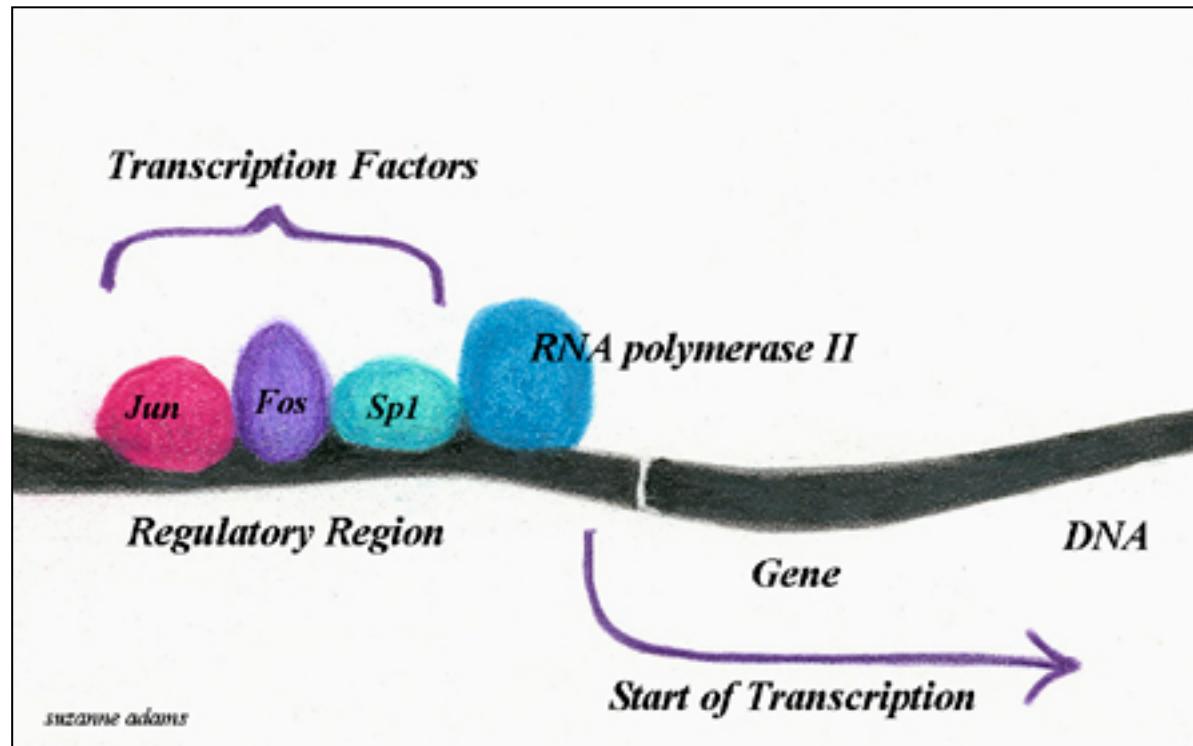
Welcome to Science Calendar Online, a publication of Johns Hopkins Medicine. This calendar lists events sponsored by JHM and other Hopkins-affiliated institutions and is open to anyone in the Hopkins family.

You can [submit an event entry online](#) or for more information, or to report an issue please submit a [web service request](#).

Event Date	Detail	Contact Information
Thursday Mar 09, 2017 12:00 PM	Establishing the C. elegans primordial germ cell niche: Tales of cell attraction, interaction, and cannibalism Jeremy Nance, Ph.D. Associate Professor, NYU School of Medicine 1830 Building, Suite 2-200 Sponsored by: Department of Cell Biology	410-502-7827
Thursday Mar 09, 2017 4:30 PM	The Sidney Kimmel Comprehensive Cancer Center Presents the Director's Visiting Professor Lecture Series featuring Peter A. Jones, B.Sc., Ph.D. "DNA Methylation as a Sculptor of the Genome and an Organizer of the Epigenome" Peter A. Jones, B.Sc., Ph.D. Chief Scientific Officer, Van Andel Research Institute, Grand Rapids, MI Albert H. Owens Jr. Auditorium Sponsored by: Sidney Kimmel Comprehensive Cancer Center	410 955 9702
Friday Mar 10, 2017 2:00 PM	Detection of ESR1 mutations and modeling hormone therapy resistance David Chu PhD Candidate CRB I Room 3M42 Sponsored by: Cellular and Molecular Medicine	410-614-3640
Monday Mar 13, 2017 8:30 AM	Pathology Grand Rounds: Roman Vishniac: The Curious Microscopist Norman Barker, MA, MS, RBP	410-955-9790

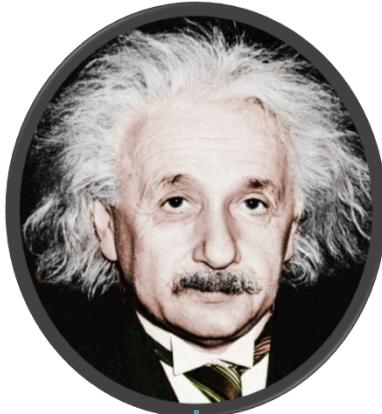
tenuredfacultymeeting....zip Show All

ChIP-seq



Genome-wide mapping of *in vivo* protein-DNA interactions.
Johnson et al (2007) Science. 316(5830):1497-502

Human Evolution

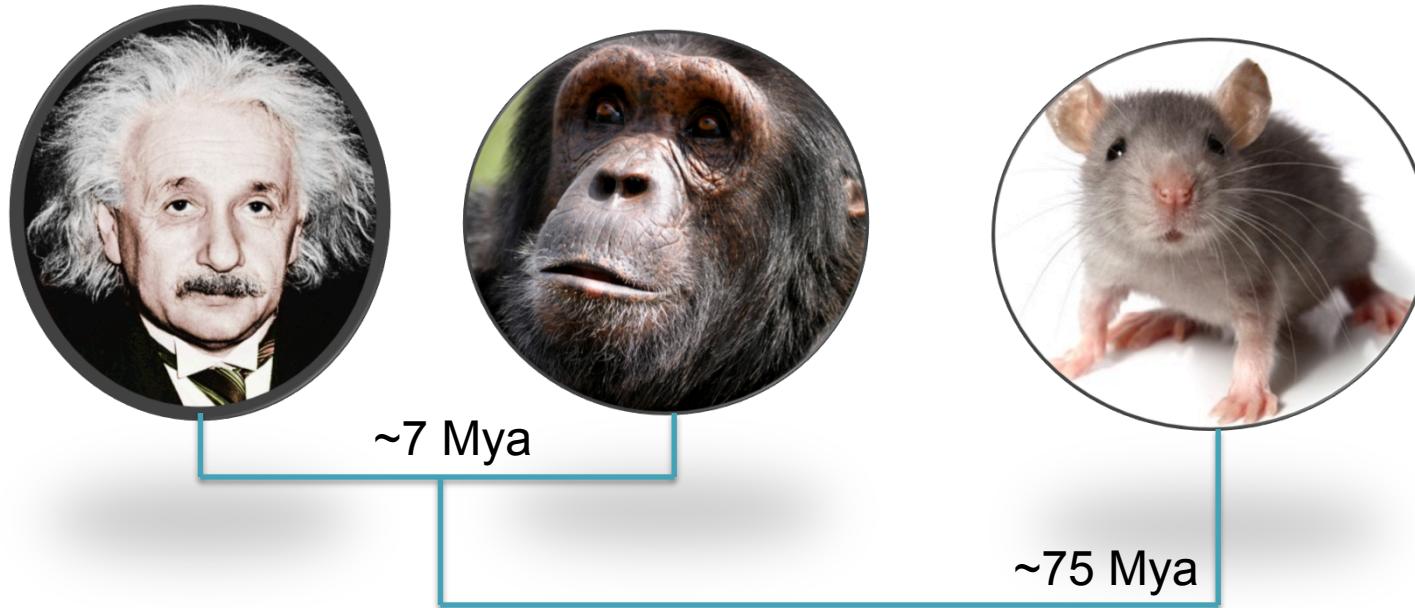


~5 Mya

- Humans and chimpanzees shared a common ancestor ~5-7 million years ago (Mya)
- Single-nucleotide substitutions occur at a mean rate of 1.23% but ~4% overall rate of mutation: comprising ~35 million single nucleotide differences and ~90 Mb of insertions and deletions
- Orthologous proteins in human and chimpanzee are extremely similar, with ~29% being identical and the typical orthologue differing by only two amino acids, one per lineage

Initial sequence of the chimpanzee genome and comparison with the human genome
(2005) *Nature* 437, 69-87 doi:10.1038/nature04072

Human Evolution

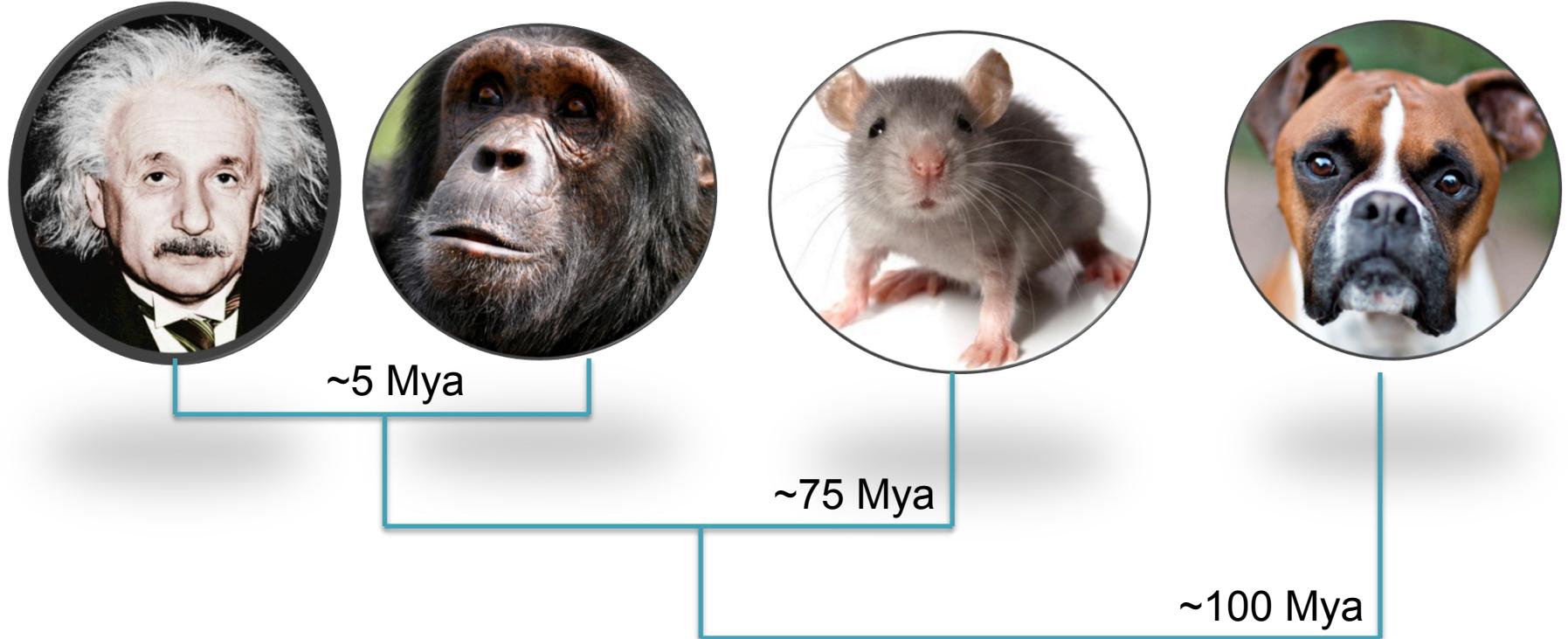


"In the roughly 75 million years since the divergence of the human and mouse lineages, the process of evolution has altered their genome sequences and caused them to diverge by ***nearly one substitution for every two nucleotides***"

The mouse and human genomes each seem to contain about 30,000 protein-coding genes. These refined estimates have been derived from both new evidence-based analyses that produce larger and more complete sets of gene predictions, and new de novo gene predictions that do not rely on previous evidence of transcription or homology. The proportion of mouse genes with a single identifiable orthologue in the human genome seems to be approximately 80%. ***The proportion of mouse genes without any homologue currently detectable in the human genome (and vice versa) seems to be less than 1%.***"

Initial sequencing and comparative analysis of the mouse genome
Chinwalla et al (2002) *Nature*. 420, 520-562 doi:10.1038/nature01262

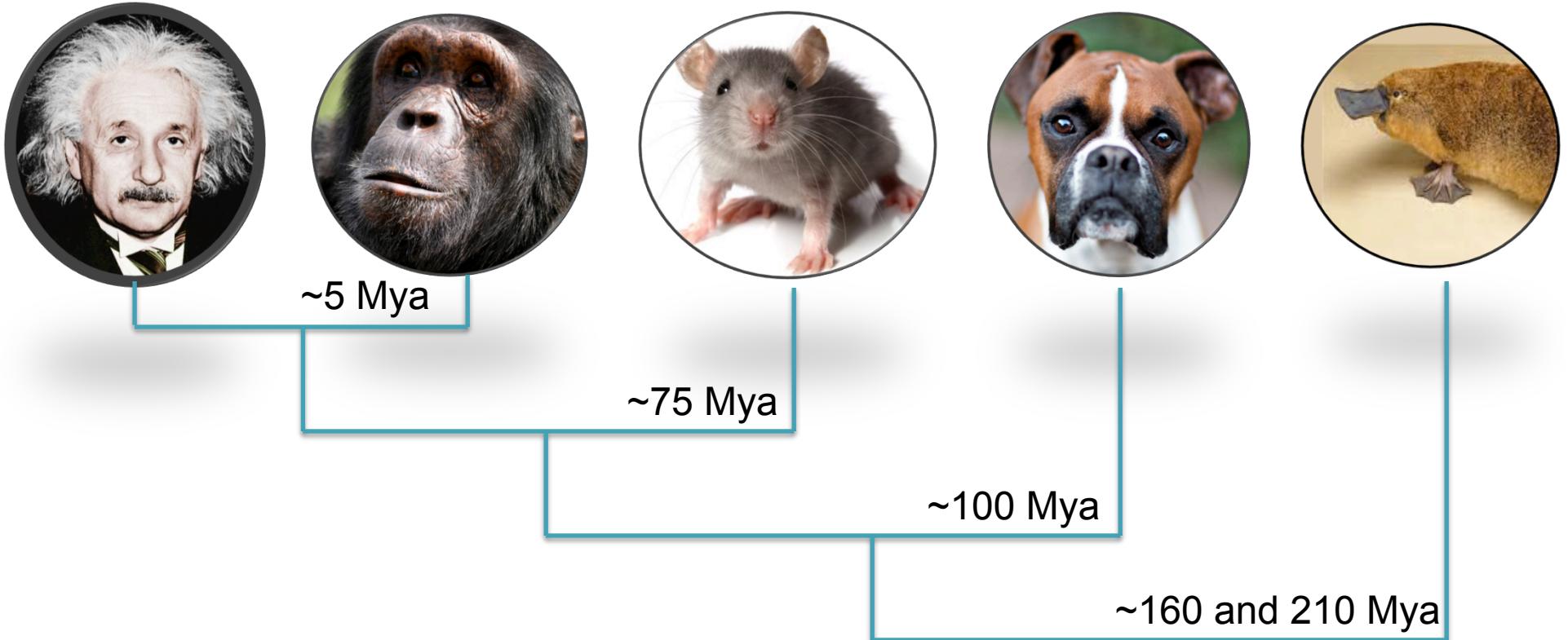
Human Evolution



"We generated gene predictions for the dog genome using an evidence-based method (see Supplementary Information). The resulting collection contains **19,300 dog gene predictions, with nearly all being clear homologues of known human genes**. The dog gene count is substantially lower than the ~22,000-gene models in the current human gene catalogue (Ensembl build 26). For many predicted human genes, we find no convincing evidence of a corresponding dog gene. Much of the excess in the human gene count is attributable to **spurious gene predictions in the human genome**"

Genome sequence, comparative analysis and haplotype structure of the domestic dog
Lindblad-Toh et al (2005) *Nature*. 438, 803-819 doi:10.1038/nature04338

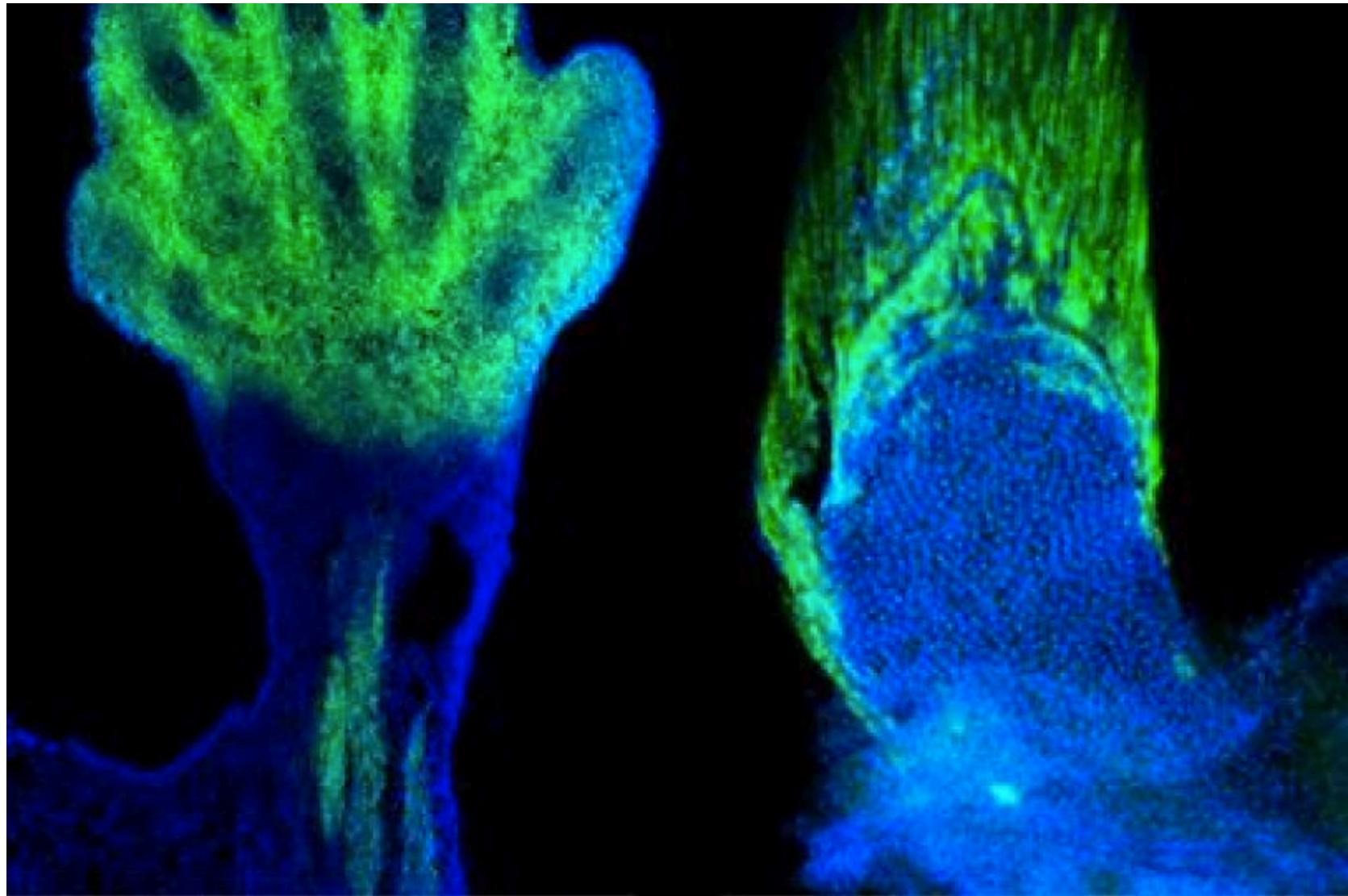
Human Evolution



As expected, the majority of platypus genes (82%; 15,312 out of 18,596) have orthologues in these five other amniotes (Supplementary Table 5). The remaining 'orphan' genes are expected to primarily reflect rapidly evolving genes, for which no other homologues are discernible, erroneous predictions, and true lineage-specific genes that have been lost in each of the other five species under consideration.

Genome analysis of the platypus reveals unique signatures of evolution
(2008) *Nature*. 453, 175-183 doi:10.1038/nature06936

Human Evolution



Digits and fin rays share common developmental histories

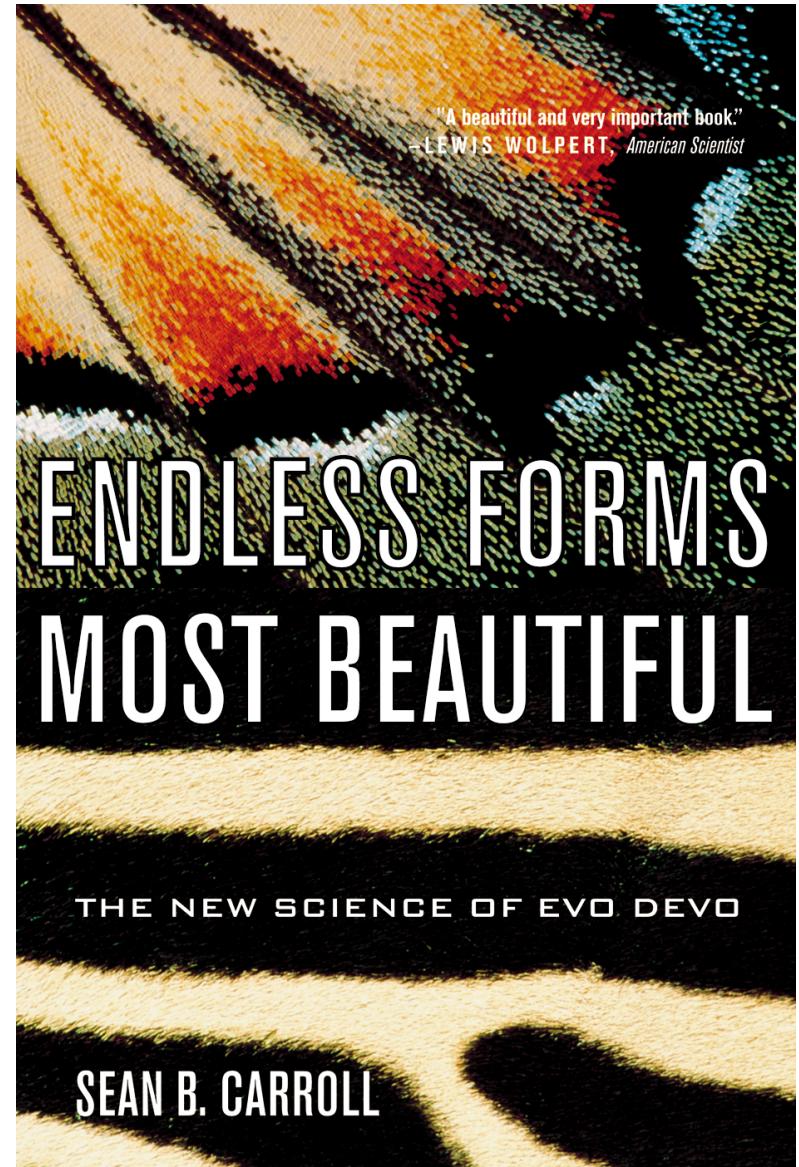
Nakamura et al (2016) *Nature*. 537, 225–228. doi:10.1038/nature19322

More Information

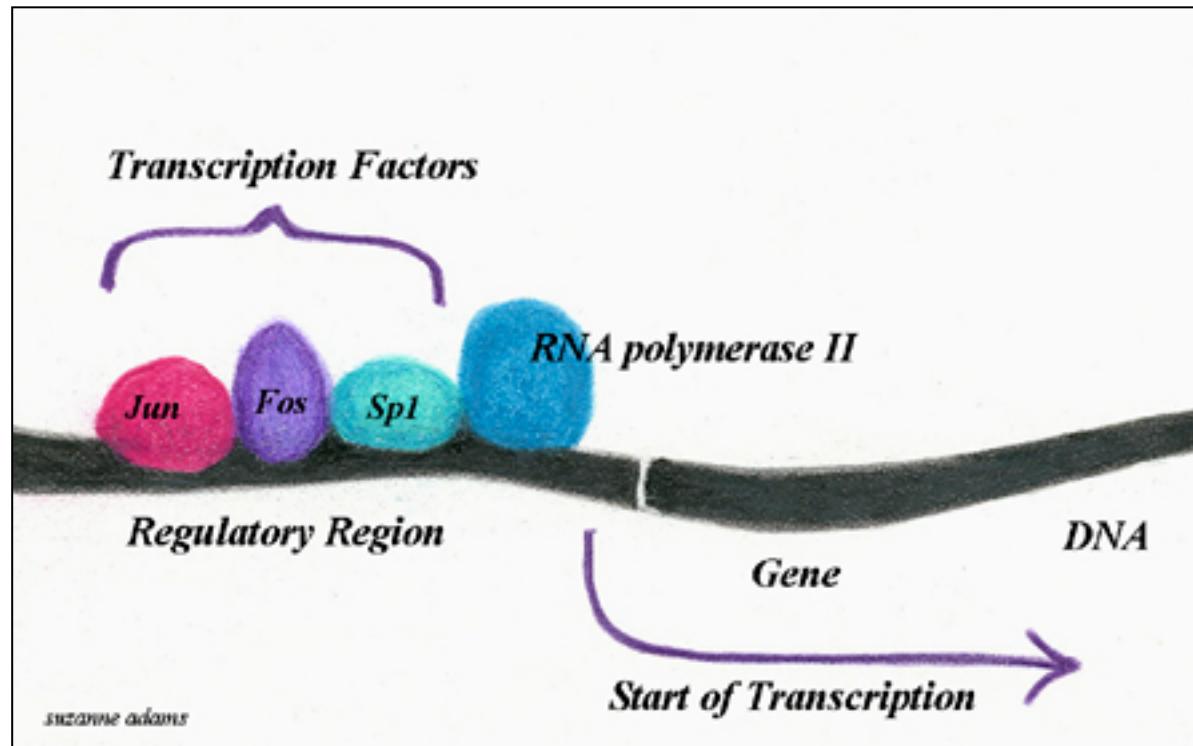


*“Anything found to be true of
E. coli must also be true of
elephants”*

-Jacques Monod

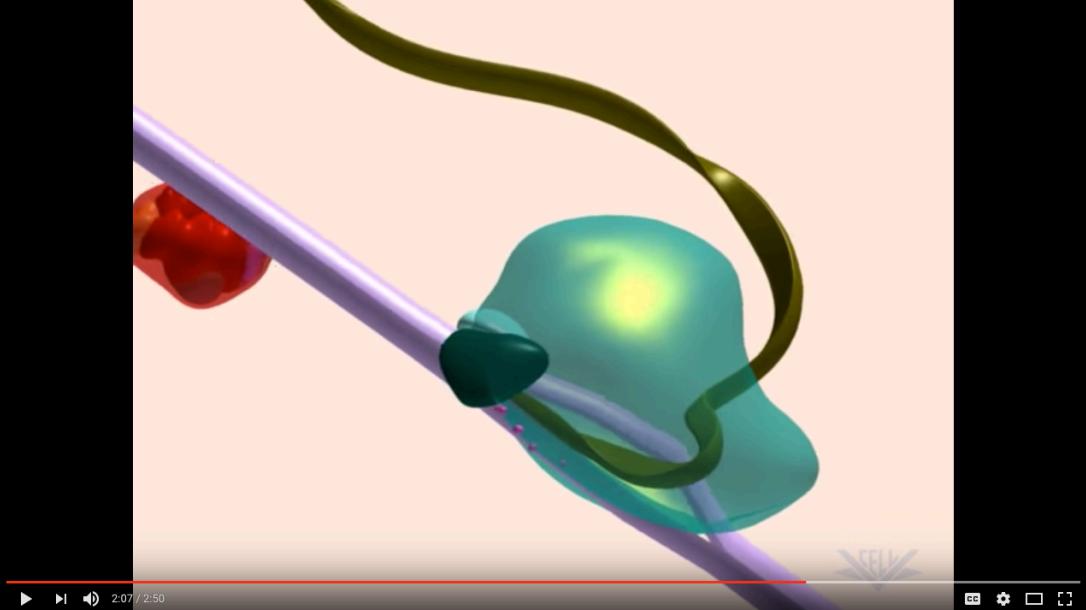


ChIP-seq



Genome-wide mapping of *in vivo* protein-DNA interactions.
Johnson et al (2007) Science. 316(5830):1497-502

Transcription



The image shows a YouTube video player window. The main video frame displays a 3D rendering of a DNA double helix with a green RNA strand being synthesized. A red ribosome-like complex is visible on the RNA. The video is titled "Transcription" and has 2,018,430 views. It was uploaded by "ndsvirtualcell" on Jan 30, 2008. Below the video, there is a URL: <https://www.youtube.com/watch?v=bKipDtJdK8Q>. To the right of the video, there is a sidebar titled "Up next" showing several other video thumbnails related to transcription and translation.

Transcription

2,018,430 views

ndsvirtualcell Uploaded on Jan 30, 2008

NDsu Virtual Cell Animations Project animation 'Transcription'. For more information please see <http://vcell.ndsu.edu/animations>

[SUBSCRIBE 45K](#)

Up next

AUTOPLAY

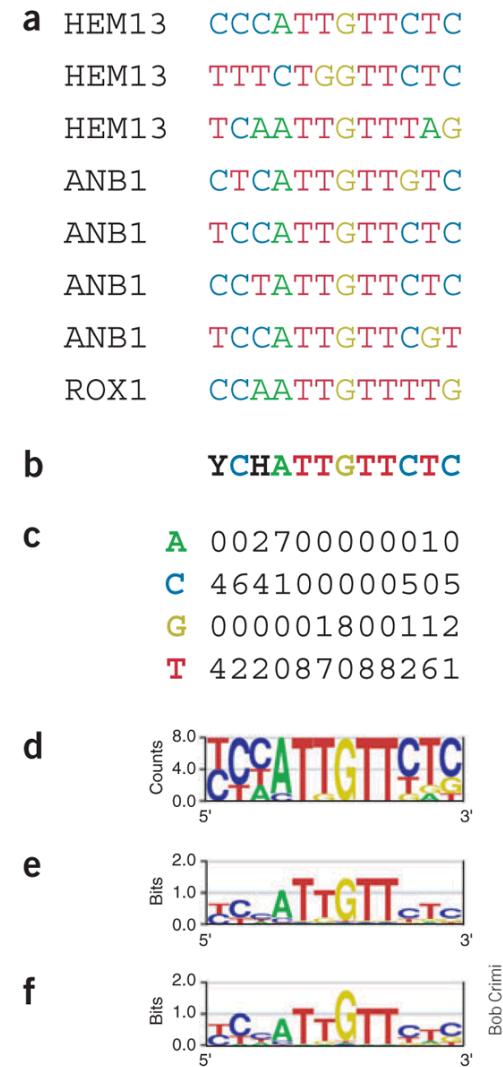
- Transcription and Translation: From DNA to Protein Professor Dave Explains 151K views RNA polymerase reads 3 6:27
- DNA - transcription and translation Wisam Kabaha 40K views 7:18
- Transcription and mRNA processing | Biomolecules | Khan Academy 106K views 10:25
- DNA transcription and translation Animation Haider abd 45K views 7:18
- Translation ndsvirtualcell 2.1M views 3:33
- Transcription and Translation Overview Armando Hasudungan 611K views 13:18
- DNA, Hot Pockets, & The Longest Word Ever: Crash CrashCourse 2.2M views 14:08
- TRANSCRIPTION 1 khanacademymedicine 263K views 12:06
- TRANSCRIPTION congthanhang 795K views 1:28
- Moana - Best Scenes (FHD)

<https://www.youtube.com/watch?v=WsofH466lqk>

Transcription Factors

A transcription factor (or sequence-specific DNA-binding factor) is a protein that controls the rate of transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence.

- Transcription factors work alone or with other proteins in a complex, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase to specific genes.
- A defining feature of transcription factors is that they contain at least one DNA-binding domain (DBD)
- Figure (a) Eight known genomic binding sites in three *S. cerevisiae* genes. (b) Degenerate consensus sequence. (c,d) Frequencies of nucleotides at each position. (e) Sequence logo (f) Energy normalized logo using relative entropy to adjust for low GC content in *S. cerevisiae*.



What are DNA sequence motifs?

D'haeseleer (2006) Nature Biotechnology 24, 423 – 425 doi:10.1038/nbt0406-423

Transcription Factors Database

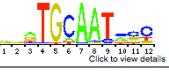
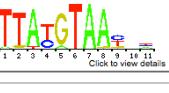
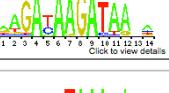
The JASPAR database

jaspar.genereg.net/cgi-bin/jaspar_db.pl?rm=browse&db=core&tax_group=vertebrates

Michael

SEARCH Name AND Species AND Class SEARCH ?

JASPAR matrix models:

TOGGLE	ID	name	species	class	family	Sequence logo
<input type="checkbox"/>	MA0004.1	Arnt	<i>Mus musculus</i>	Basic helix-loop-helix factors (bHLH)	PAS domain factors	
<input type="checkbox"/>	MA0006.1	Ahr::Arnt	<i>Mus musculus</i>	Basic helix-loop-helix factors (bHLH); Basic helix-loop-helix factors (bHLH)	PAS domain factors::PAS domain factors	
<input type="checkbox"/>	MA0019.1	Ddit3::Cebpa	<i>Rattus norvegicus</i>	Basic leucine zipper factors (bZIP); Basic leucine zipper factors (bZIP)	C/EBP-related	
<input type="checkbox"/>	MA0025.1	NFIL3	<i>Homo sapiens</i>	Basic leucine zipper factors (bZIP)	C/EBP-related	
<input type="checkbox"/>	MA0029.1	Mecom	<i>Mus musculus</i>	C2H2 zinc finger factors	Factors with multiple dispersed zinc fingers	
<input type="checkbox"/>	MA0030.1	FOXF2	<i>Homo sapiens</i>	Fork head / winged helix factors	Forkhead box (FOX) factors	
<input type="checkbox"/>	MA0031.1	FOXD1	<i>Homo sapiens</i>	Fork head / winged helix factors	Forkhead box (FOX) factors	
<input type="checkbox"/>	MA0038.1	Gfi1	<i>Rattus norvegicus</i>	C2H2 zinc finger factors	More than 3 adjacent zinc finger factors	
<input type="checkbox"/>	MA0040.1	Foxq1	<i>Rattus norvegicus</i>	Fork head / winged helix factors	Forkhead box (FOX) factors	

ANALYZE selected matrix models:

CLUSTER ? selected models using STAMP

Create RANDOM matrix models based on selected models

Number of matrices: 200 Format: Raw RANDOMIZE ?

Create models with PERMUTED columns from selected:

Type: Within each matrix Format: Raw PERMUTE ?

SCAN this (fasta-formatted) sequence with selected matrix models

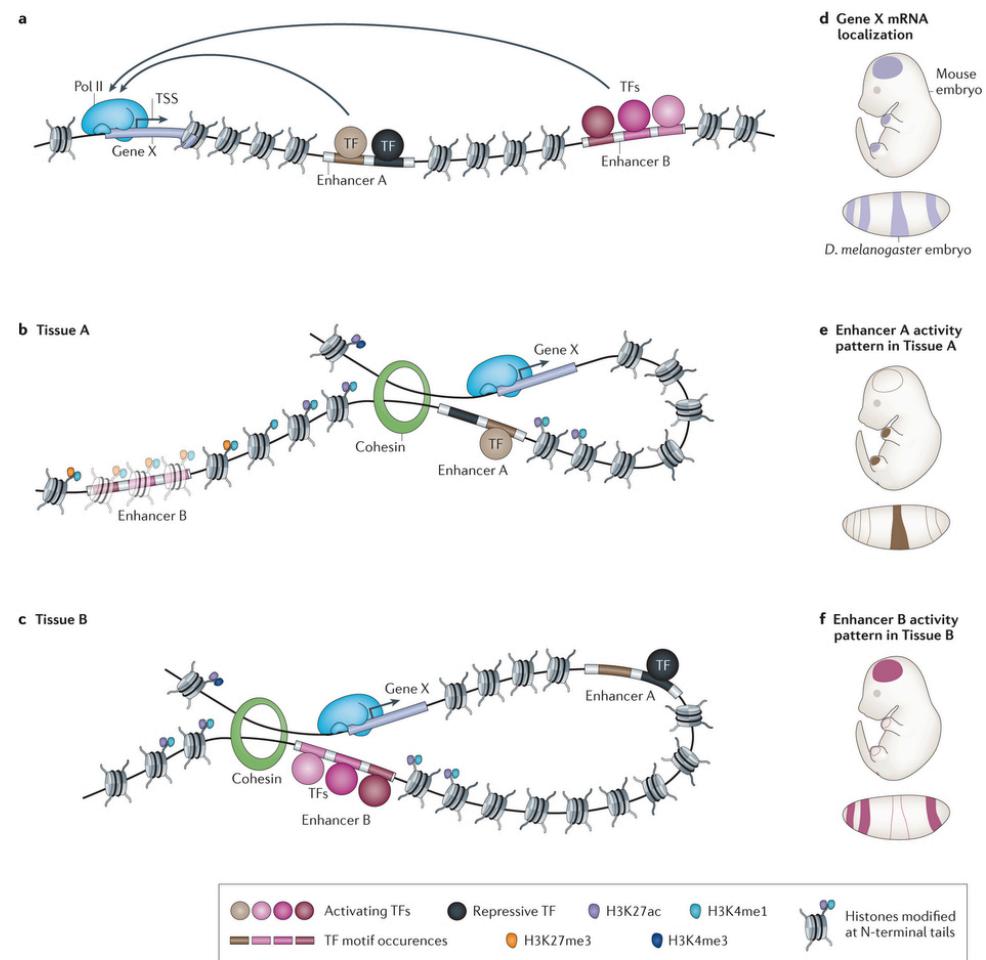
Relative profile score threshold 80 % SCAN ?

JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles
Anthony Mathelier (2014) Nucleic Acids Res. 42 (D1): D142-D147. DOI: <https://doi.org/10.1093/nar/gkt997>

Enhancers

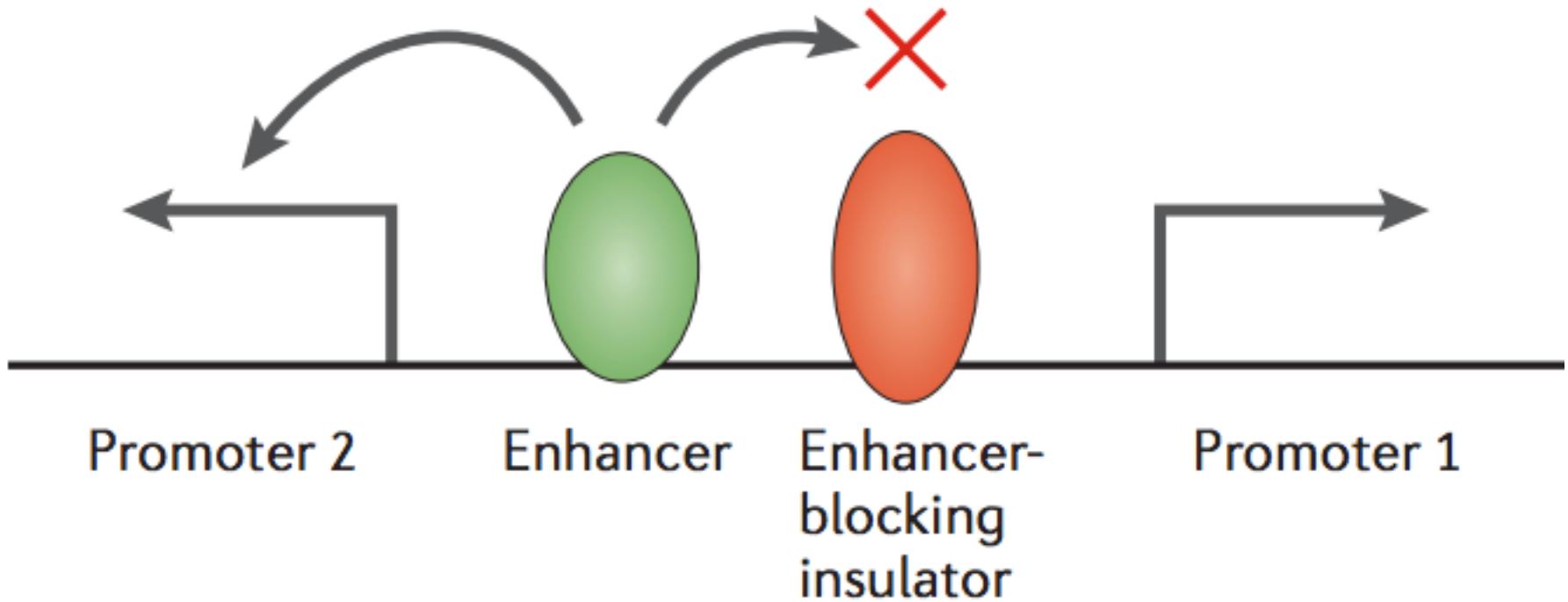
Enhancers are genomic regions that contain binding sites for transcription factors (TFs) and that can upregulate (enhance) the transcription of a target gene.

- Enhancers can be located at any distance from their target genes (up to ~1Mbp)
- In a given tissue, active enhancers (Enhancer A in part b or Enhancer B in part c) are bound by activating TFs and are brought into proximity of their respective target promoters by looping
- Active and inactive gene regulatory elements are marked by various biochemical features
- Complex patterns of gene expression result from the additive action of different enhancers with cell-type- or tissue-specific activities



Transcriptional enhancers: from properties to genome-wide predictions
Shlyueva et al (2014) *Nature Reviews Genetics* 15, 272–286

Insulators



Insulators are DNA sequence elements that prevent “inappropriate interactions” between adjacent chromatin domains.

- One type of insulator establishes domains that separate enhancers and promoters to block their interaction,
- Second type creates a barrier against the spread of heterochromatin.

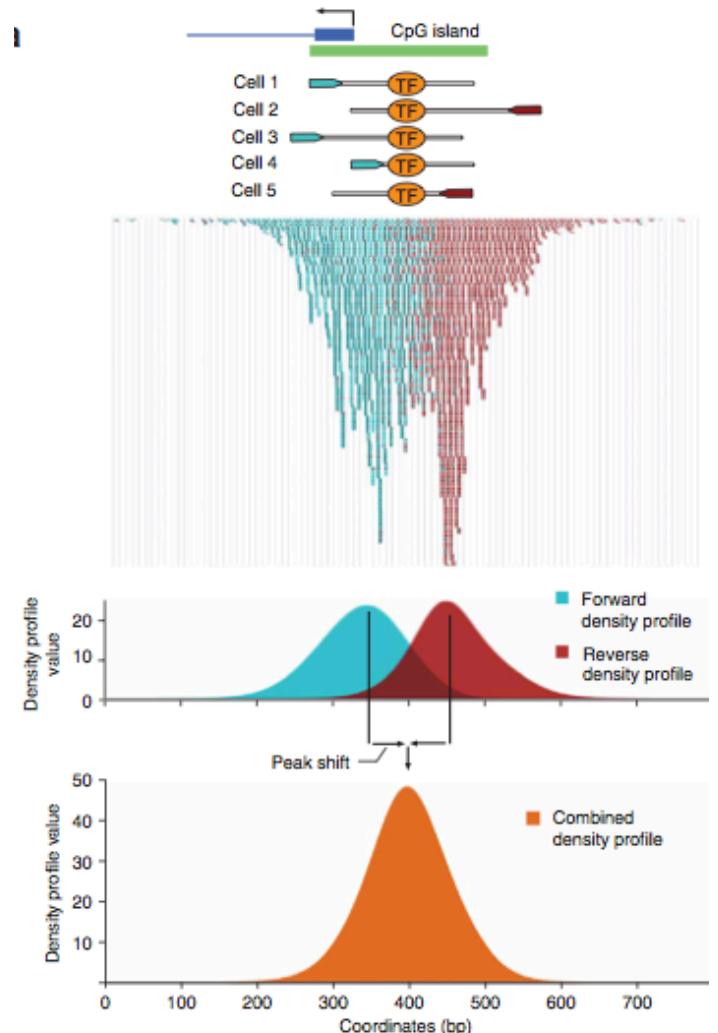
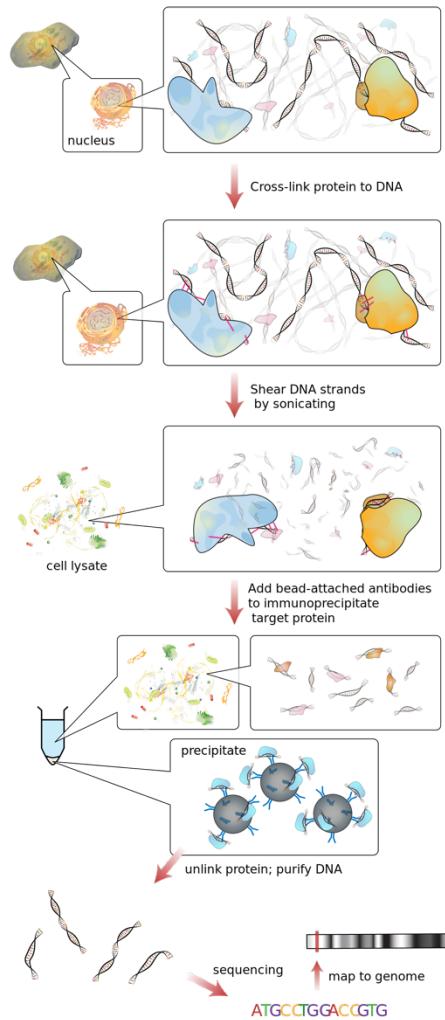
Insulators: exploiting transcriptional and epigenetic mechanisms

Gaszner & Felsenfeld (2006) *Nature Reviews Genetics* 7, 703-713. doi:10.1038/nrg1925

ChIP-seq:TF Binding

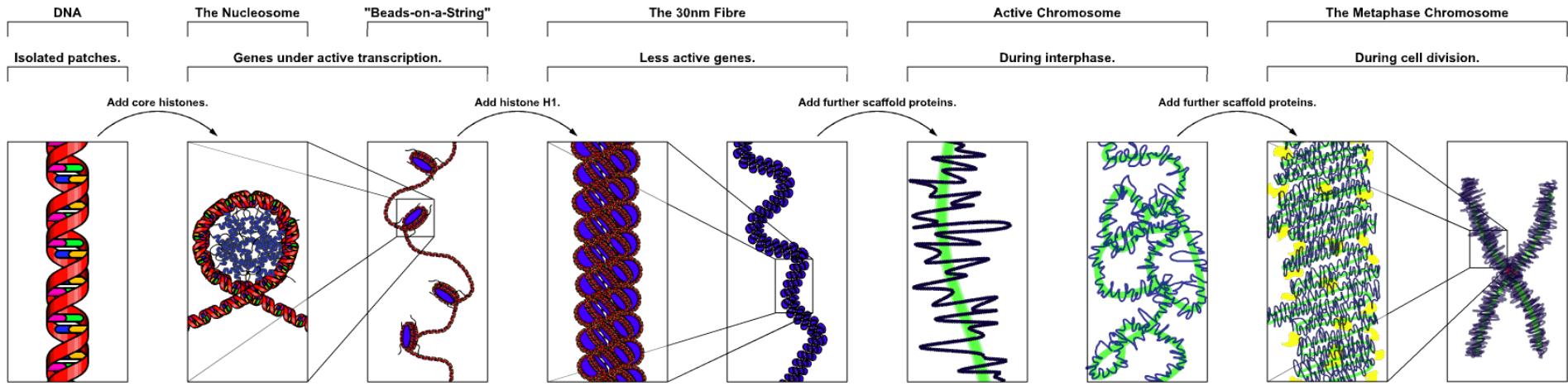
Goals:

- Where are transcription factors and other proteins binding to the DNA?
- How strongly are they binding?
- Do the protein binding patterns change over developmental stages or when the cells are stressed?



Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data
Valouev et al (2008) *Nature Methods*. 5, 829 - 834

Chromatin compaction model



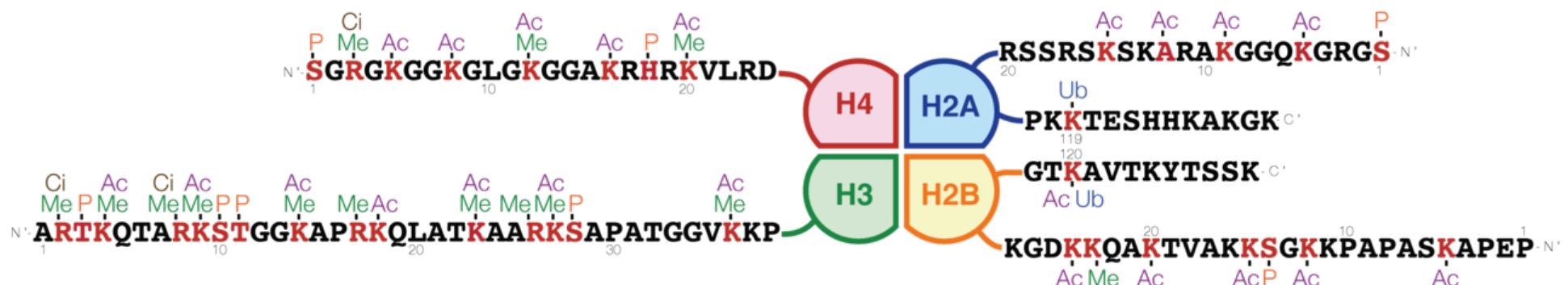
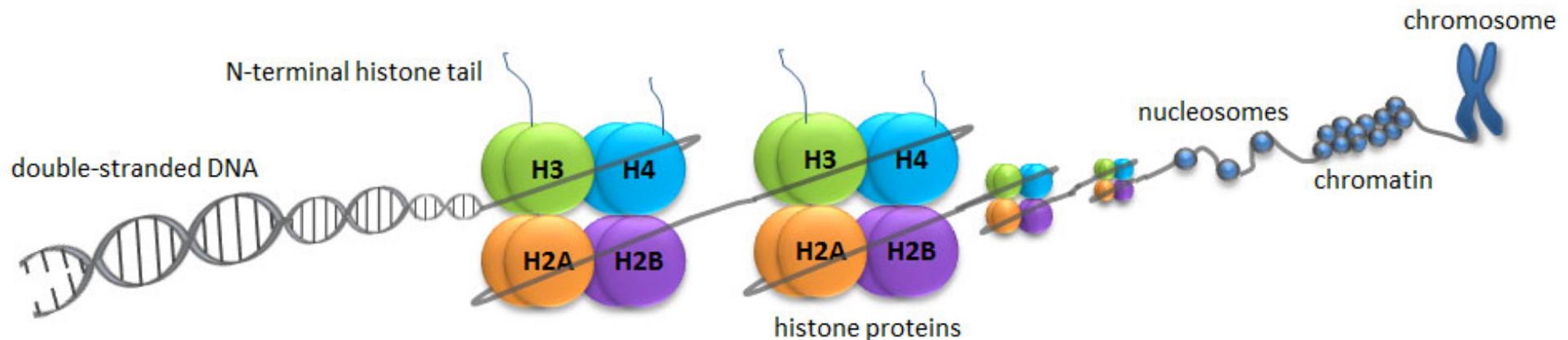
Nucleosome is a basic unit of DNA packaging in eukaryotes

- Consists of a segment of 146bp DNA wound in sequence around eight histone protein cores (thread wrapped around a spool) followed by a ~38bp linker
- Under active transcription, nucleosomes appear as “beads-on-a-string”, but are more densely packed for less active genes

Nucleosomes form the fundamental repeating units of eukaryotic chromatin

- Used to pack the large eukaryotic genomes into the nucleus while still ensuring appropriate access to it (in mammalian cells approximately 2 m of linear DNA have to be packed into a nucleus of roughly 10 μm diameter).

ChIP-seq: Histone Modifications



Me Methylation

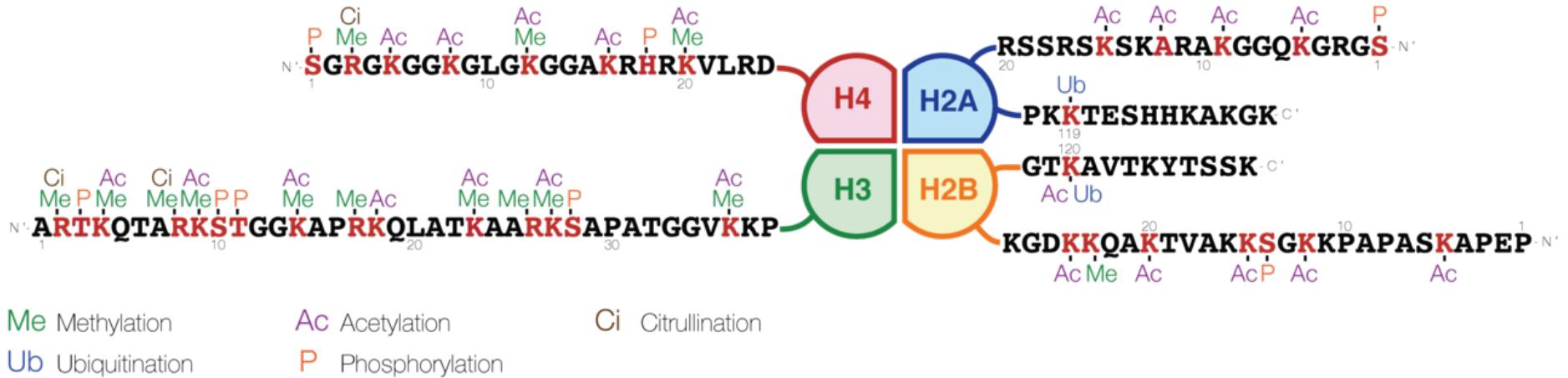
Ub Ubiquitination

Ac Acetylation

P Phosphorylation

Ci Citrullination

ChIP-seq: Histone Modifications

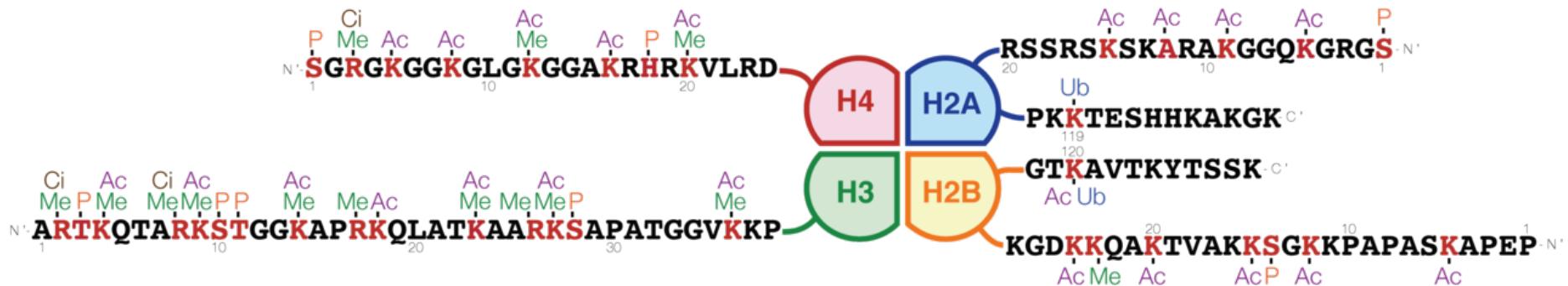


The common nomenclature of histone modifications is:

- The name of the histone (e.g., H3)
- The single-letter amino acid abbreviation (e.g., K for Lysine) and the amino acid position in the protein
- The type of modification (Me: methyl, P: phosphate, Ac: acetyl, Ub: ubiquitin)
- The number of modifications (only Me is known to occur in more than one copy per residue. 1, 2 or 3 is mono-, di- or tri-methylation)

So H3K4me1 denotes the monomethylation of the 4th residue (a lysine) from the start (i.e., the N-terminal) of the H3 protein.

ChIP-seq: Histone Modifications

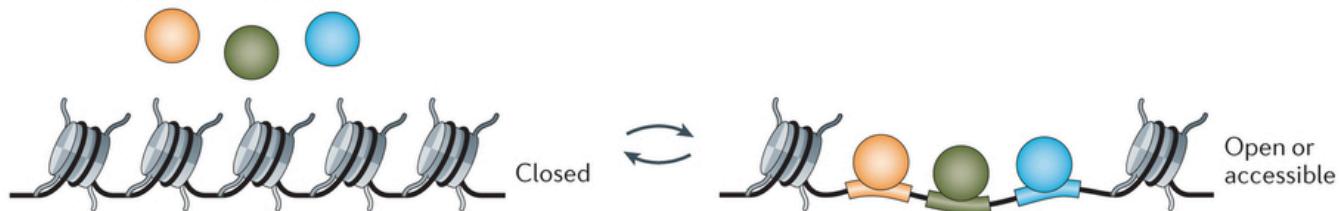


Type of modification	Histone							
	H3K4	H3K9	H3K14	H3K27	H3K79	H3K122	H4K20	H2BK5
mono-methylation	activation ^[6]	activation ^[7]		activation ^[7]	activation ^{[7][8]}		activation ^[7]	activation ^[7]
di-methylation	activation	repression ^[3]		repression ^[3]	activation ^[8]			
tri-methylation	activation ^[9]	repression ^[7]		repression ^[7]	activation, ^[8] repression ^[7]			repression ^[3]
acetylation		activation ^[9]	activation ^[9]	activation ^[10]		activation ^[11]		

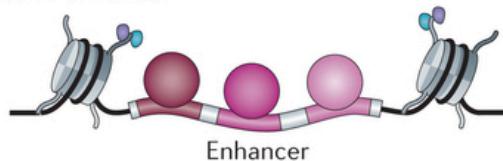
- H3K4me3 is enriched in transcriptionally active promoters.^[12]
- H3K9me3 is found in constitutively repressed genes.
- H3K27me is found in facultatively repressed genes.^[7]
- H3K36me3 is found in actively transcribed gene bodies.
- H3K9ac is found in actively transcribed promoters.
- H3K14ac is found in actively transcribed promoters.
- H3K27ac distinguishes active enhancers from poised enhancers.
- H3K122ac is enriched in poised promoters and also found in a different type of putative enhancer that lacks H3K27ac.

Enhancer States

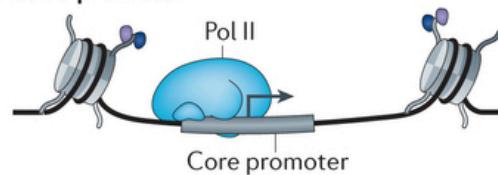
a Chromatin as accessibility barrier



b Active enhancer



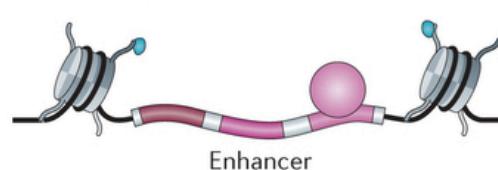
c Active promoter



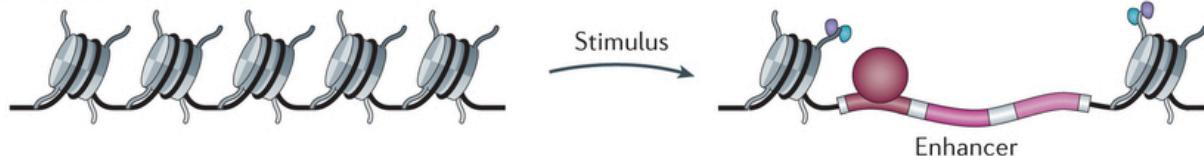
d Closed or poised enhancer



e Primed enhancer



f Latent enhancer

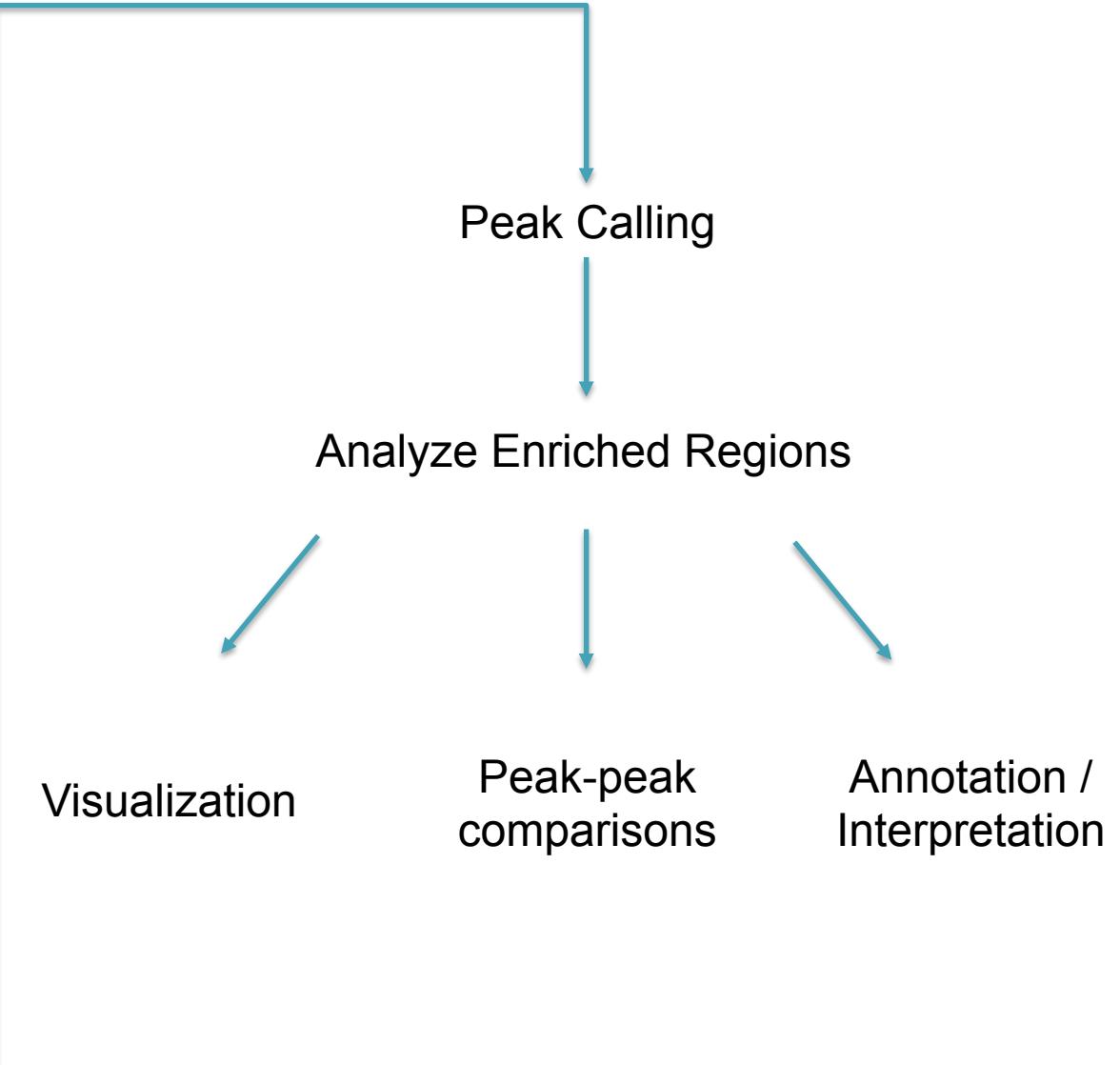
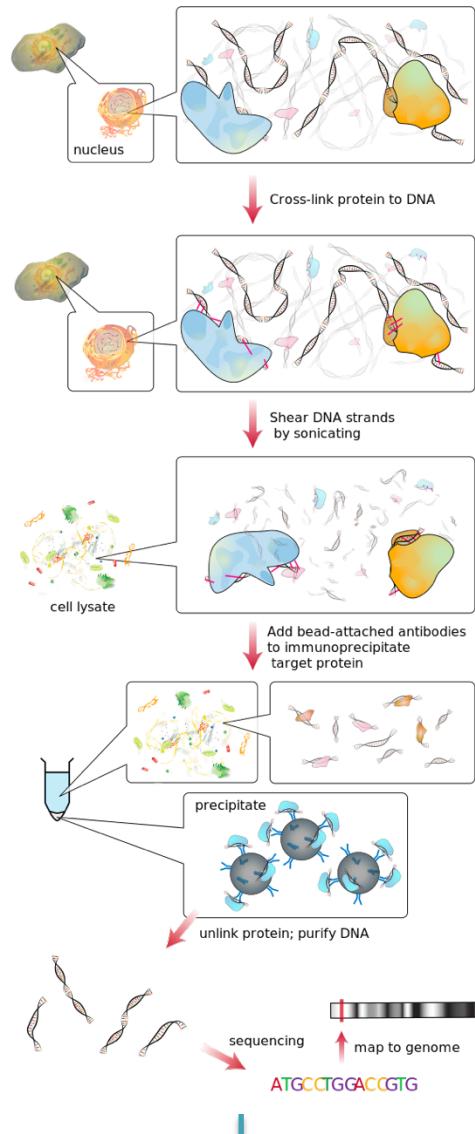


DNA-binding proteins:
TFs, CTCF, repressors
and polymerases

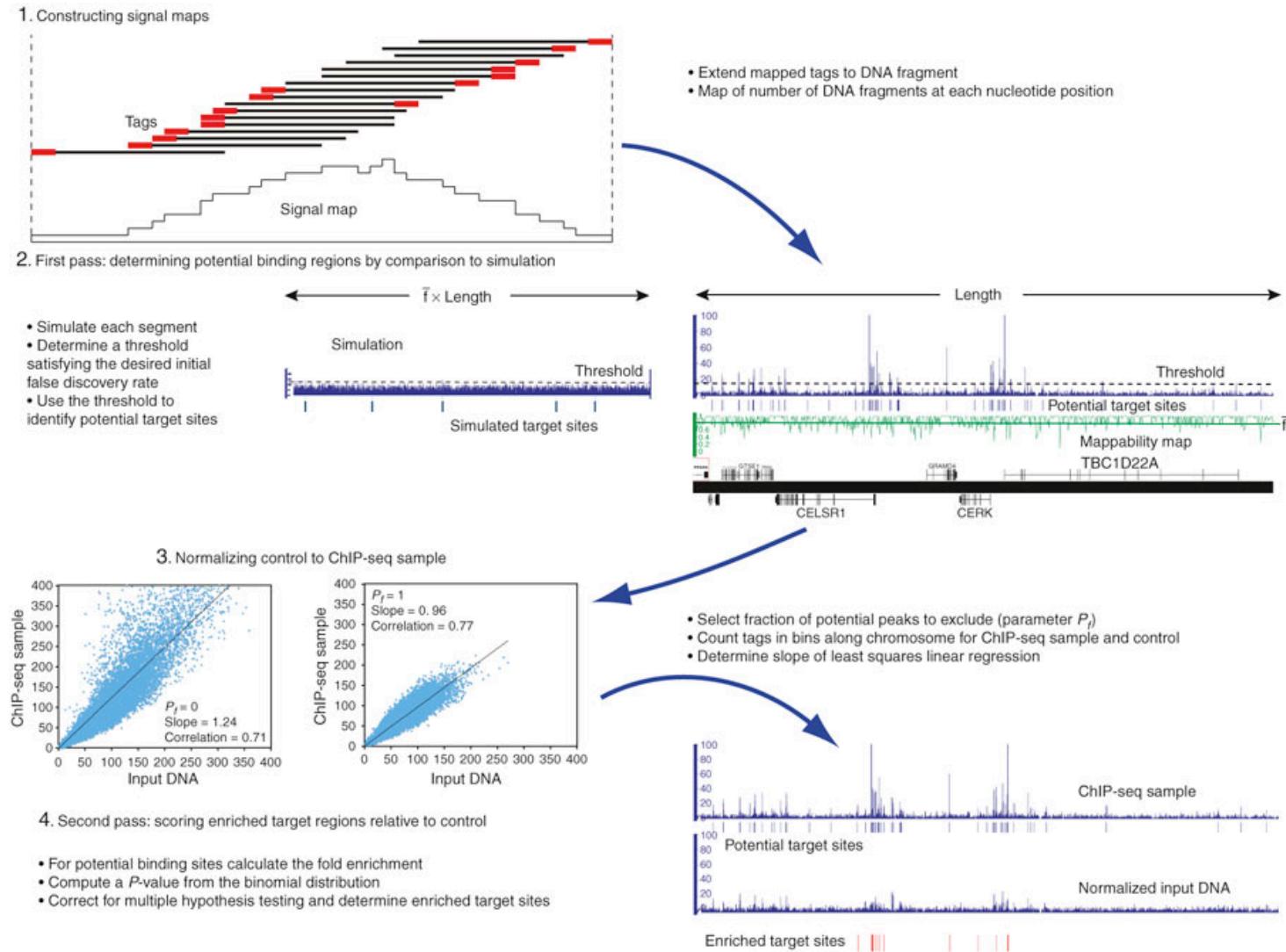
H3K4me1
H3K4me3

H3K27ac
H3K27me3

General Flow of ChIP-seq Analysis

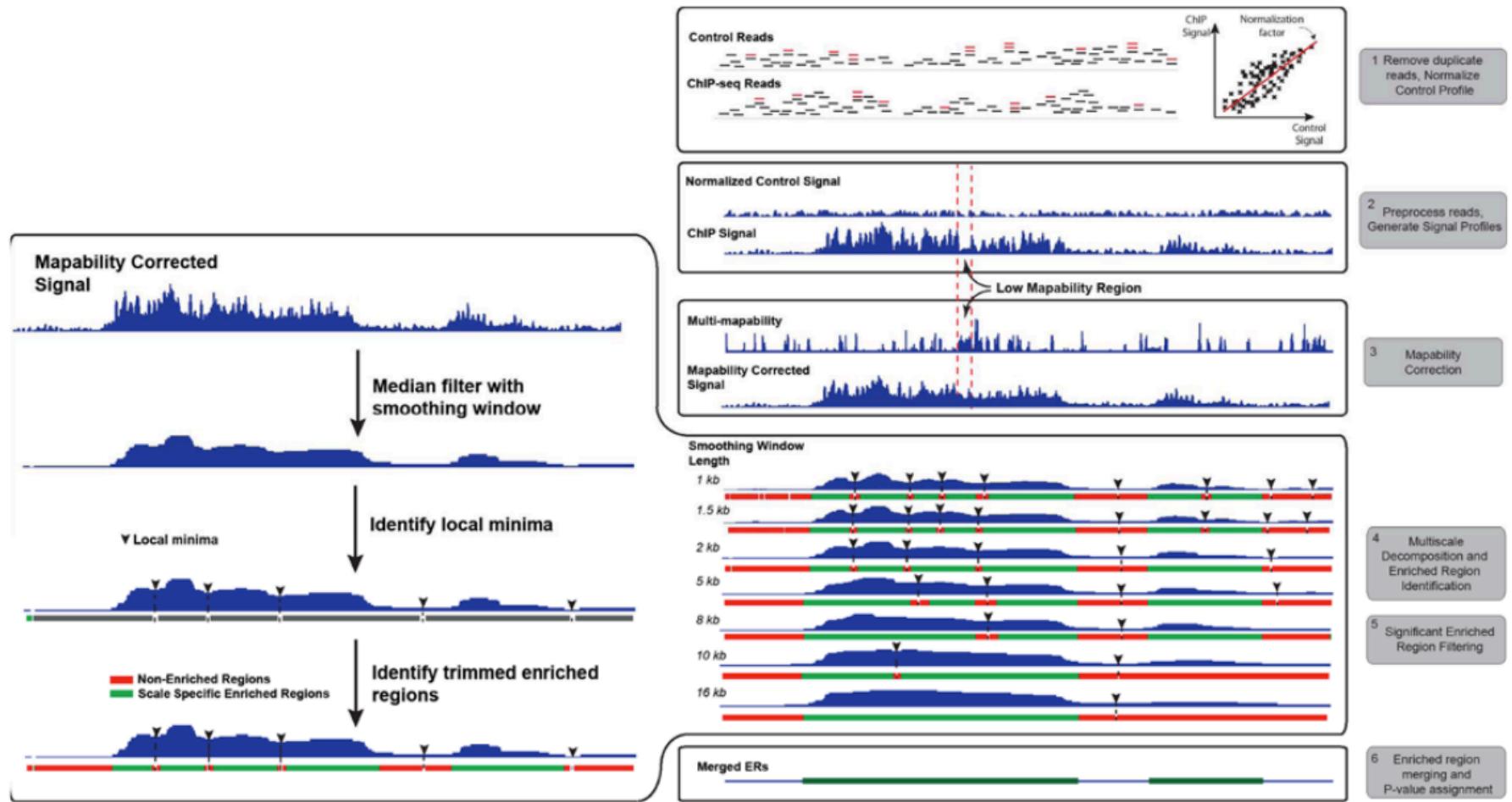


PeakSeq



PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls
 Rozowsky et al (2009) Nature Biotechnology 27, 66 - 75

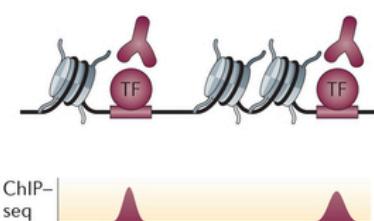
MUSIC



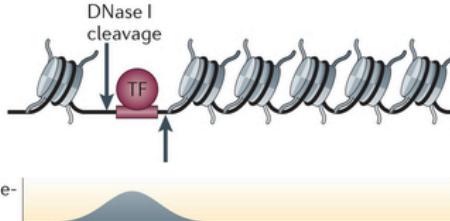
MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework
Harmanci et al. Genome Biology 2014, 15:474

Related Assays

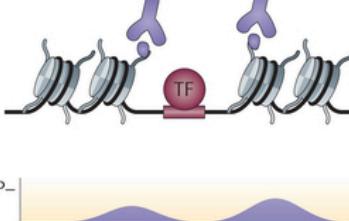
a ChIP-seq for a TF



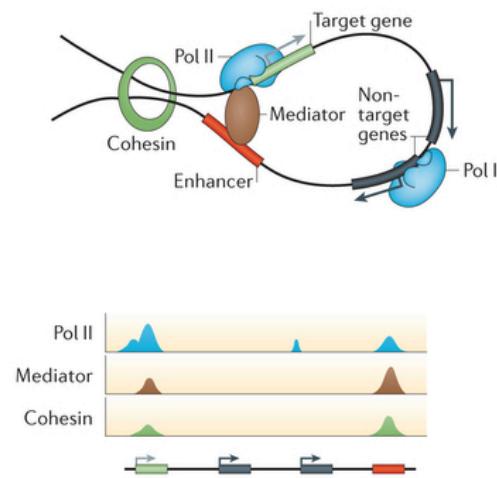
b DNase-seq



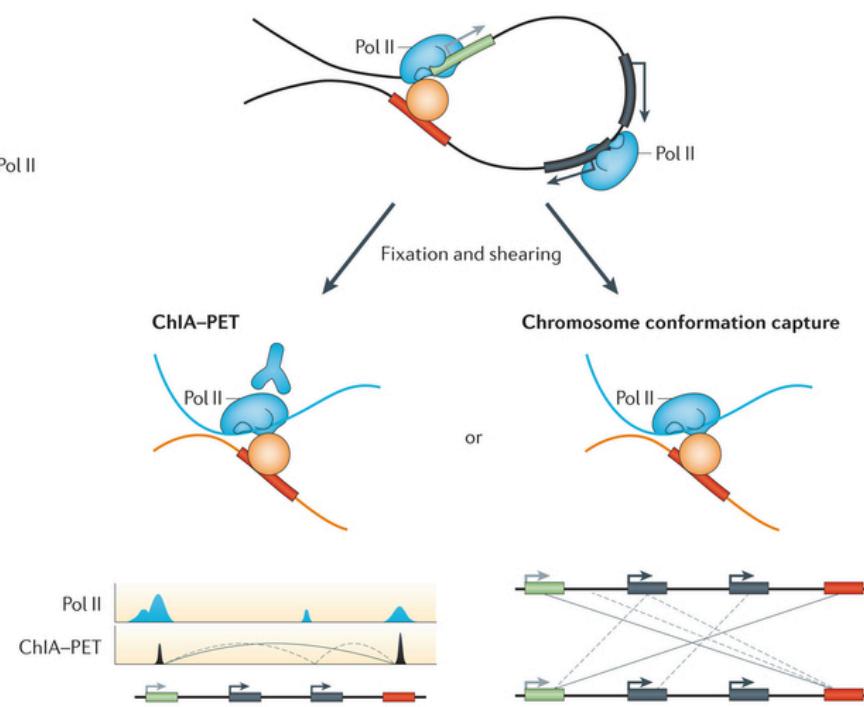
c ChIP-seq for chromatin marks



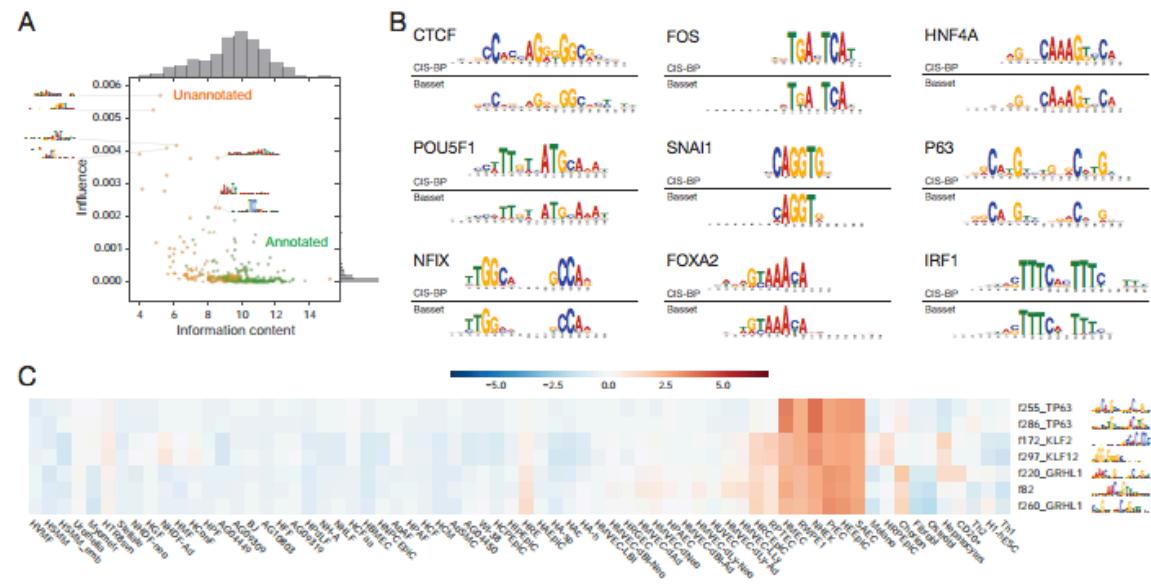
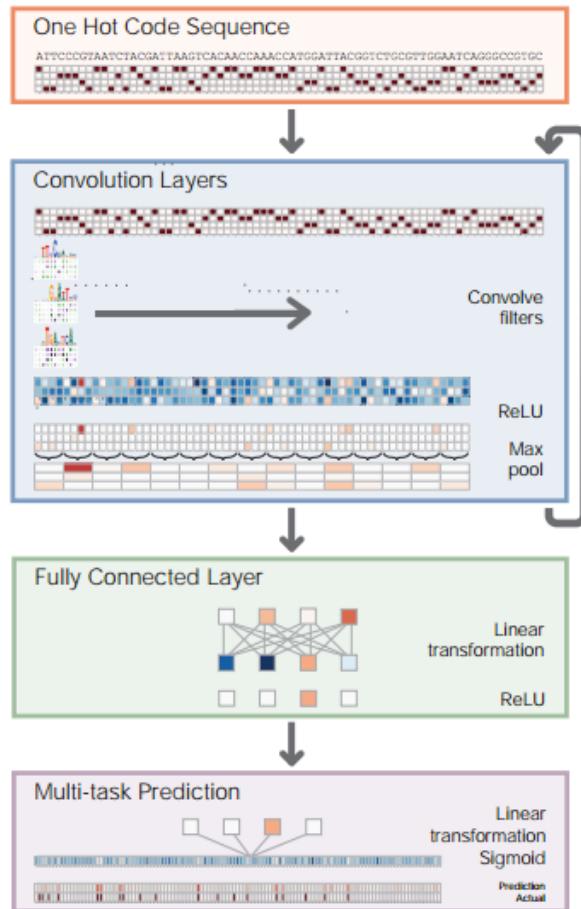
d ChIP-seq for Mediator and cohesin



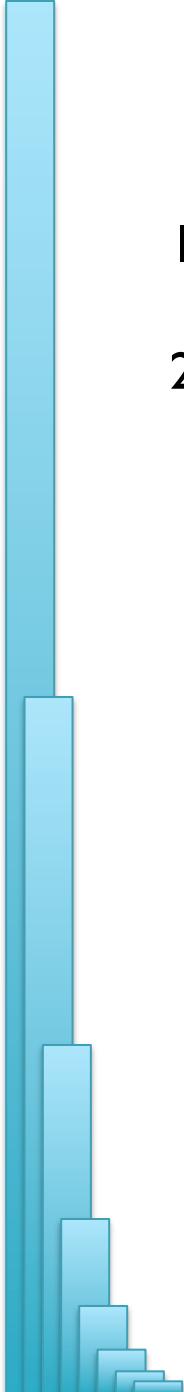
e ChIA-PET and chromosome conformation capture methods



Basset



Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks
Kelley et al. (2016) Genome Research doi: 10.1101/gr.200535.115



Next Steps

1. Questions on assignment 2?
2. Check out the course webpage



Welcome to Applied Comparative Genomics

<https://github.com/schatzlab/appliedgenomics>

Questions?