

Gene Regulation

Michael Schatz

October 30, 2023

Lecture 18. Applied Comparative Genomics



Upcoming events

Mon Oct 30: Regular class | Prelim report assigned

Wed Nov 1: Review class

Mon Nov 6: Midterm exam (1 page of notes allowed)

Wed Nov 8: Regular class | Review exam

Mon Nov 13: Regular class | Prelim report due

Wed Nov 15: Regular class

Mon Nov 20: Thanksgiving break

Wed Nov 22: Thanksgiving break

Mon Nov 27: In class presentation (random order)

Wed Nov 29: In class presentation (random order)

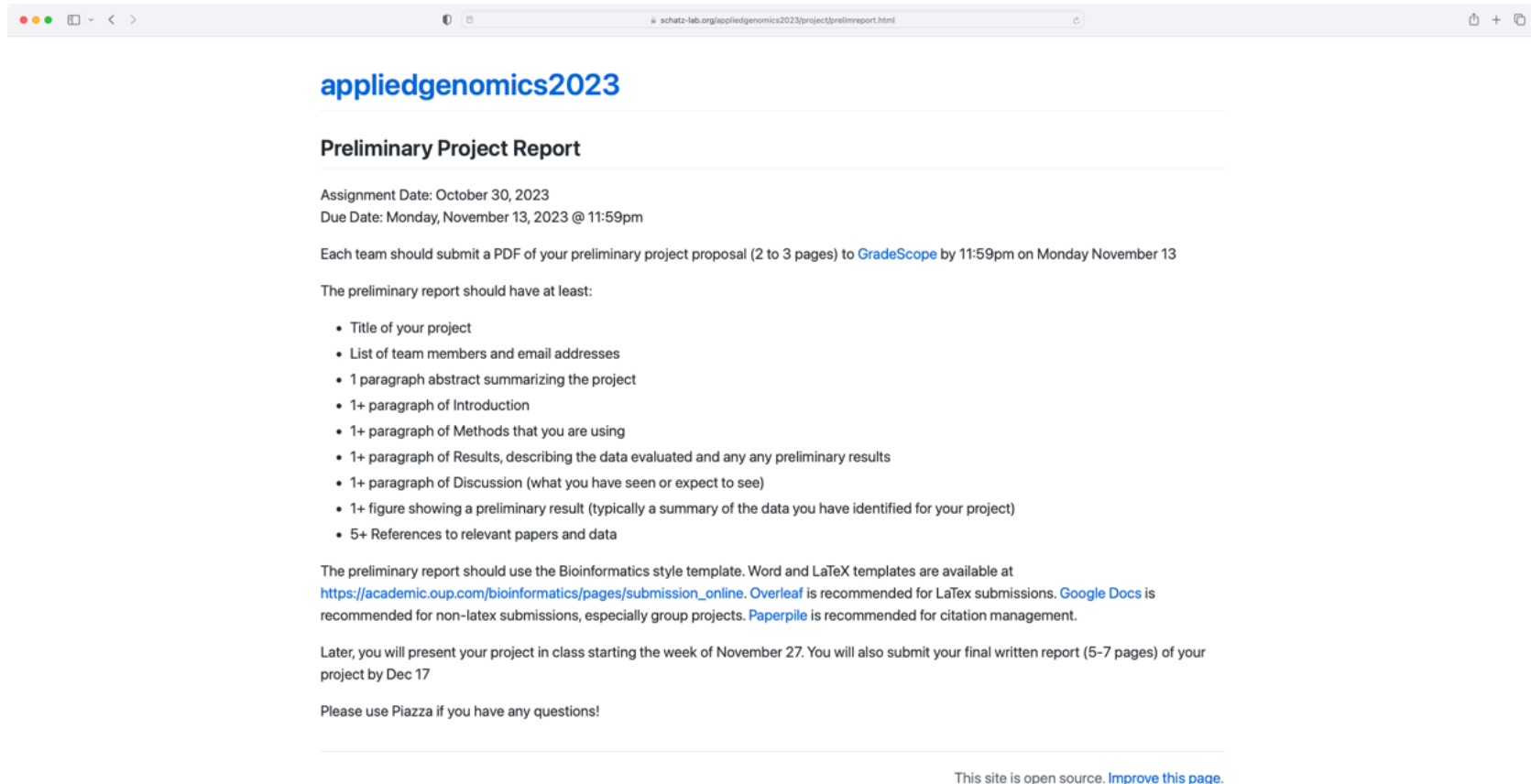
Mon Dec 4: In class presentation (random order)

Wed Dec 6: No class!

Sun Dec 17: Final report due!!!!

Preliminary Report

Due Monday November 13



appliedgenomics2023

Preliminary Project Report

Assignment Date: October 30, 2023
Due Date: Monday, November 13, 2023 @ 11:59pm

Each team should submit a PDF of your preliminary project proposal (2 to 3 pages) to [GradeScope](#) by 11:59pm on Monday November 13

The preliminary report should have at least:

- Title of your project
- List of team members and email addresses
- 1 paragraph abstract summarizing the project
- 1+ paragraph of Introduction
- 1+ paragraph of Methods that you are using
- 1+ paragraph of Results, describing the data evaluated and any any preliminary results
- 1+ paragraph of Discussion (what you have seen or expect to see)
- 1+ figure showing a preliminary result (typically a summary of the data you have identified for your project)
- 5+ References to relevant papers and data

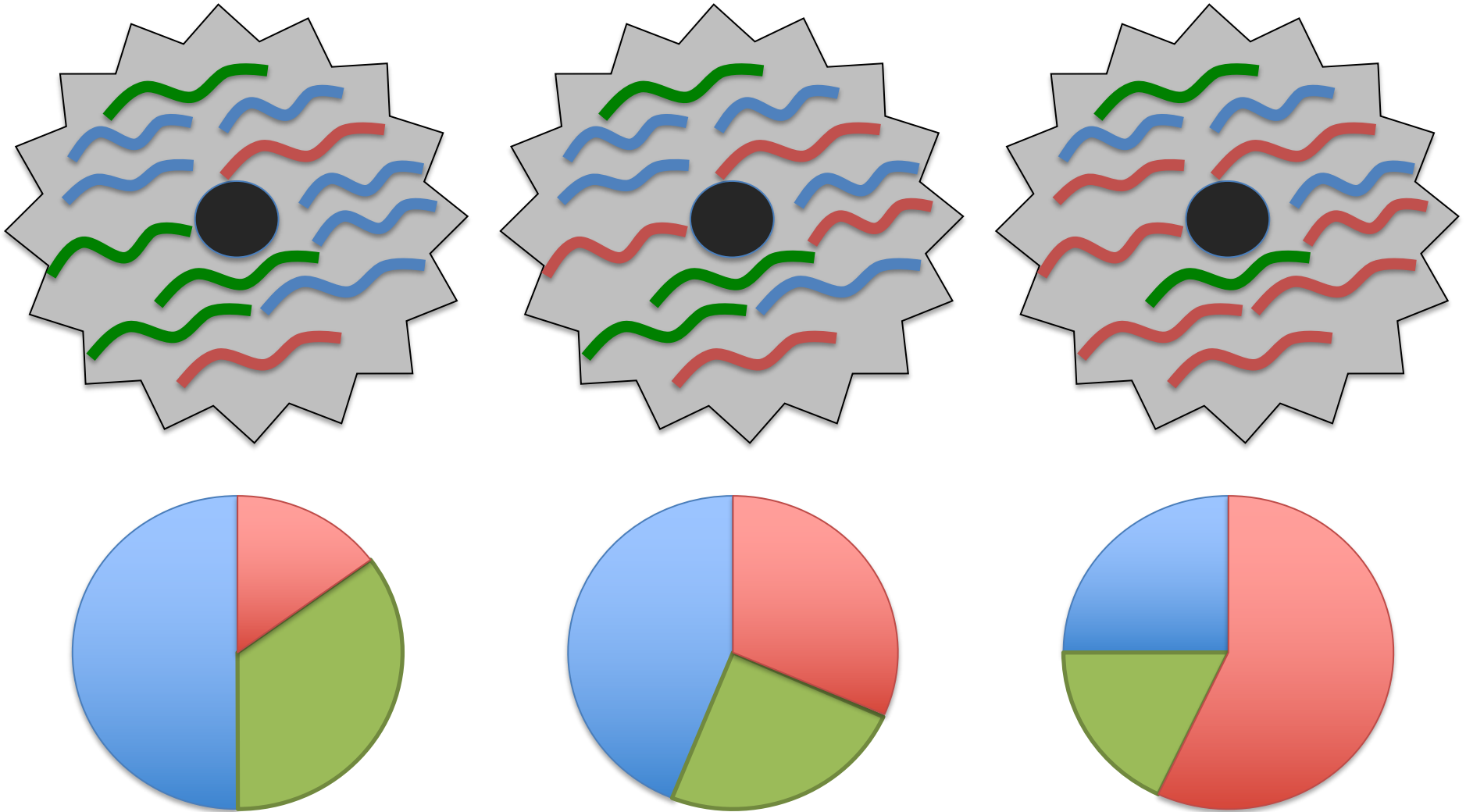
The preliminary report should use the Bioinformatics style template. Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online. [Overleaf](#) is recommended for LaTeX submissions. [Google Docs](#) is recommended for non-latex submissions, especially group projects. [Paperpile](#) is recommended for citation management.

Later, you will present your project in class starting the week of November 27. You will also submit your final written report (5-7 pages) of your project by Dec 17

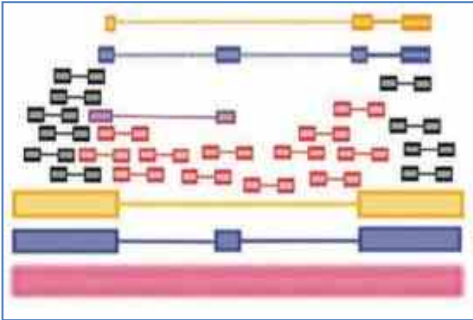
Please use Piazza if you have any questions!

This site is open source. [Improve this page](#).

RNA-seq Overview



RNA-seq Challenges

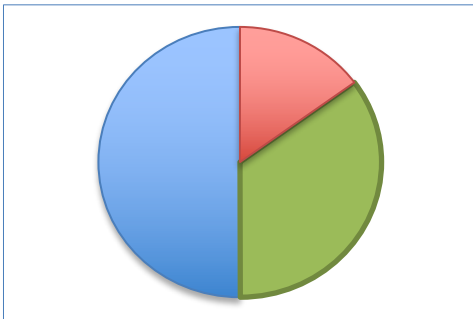


Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

TopHat: discovering spliced junctions with RNA-Seq.

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

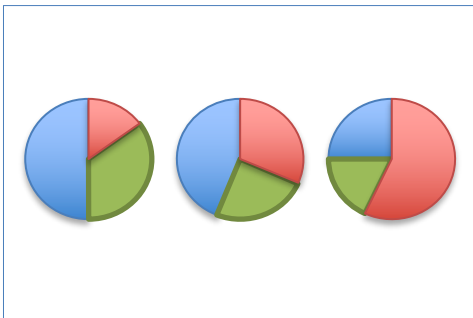


Challenge 2: Read Count != Transcript abundance

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

Transcript assembly and quantification by RNA-seq

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



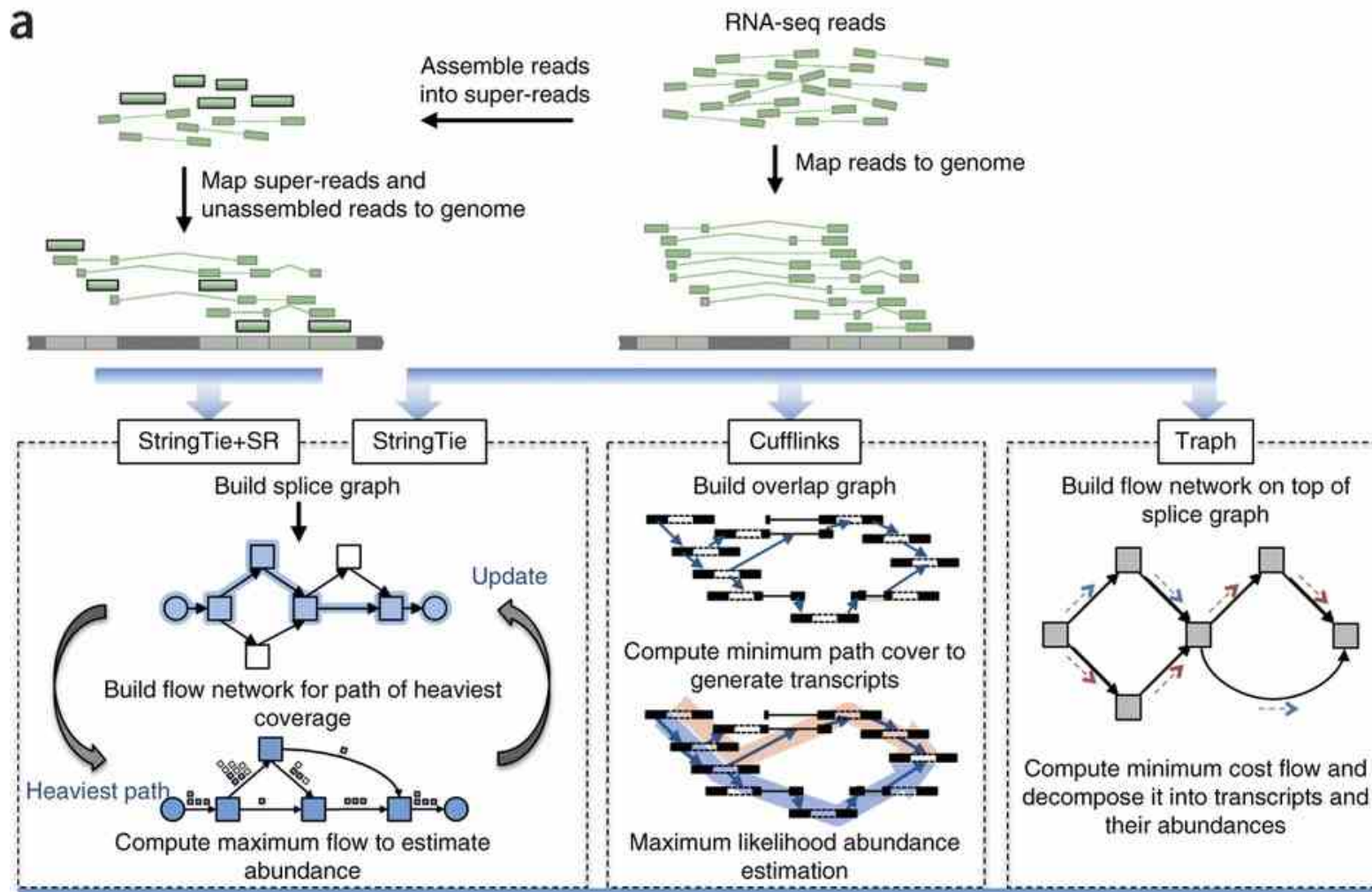
Challenge 3: Transcript abundances are stochastic

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

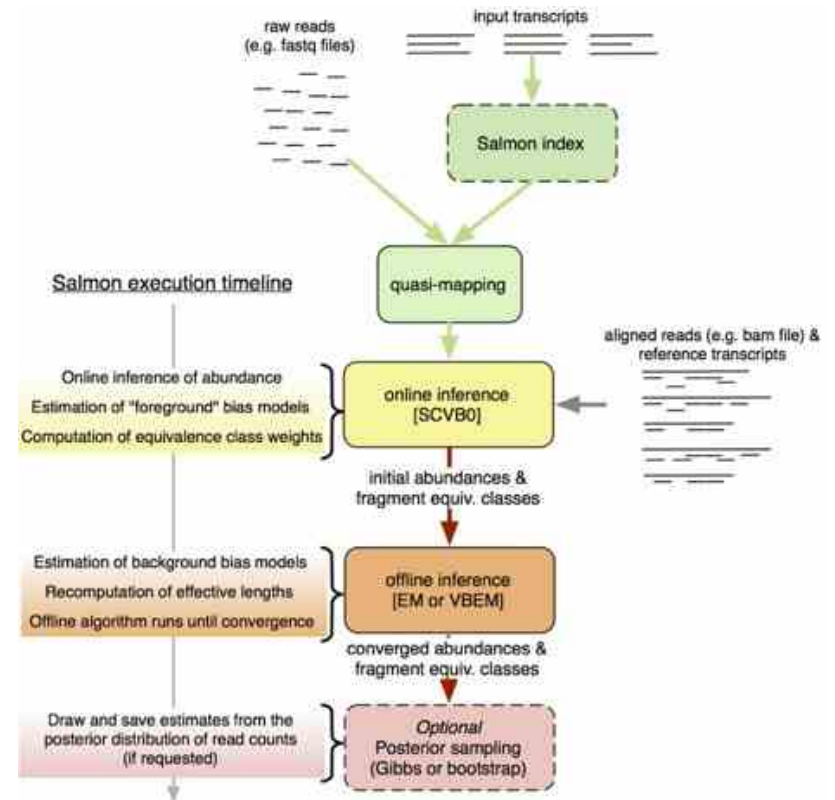
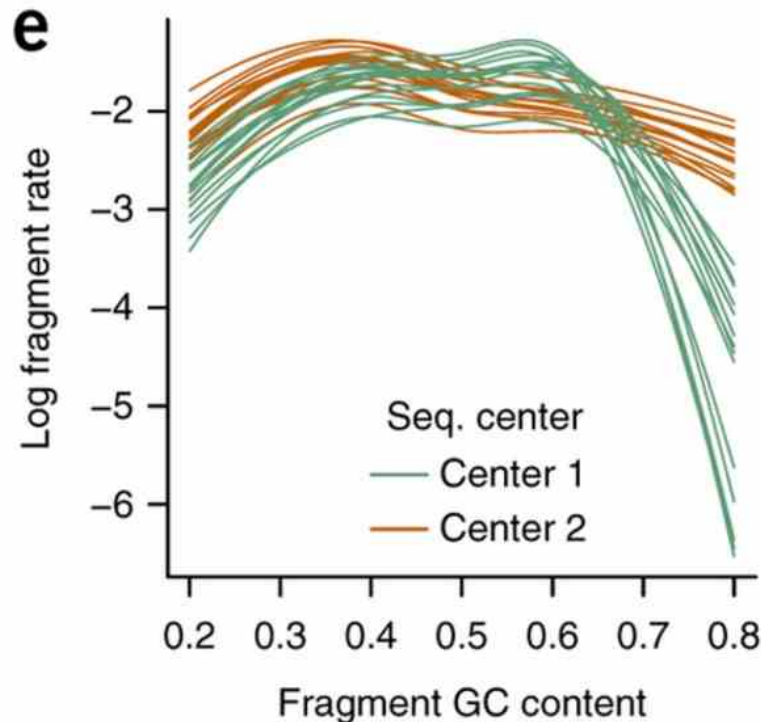
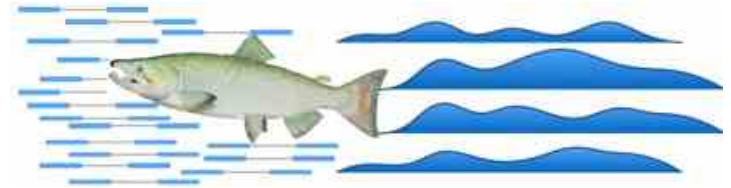
Isoform Quantification Approaches



StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.

Pertea M, et al. (2015) Nature Biotechnology. doi: 10.1038/nbt.3122.

Salmon: The ultimate RNA-seq Pipeline?



Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation

Love et al (2016) Nature Biotechnology 34, 1287–1291 (2016) doi:10.1038/nbt.3682

Salmon provides fast and bias-aware quantification of transcript expression

Patro et al (2017) Nature Methods (2017) doi:10.1038/nmeth.4197

Annotation Summary

- Three major approaches to annotate a genome

1. Alignment:

- Does this sequence align to any other sequences of known function?
- Great for projecting knowledge from one species to another

2. Prediction:

- Does this sequence statistically resemble other known sequences?
- Potentially most flexible but dependent on good training data

3. Experimental:

- Lets test to see if it is transcribed/methylated/bound/etc
- Strongest but expensive and context dependent

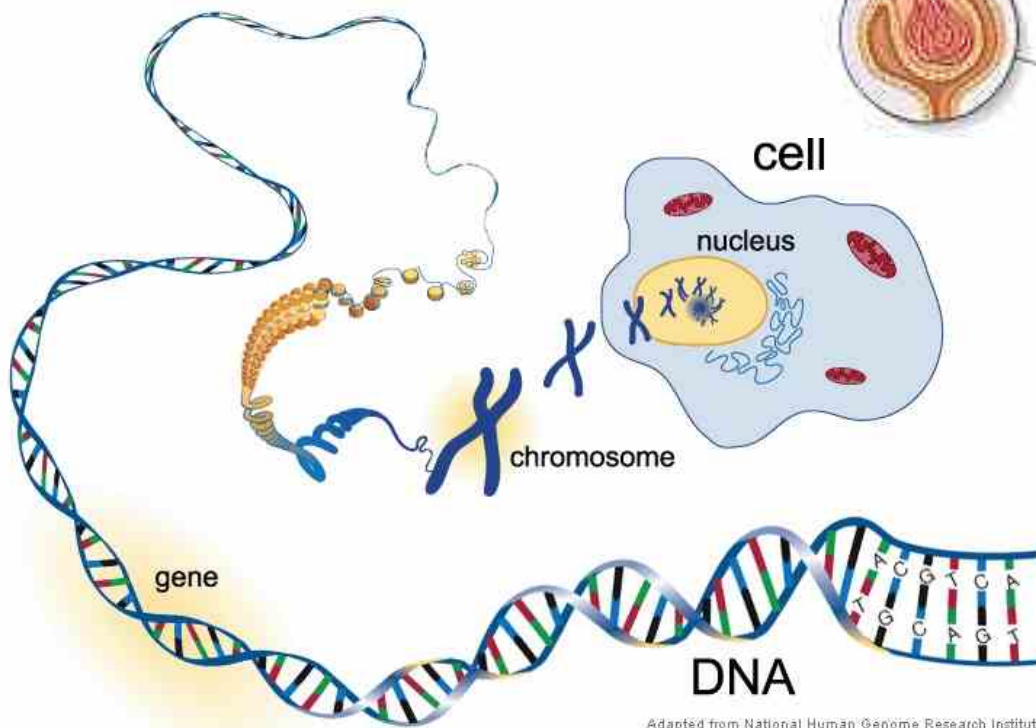
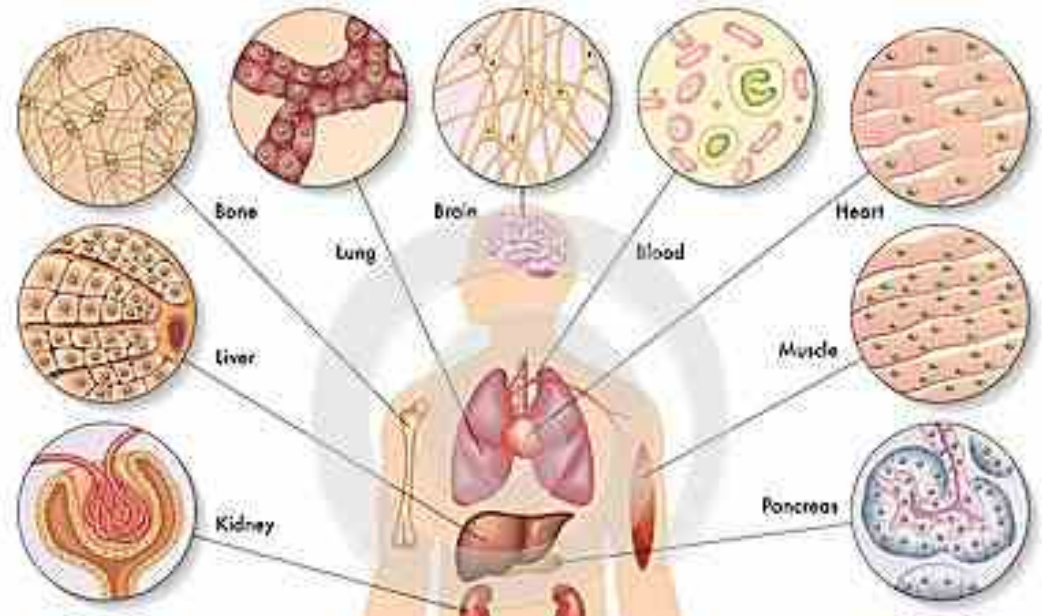
- Many great resources available

- Learn to love the literature and the databases
- Standard formats let you rapidly query and cross reference
- Google is your number one resource 😊



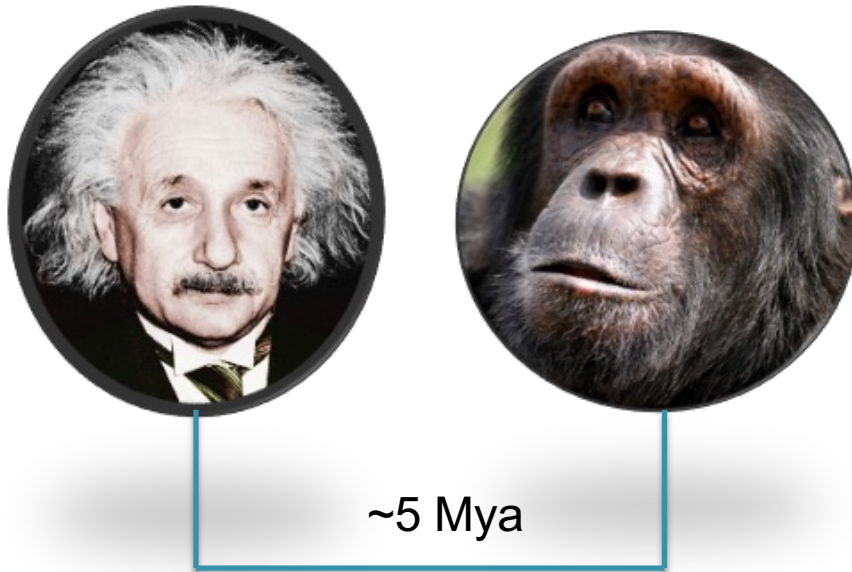
Why Genes?

Each cell of your body contains an exact copy of your 3 billion base pair genome.



Your body has a few hundred (thousands?) major cell types, largely defined by the gene expression patterns

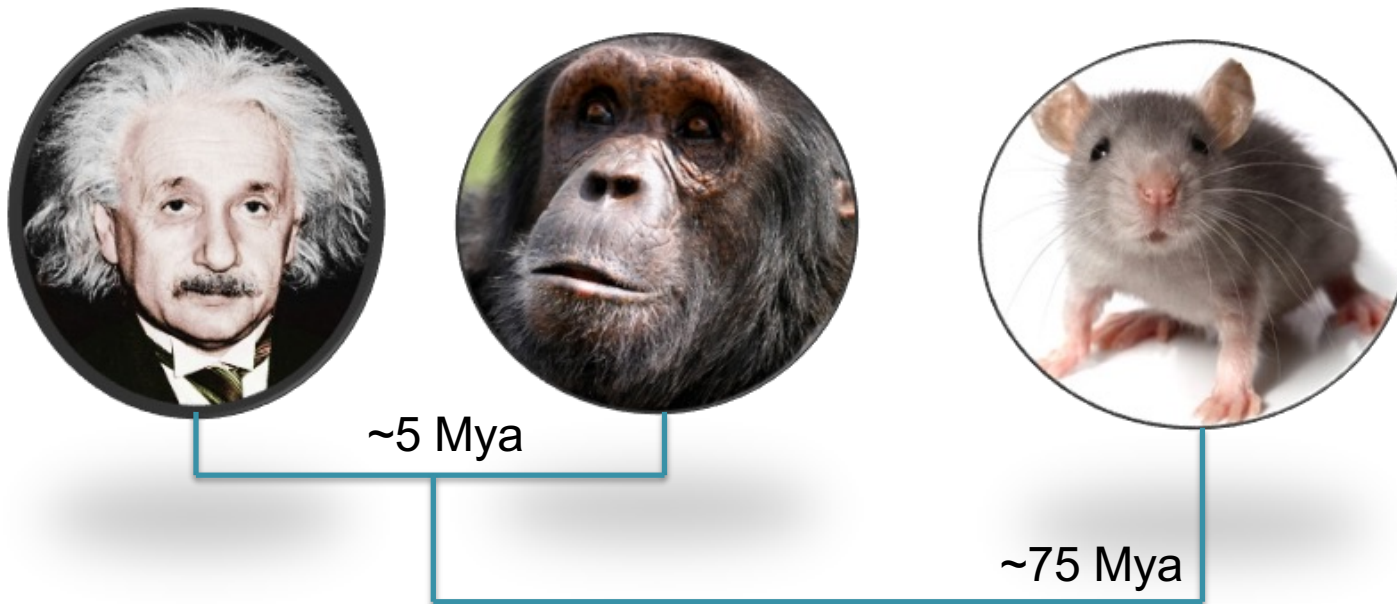
Human Evolution



- 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



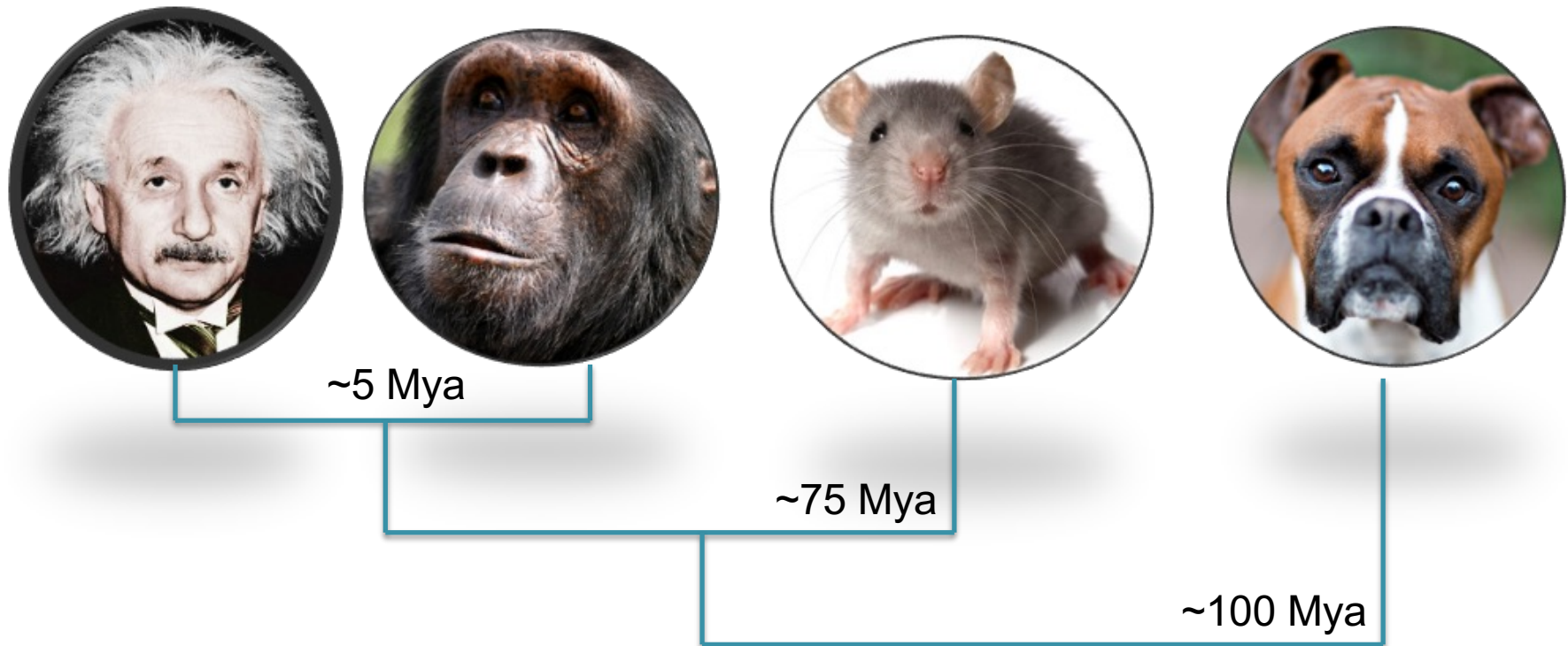
“In the roughly 75 million years since the divergence of the human and mouse lineages, the process of evolution has altered their genome sequences and caused them to diverge by ***nearly one substitution for every two nucleotides***”

“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

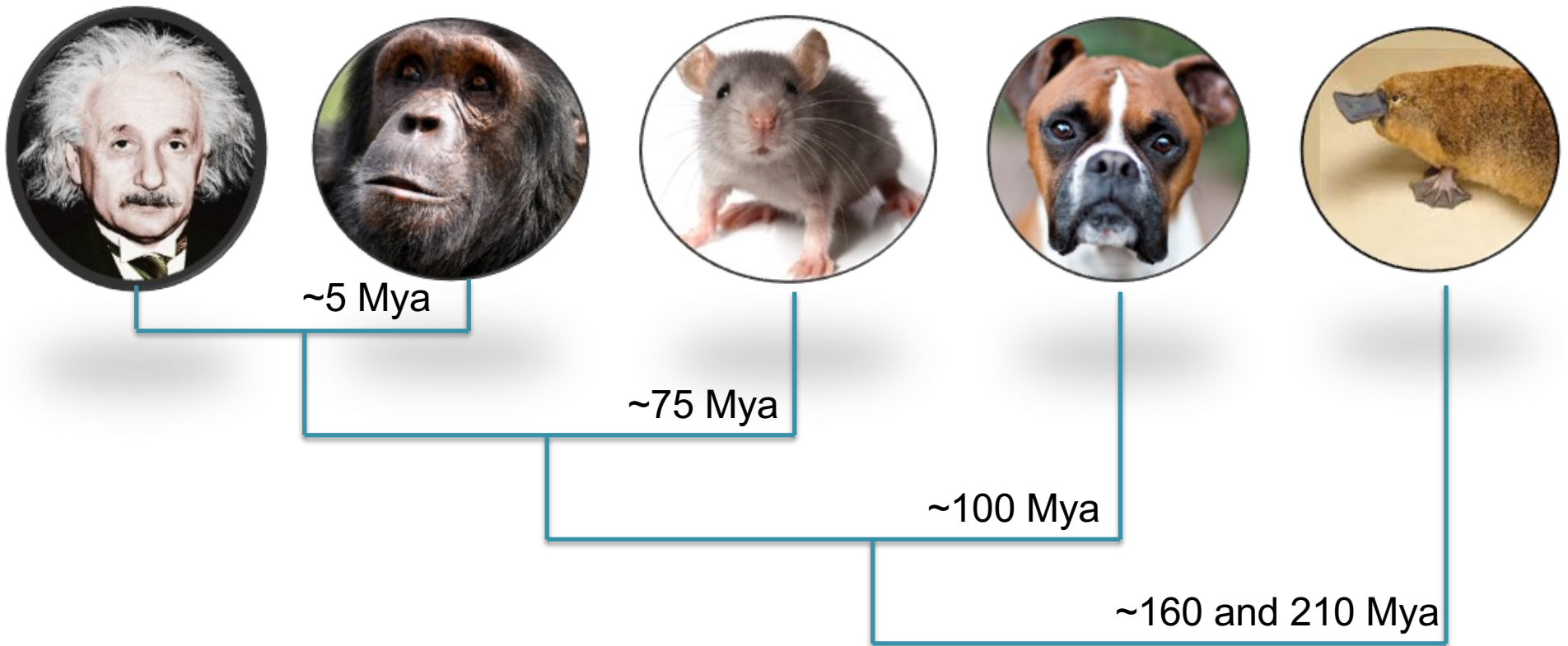


“We generated gene predictions for the dog genome using an evidence-based method (see Supplementary Information). The resulting collection contains **19,300 dog gene predictions, with nearly all being clear homologues of known human genes**. The dog gene count is substantially lower than the ~22,000-gene 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

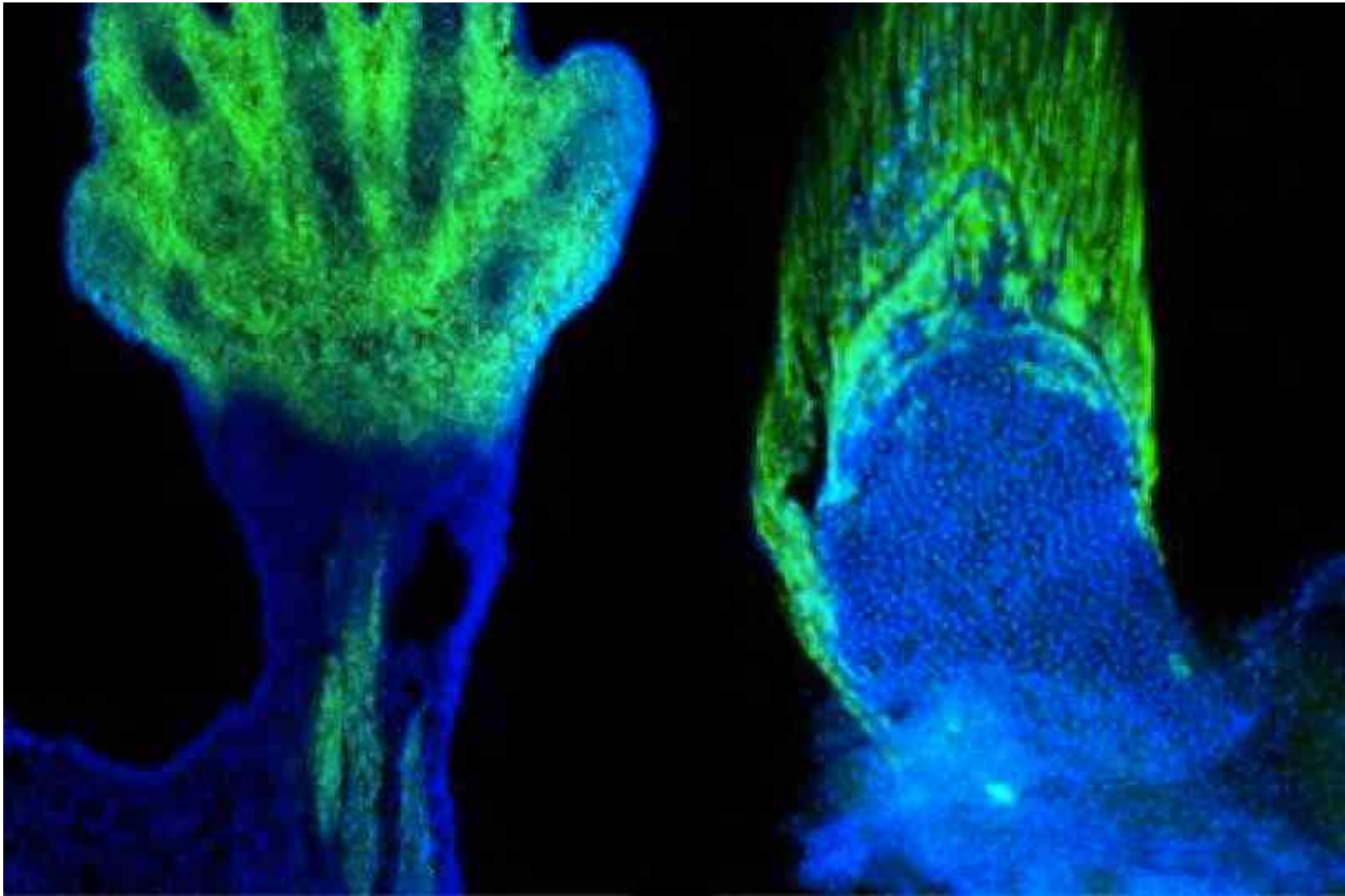
Human Evolution



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

Animal Evolution



Digits and fin rays share common developmental histories

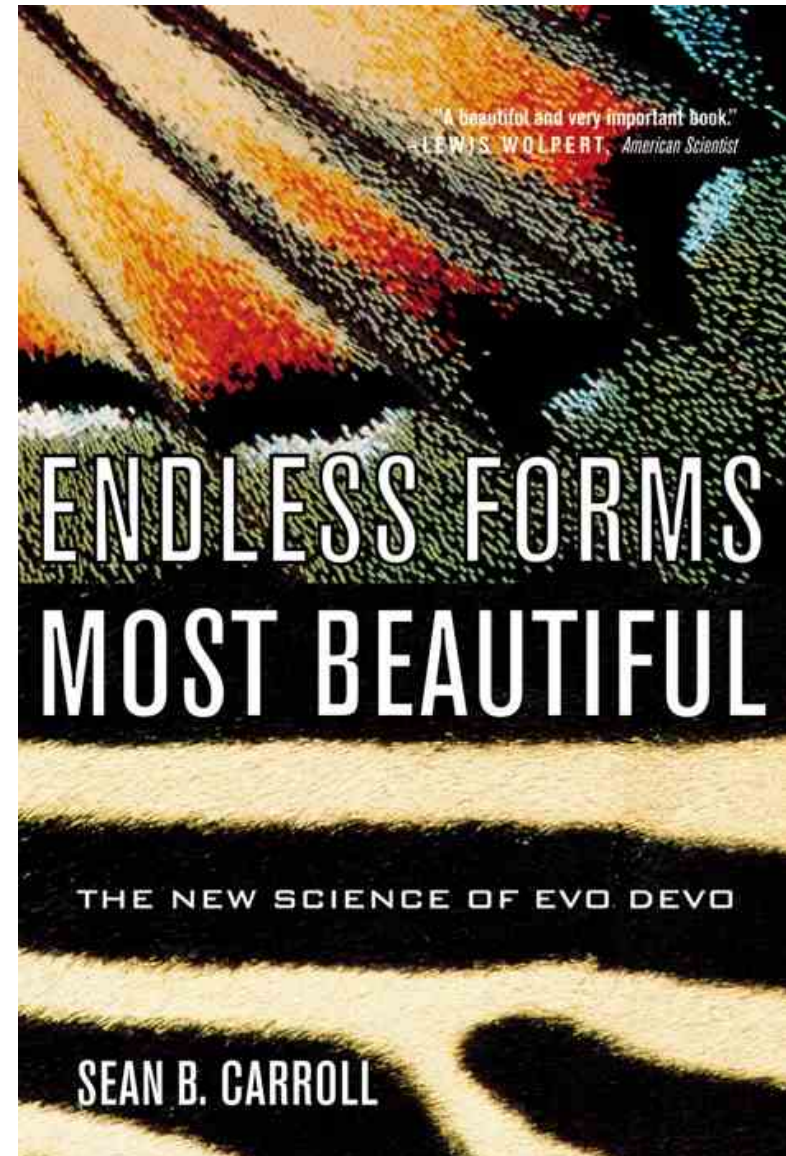
Nakamura et al (2016) *Nature*. 537, 225–228. doi:10.1038/nature19322

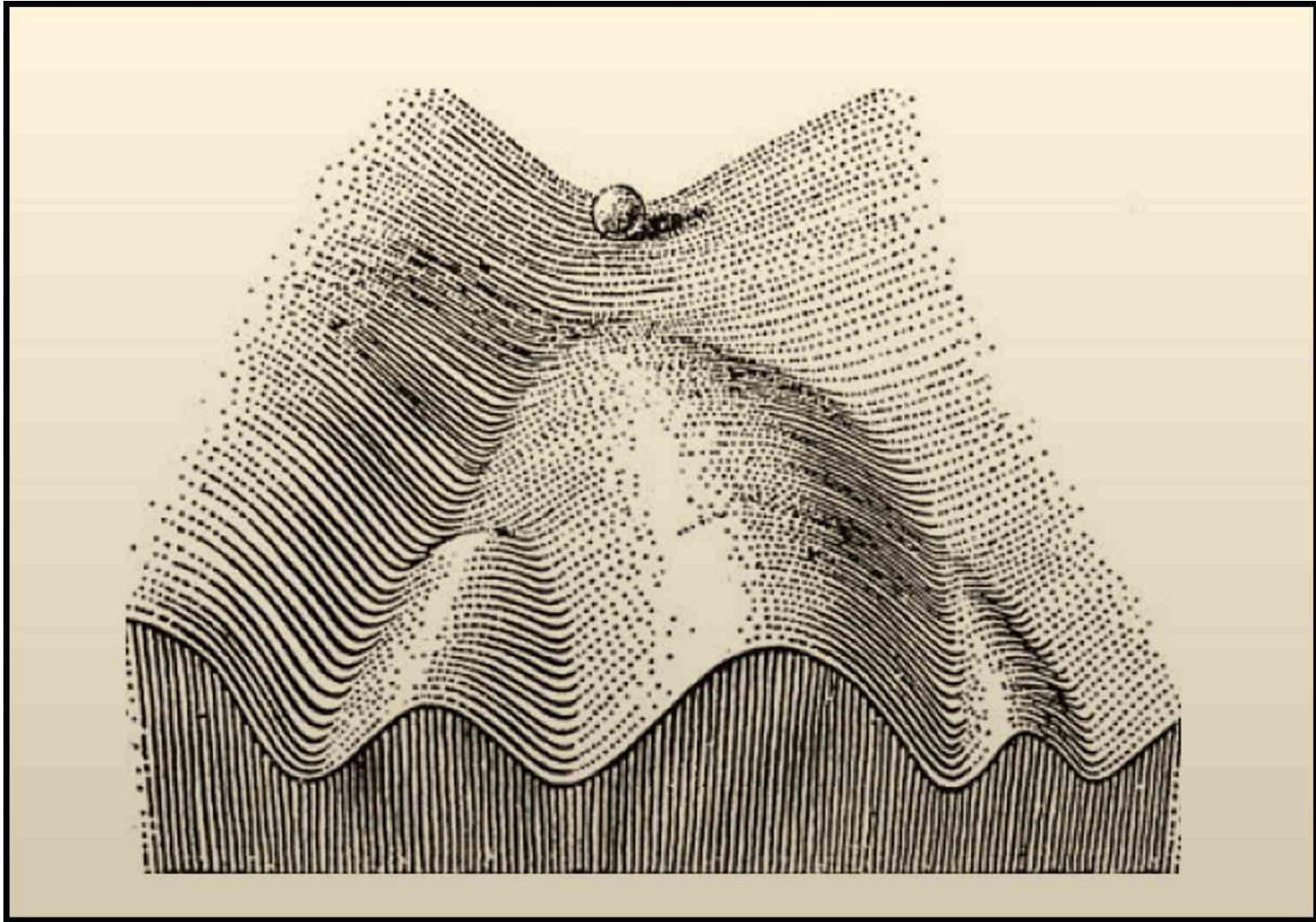
More Information



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

-Jacques Monod

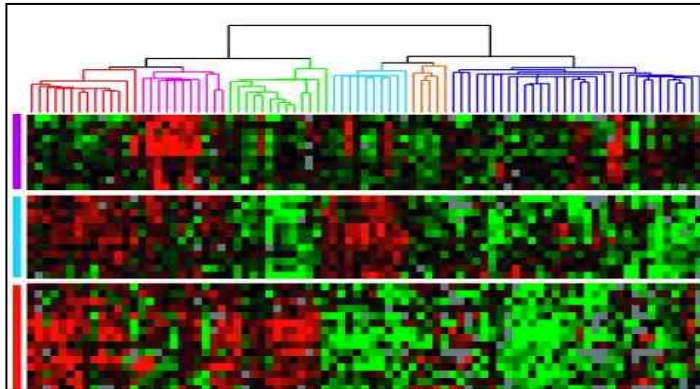




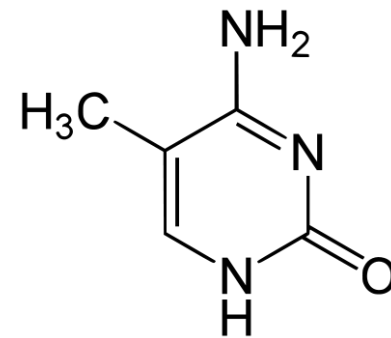
The Strategy of the Genes
CH Waddington (1957)

*-seq in 4 short vignettes

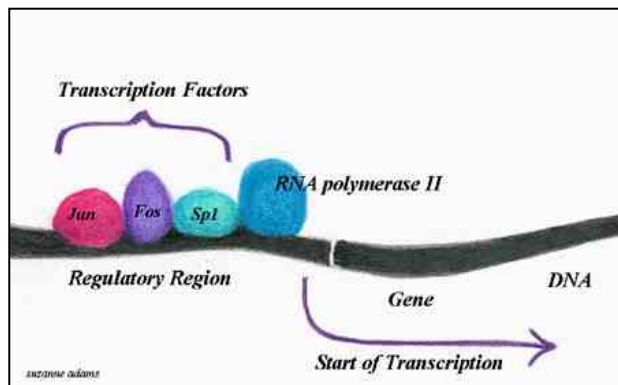
RNA-seq



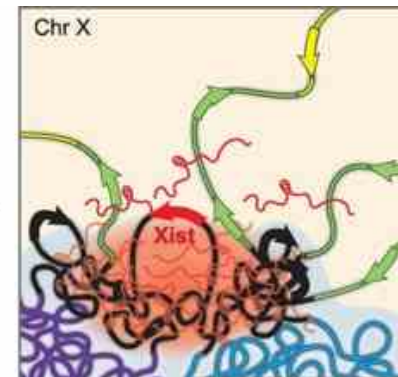
Methyl-seq



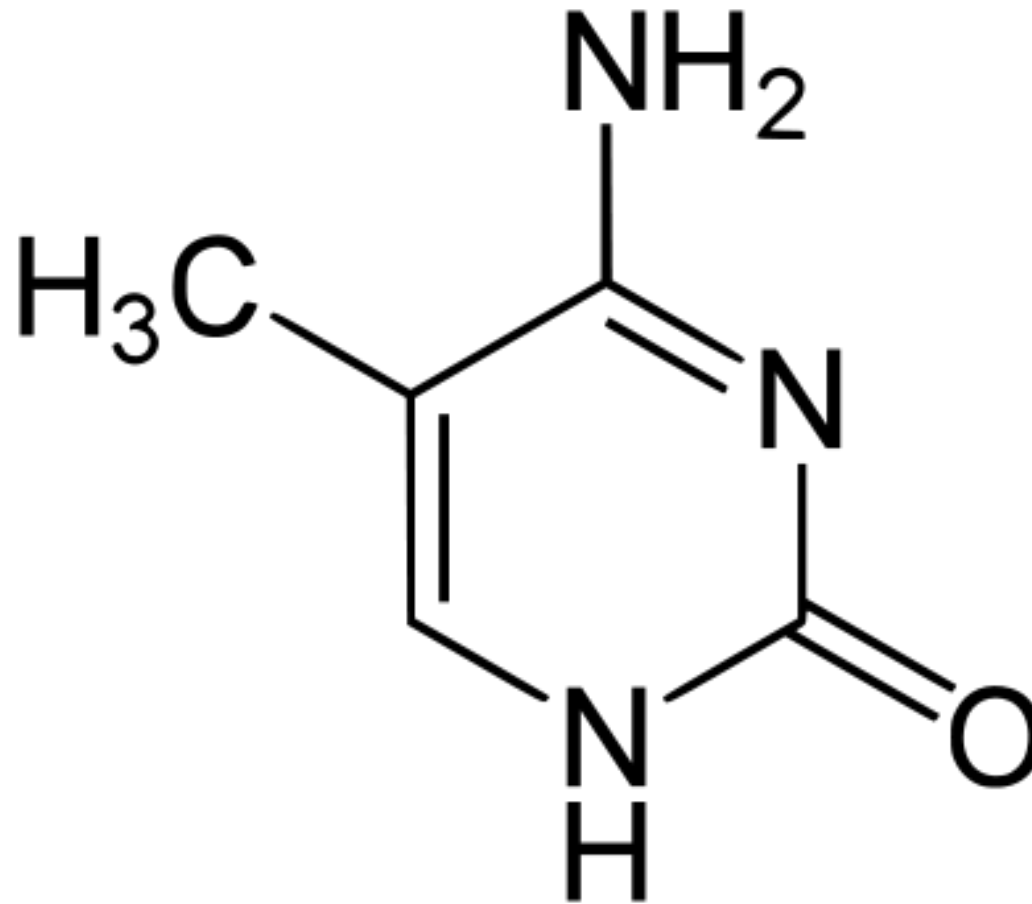
ChIP-seq



Hi-C



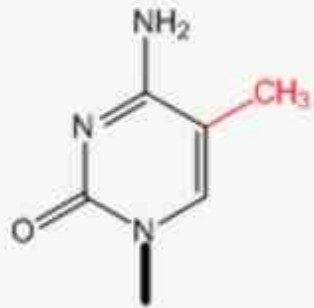
Methyl-seq



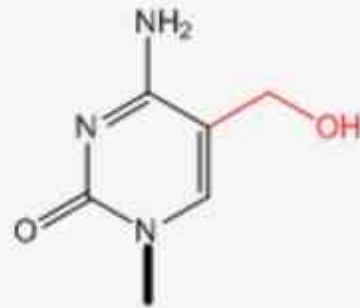
Finding the fifth base: Genome-wide sequencing of cytosine methylation

Lister and Ecker (2009) *Genome Research*. 19: 959-966

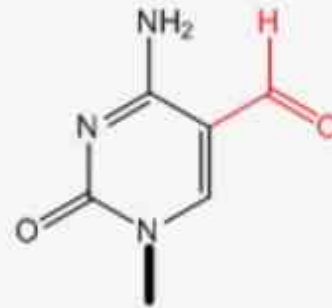
Epigenetic Modifications to DNA



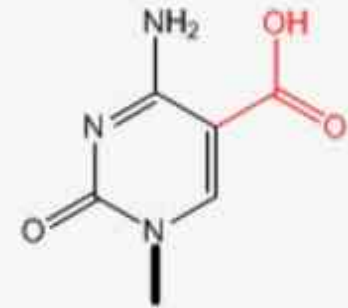
5-mC



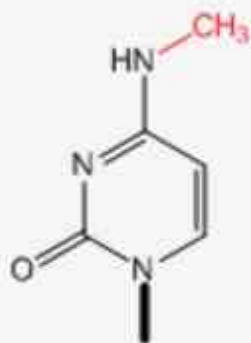
5-hmC



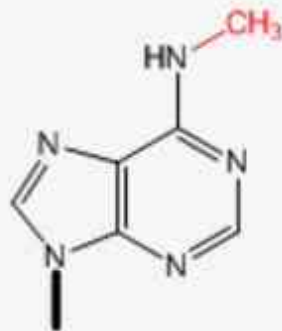
5-fC



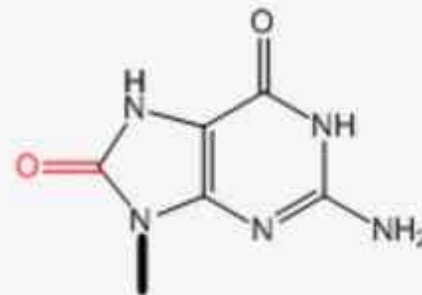
5-caC



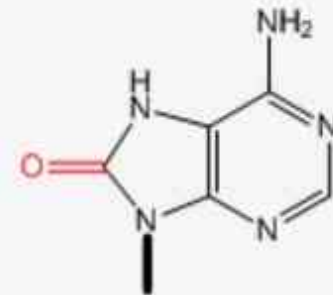
4-mC



6-mA



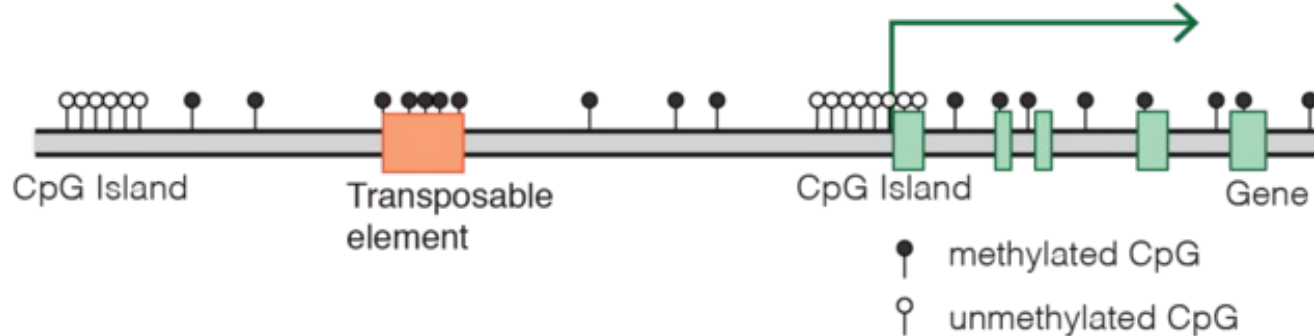
8-oxoG



8-oxoA

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.

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





Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm

Ong-Abdullah, et al (2015) *Nature*. doi:10.1038/nature15365



Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm

Ong-Abdullah, et al (2015) *Nature*. doi:10.1038/nature15365



Somaclonal variation arises in plants and animals when differentiated somatic cells are induced into a pluripotent state, but the resulting clones differ from each other and from their parents. In agriculture, somaclonal variation has hindered the micropropagation of elite hybrids and genetically modified crops, but the mechanism responsible remains unknown. The oil palm fruit 'mantled' abnormality is a somaclonal variant arising from tissue culture that drastically reduces yield, and has largely halted efforts to clone elite hybrids for oil production. Widely regarded as an epigenetic phenomenon, 'mantling' has defied explanation, but here we identify the MANTLED locus using epigenome-wide association studies of the African oil palm *Elaeis guineensis*. DNA hypomethylation of a LINE retrotransposon related to rice Karma, in the intron of the homeotic gene *DEFICIENS*, is common to all mantled clones and is associated with alternative splicing and premature termination. **Dense methylation near the Karma splice site (termed the Good Karma epiallele) predicts normal fruit set, whereas hypomethylation (the Bad Karma epiallele) predicts homeotic transformation, parthenocarpy and marked loss of yield.** Loss of Karma methylation and of small RNA in tissue culture contributes to the origin of mantled, while restoration in spontaneous revertants accounts for non-Mendelian inheritance. The ability to predict and cull mantling at the plantlet stage will facilitate the introduction of higher performing clones and optimize environmentally sensitive land resources.

Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm

Ong-Abdullah, et al (2015) *Nature*. doi:10.1038/nature15365

Hypomethylation distinguishes genes of some human cancers from their normal counterparts

Andrew P. Feinberg & Bert Vogelstein

Cell Structure and Function Laboratory, The Oncology Center, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

It has been suggested that cancer represents an alteration in DNA, heritable by progeny cells, that leads to abnormally regulated expression of normal cellular genes; DNA alterations such as mutations^{1,2}, rearrangements³⁻⁵ and changes in methylation⁶⁻⁸ have been proposed to have such a role. Because of increasing evidence that DNA methylation is important in gene expression (for review see refs 7, 9-11), several investigators have studied DNA methylation in animal tumours, transformed cells and leukaemia cells in culture^{8,12-30}. The results of these studies have varied; depending on the techniques and systems used, an increase¹²⁻¹⁹, decrease²⁰⁻²⁴, or no change²⁵⁻²⁹ in the degree of methylation has been reported. To our knowledge, however, primary human tumour tissues have not been used in such studies. We have now examined DNA methylation in human cancer with three considerations in mind: (1) the methylation pattern of specific genes, rather than total levels of methylation, was determined; (2) human cancers and adjacent analogous normal tissues, unconditioned by culture media, were analysed; and (3) the cancers were taken from patients who had received neither radiation nor chemotherapy. In four of five patients studied, representing two histological types of cancer, substantial hypomethylation was found in genes of cancer cells compared with their normal counterparts. This hypomethylation was progressive in a metastasis from one of the patients.

and (3) *Hpa*II and *Hha*I cleavage sites should be present in the regions of the genes.

The first cancer studied was a grade D (ref. 43), moderately well differentiated adenocarcinoma of the colon from a 67-yr-old male. Tissue was obtained from the cancer itself and also from colonic mucosa stripped from the colon at a site just outside the histologically proven tumour margin. Figure 1 shows the pattern of methylation of the studied genes. Before digestion with restriction enzymes, all DNA samples used in the study had a size >25,000 base pairs (bp). After *Hpa*II cleavage, hybridization with a probe made from a cDNA clone of human growth hormone (HGH) showed that significantly more of the DNA was digested to low-molecular weight fragments in DNA from the cancer (labelled C in Fig. 1) than in DNA from the normal colonic mucosa (labelled N). In the hybridization conditions used, the HGH probe detected the human growth hormone genes as well as the related chorionic somatotropin

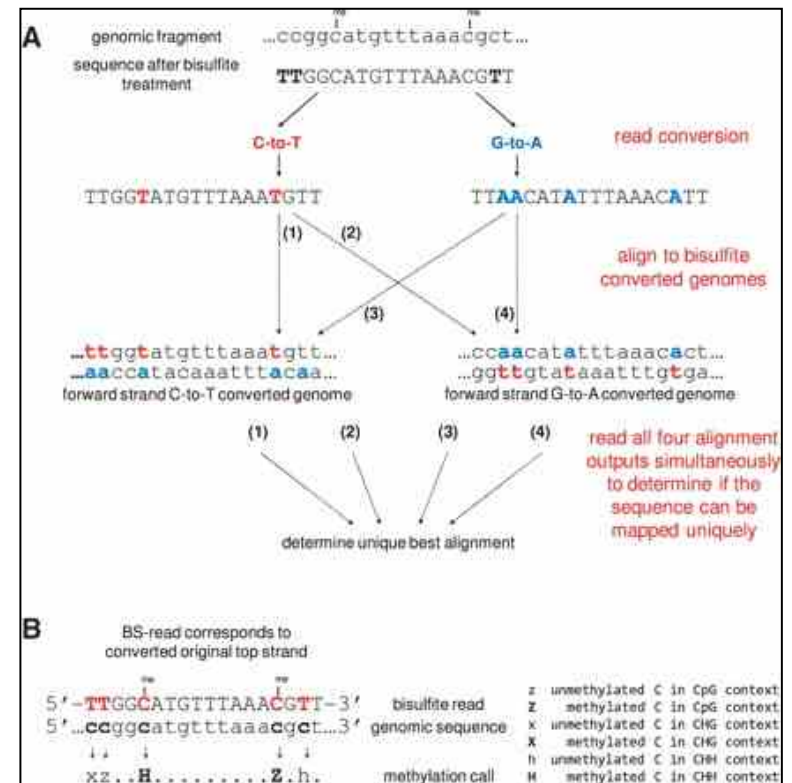
Table 1 Quantitation of methylation of specific genes in human cancers and adjacent analogous normal tissues

Patient	Carcinoma	Probe	Enzyme	% Hypomethylated fragments		
				N	C	M
1	Colon	HGH	{ <i>Hpa</i> II	<10	35	—
			{ <i>Hha</i> I	<10	39	—
		γ -Globin	{ <i>Hpa</i> II	<10	52	—
			{ <i>Hha</i> I	<10	39	—
		α -Globin	{ <i>Hpa</i> II	<10	<10	—
			{ <i>Hha</i> I	<10	<10	—
2	Colon	HGH	{ <i>Hpa</i> II	<10	76	—
			{ <i>Hha</i> I	<10	85	—
		γ -Globin	{ <i>Hpa</i> II	<10	58	—
			{ <i>Hha</i> I	<10	23	—
		α -Globin	{ <i>Hpa</i> II	<10	<10	—
			{ <i>Hha</i> I	<10	<10	—
3	Colon	HGH	{ <i>Hpa</i> II	<10	41	—
			{ <i>Hha</i> I	<10	38	—
		γ -Globin	{ <i>Hpa</i> II	<10	50	—
			{ <i>Hha</i> I	<10	22	—

Bisulfite Conversion

Treating DNA with sodium bisulfite will convert unmethyated **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

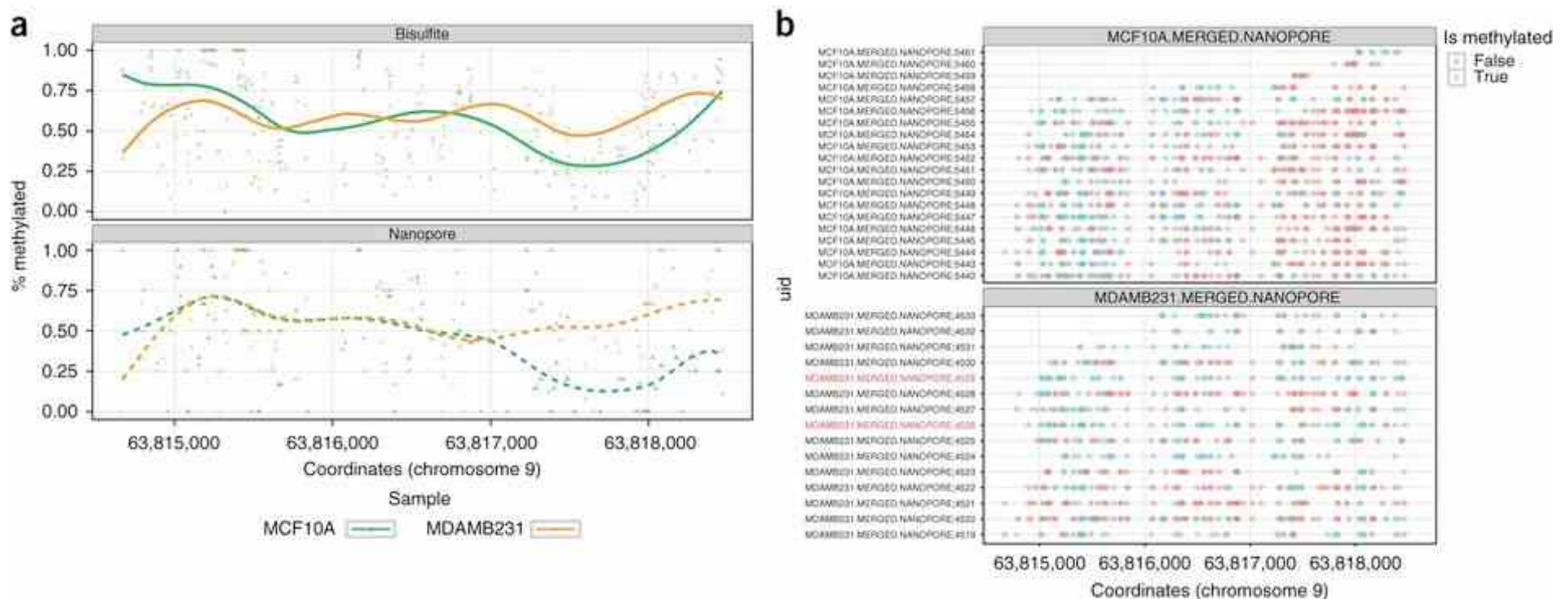
Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.

Bisulfite Conversion



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

Methylation changes in cancer detected by Nanopore Sequencing

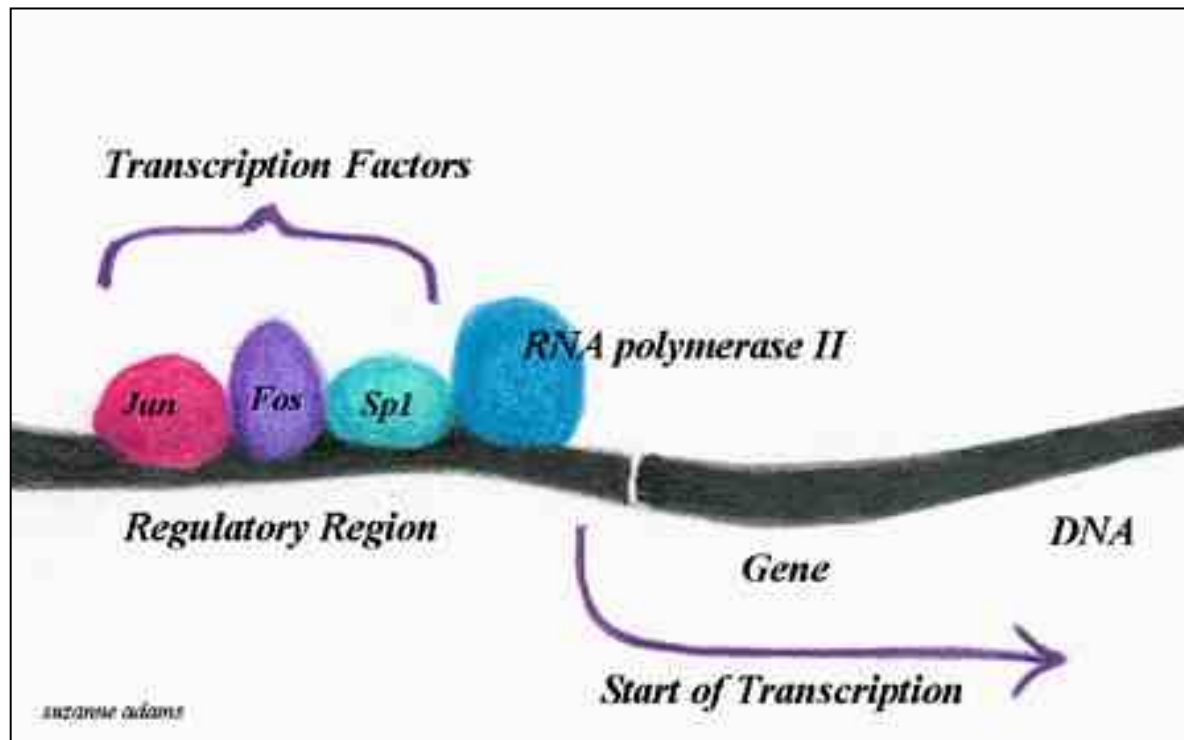


Comparison of bisulfite sequencing and nanopore-based R7.3 data in reduced representation data sets from cancer and normal cells. (a) Raw data (points) and smoothed data (lines) for methylation, as determined by bisulfite sequencing (top) and nanopore-based sequencing using an R7.3 pore (bottom), in a genomic region from the human mammary epithelial cell line MCF10A (green) and metastatic mammary epithelial cell line MDA-MB-231 (orange). (b) Same region as in a but with individual nanopore reads plotted separately. Each CpG that can be called is a point. Blue indicates methylated; red indicates unmethylated.

Detecting DNA cytosine methylation using nanopore sequencing

Simpson, Workman, Zuzarte, David, Dursi, Timp (2017) Nature Methods. doi:10.1038/nmeth.4184

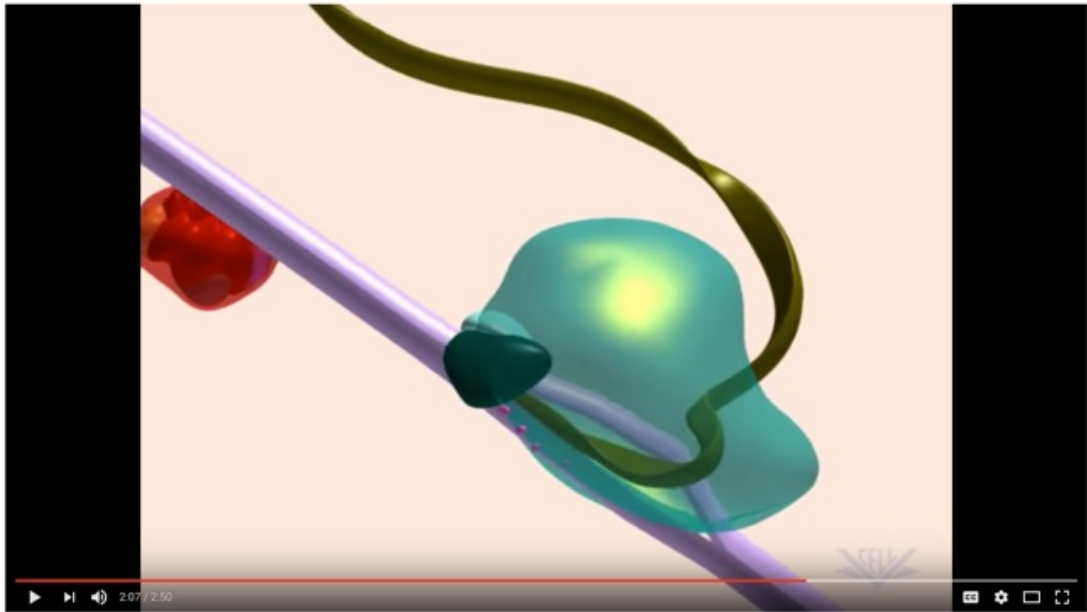
ChIP-seq



Genome-wide mapping of in vivo protein-DNA interactions.

Johnson et al (2007) *Science*. 316(5830):1497-502

Transcription



Transcription

2,018,430 views

ndsuvirtualcell
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

Transcription and Translation: From DNA to Protein
Professor Dave Explains
151K views
6:27

DNA - transcription and translation
Wisam Kabaha
40K views
7:18

Transcription and mRNA processing | Biomolecules | Khan Academy
105K views
10:25

DNA transcription and translation Animation
Haider abd
45K views
7:18

Translation
ndsuvirtualcell
2.1M views
3:33

Transcription and Translation Overview
Armando Hasudungan
611K views
13:18

DNA, Hot Pockets, & The Longest Word Ever: Crash Course
CrashCourse
2.2M views
14:08

Transcription 1
khanacademymedicine
263K views
12:06

TRANSCRIPTION
congthanh
795K views
1:28

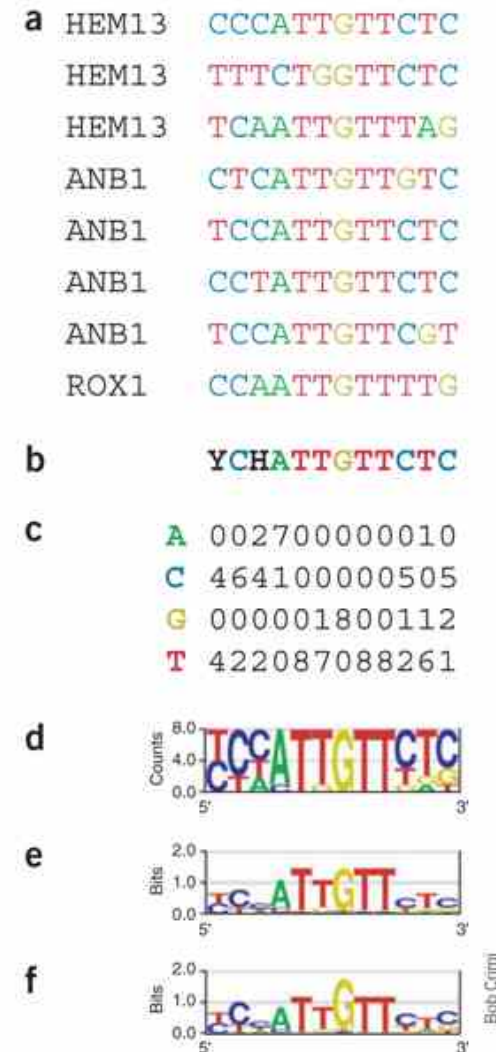
Moana - Best Scenes (FHD)

<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*.



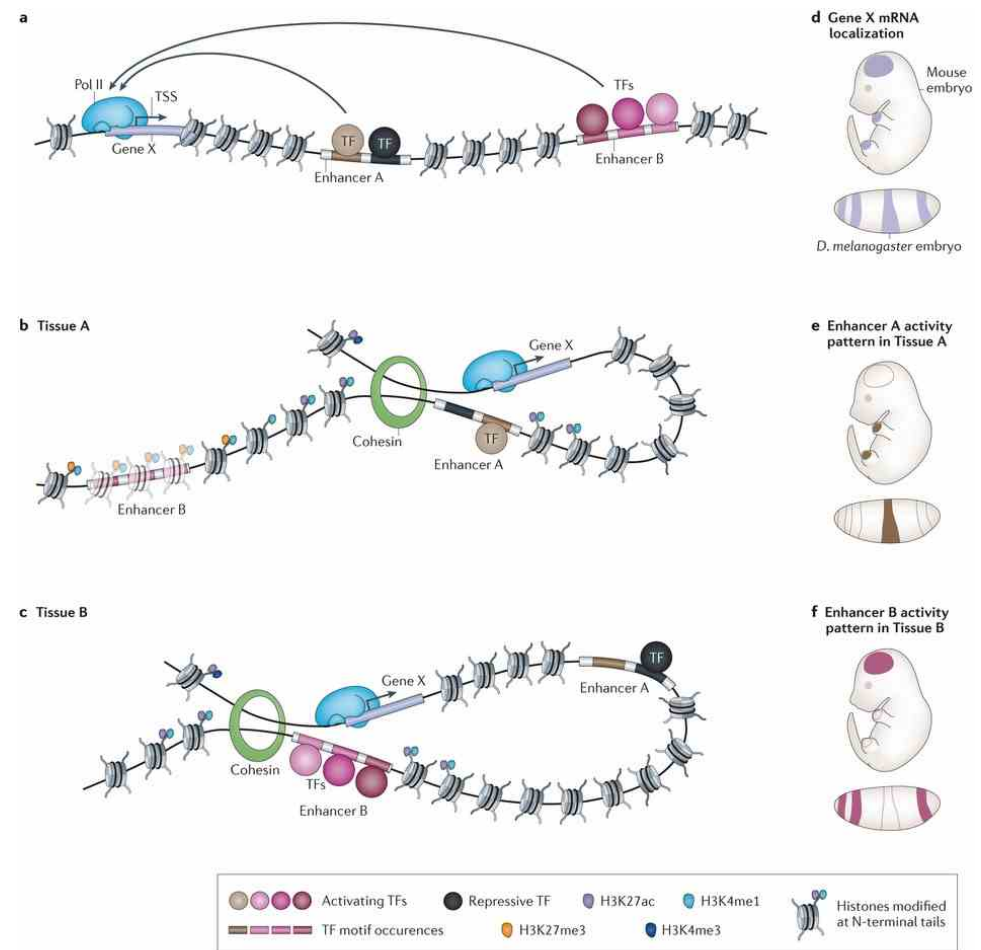
What are DNA sequence motifs?

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

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

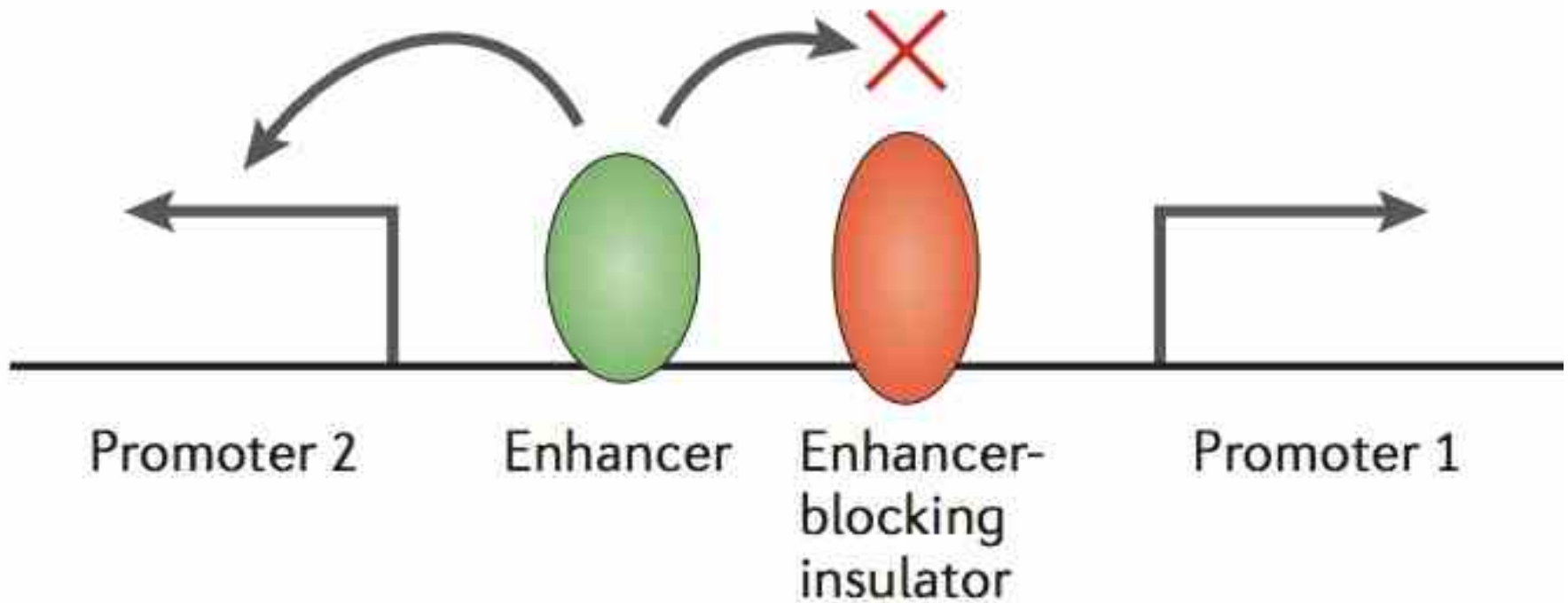


Nature Reviews | Genetics

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.

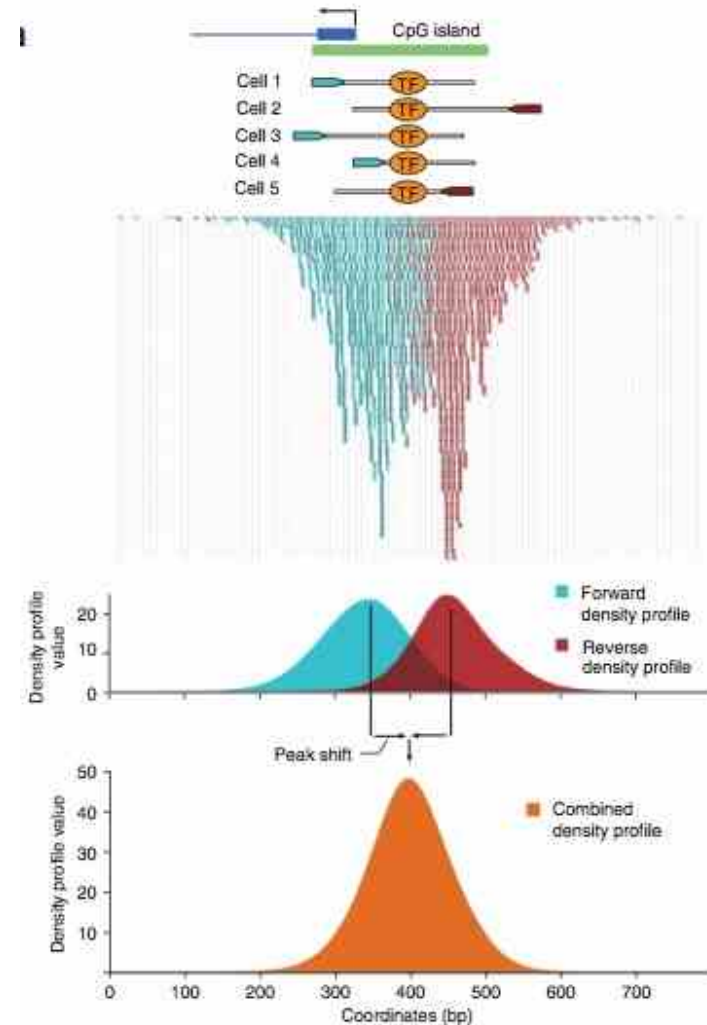
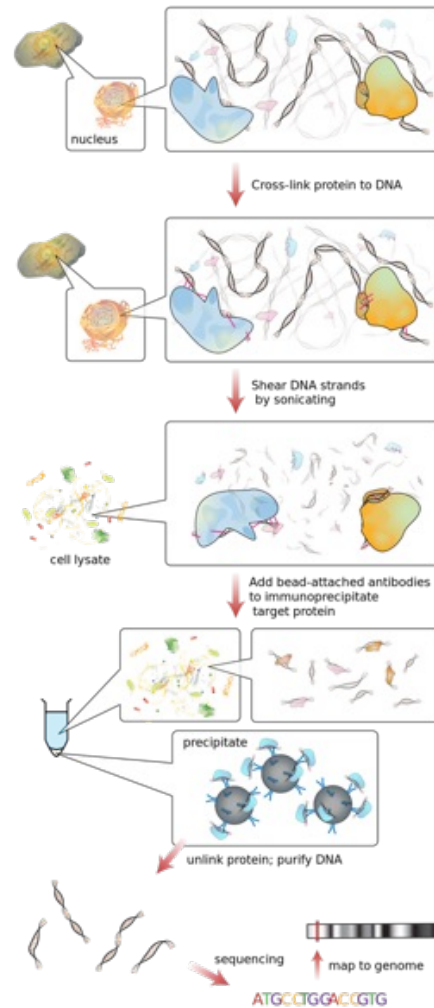
Insulators: exploiting transcriptional and epigenetic mechanisms

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

ChIP-seq:TF Binding

Goals:

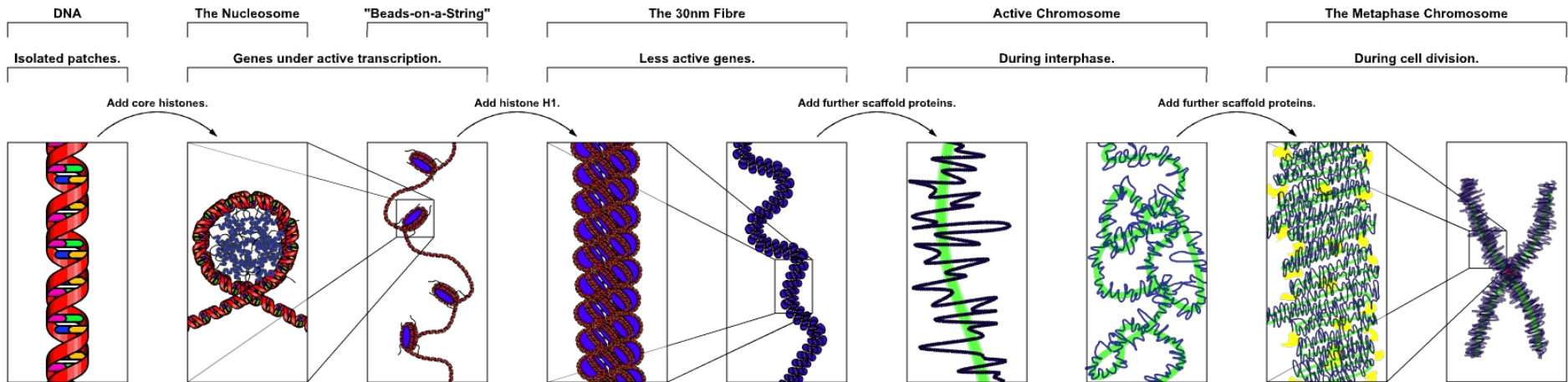
- Where are transcription factors and other proteins binding to the DNA?
- How strongly are they binding?
- Do the protein binding patterns change over developmental stages or when the cells are stressed?



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



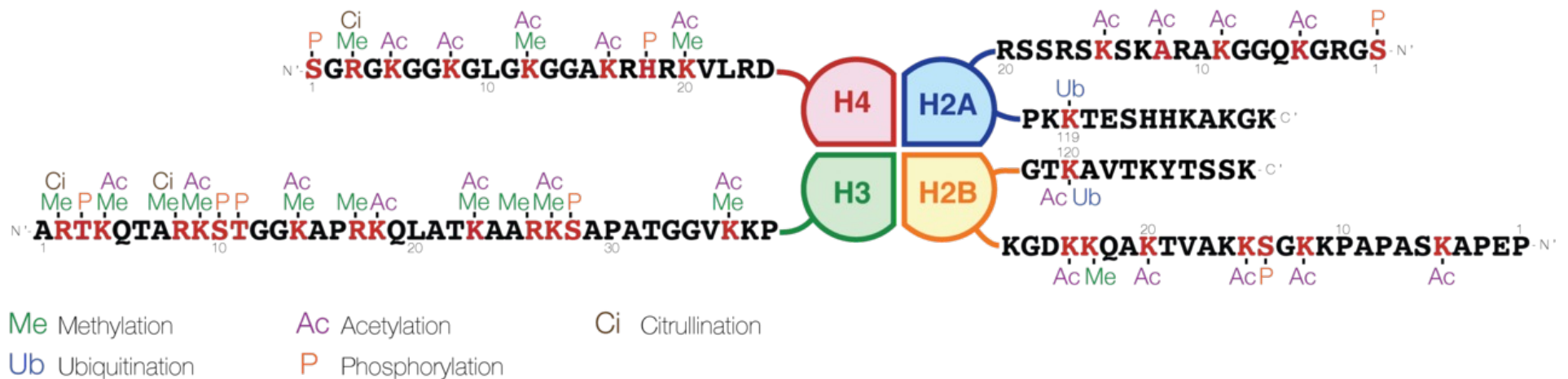
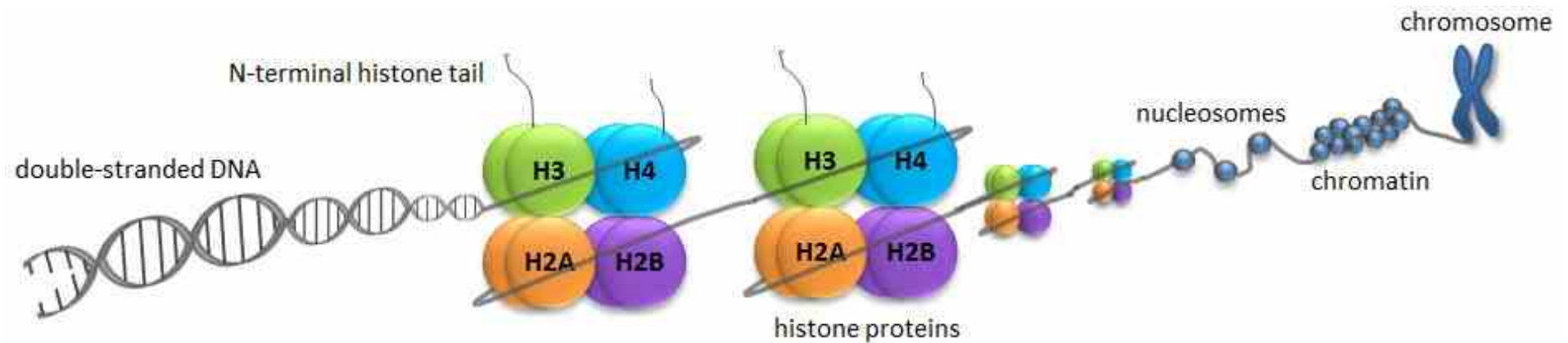
Nucleosome is a basic unit of DNA packaging in eukaryotes

- 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

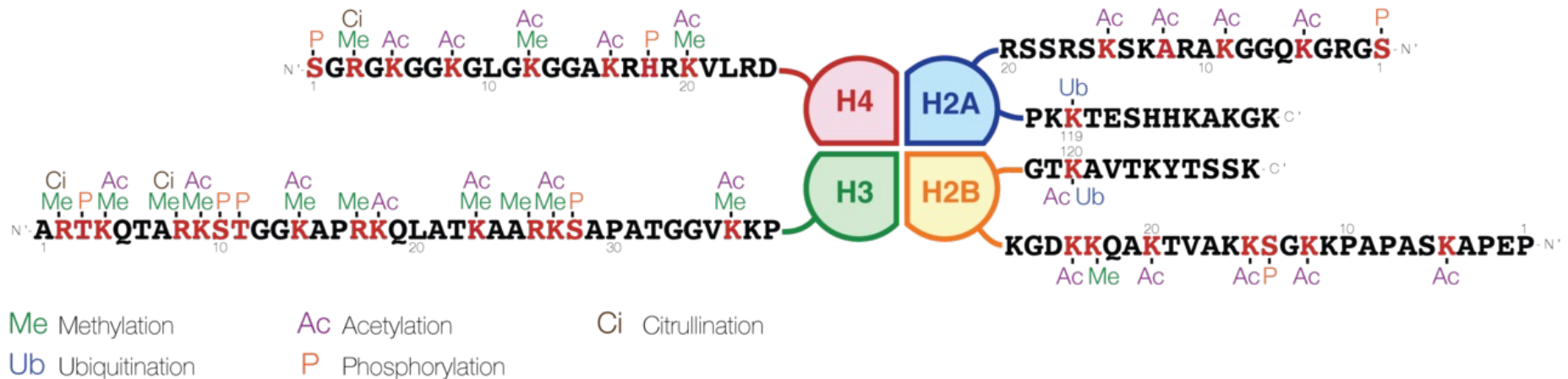
Nucleosomes form the fundamental repeating units of eukaryotic chromatin

- 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



ChIP-seq: Histone Modifications

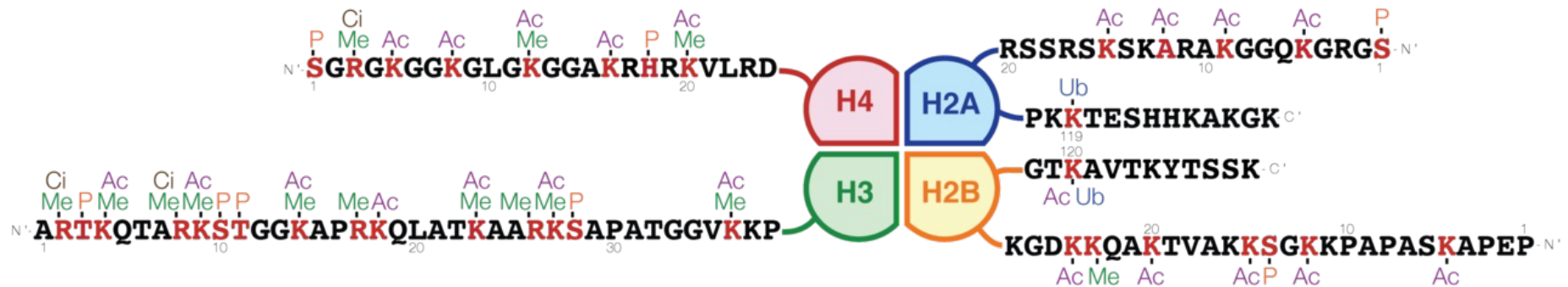


The common nomenclature of histone modifications is:

- 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.

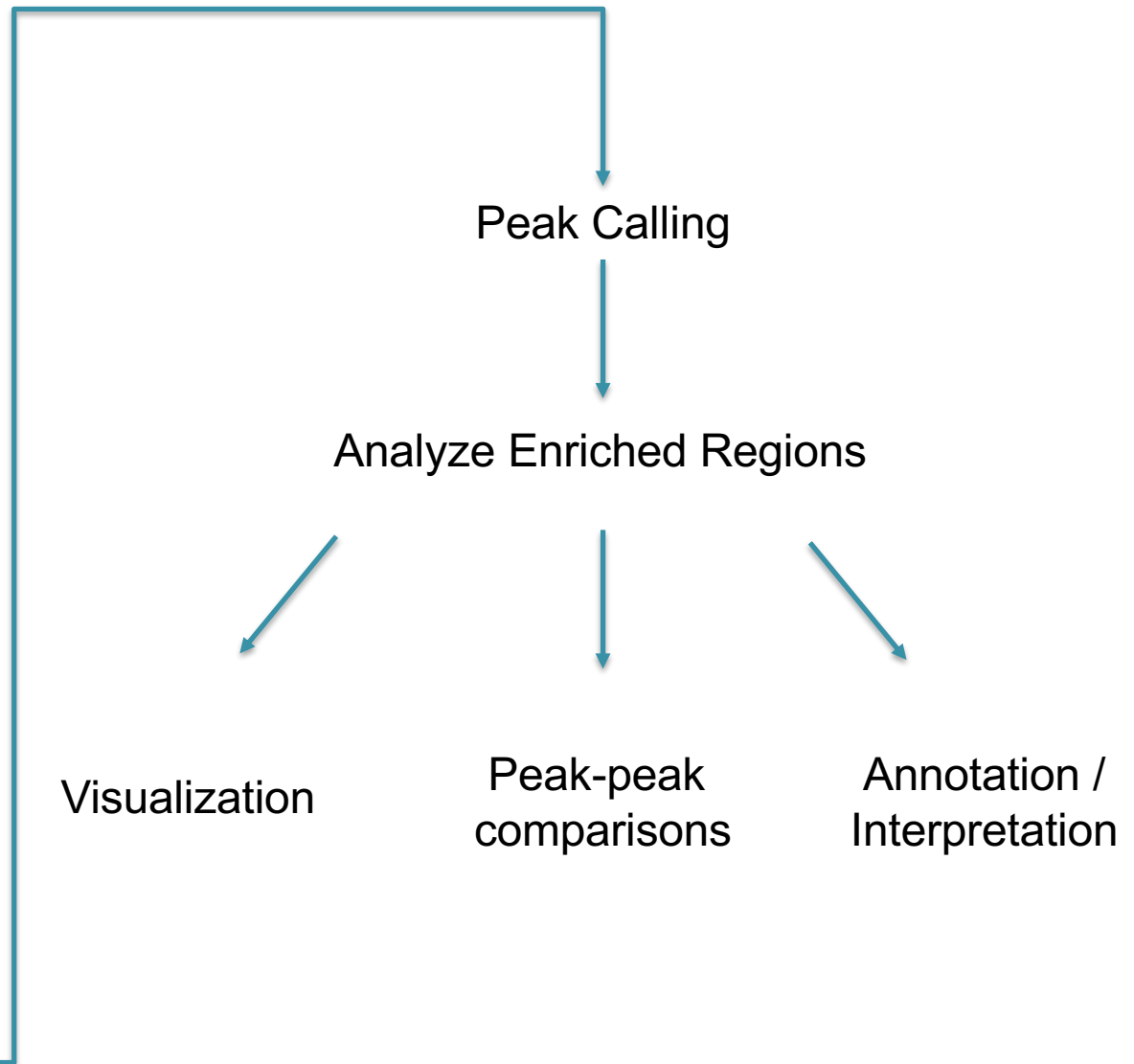
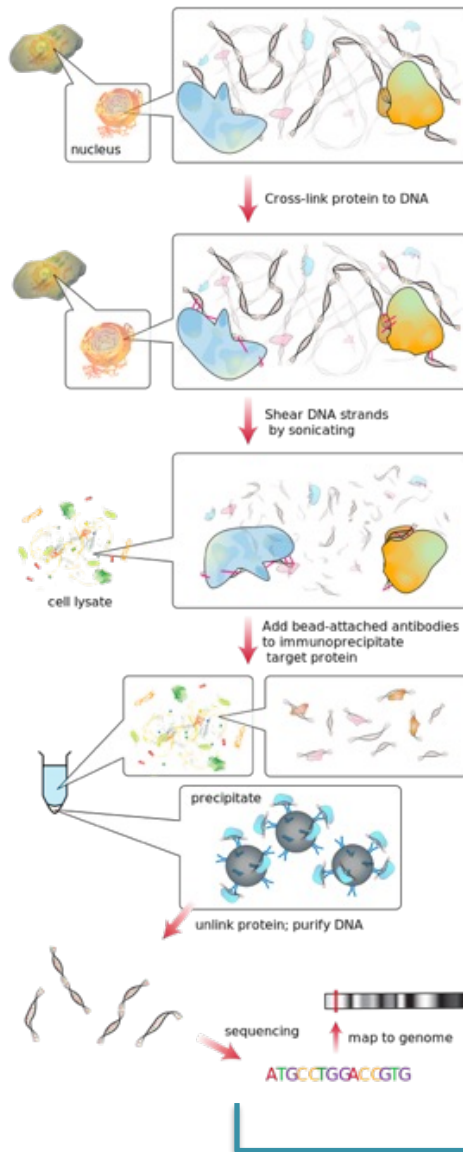
ChIP-seq: Histone Modifications



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]		

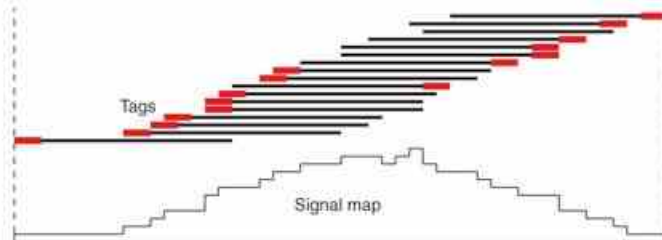
- H3K4me3 is enriched in transcriptionally active promoters.^[12]
- H3K9me3 is found in constitutively repressed genes.
- H3K27me is found in facultatively repressed genes.^[7]
- H3K36me3 is found in actively transcribed gene bodies.
- H3K9ac is found in actively transcribed promoters.
- H3K14ac is found 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.

General Flow of ChIP-seq Analysis



PeakSeq

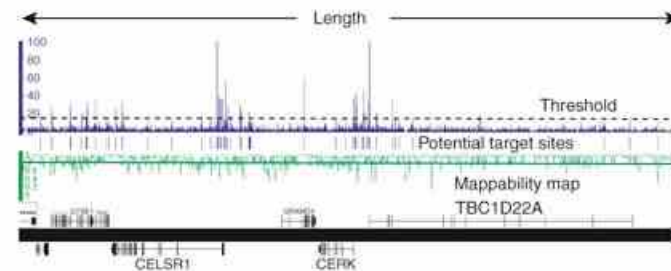
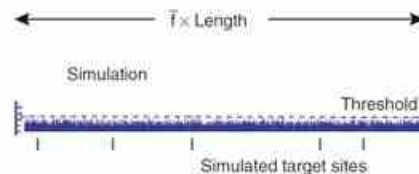
1. Constructing signal maps



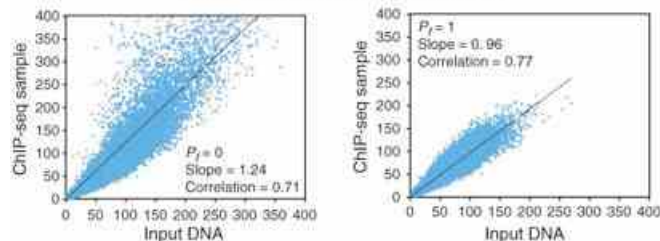
- Extend mapped tags to DNA fragment
- Map of number of DNA fragments at each nucleotide position

2. First pass: determining potential binding regions by comparison to simulation

- Simulate each segment
- Determine a threshold satisfying the desired initial false discovery rate
- Use the threshold to identify potential target sites



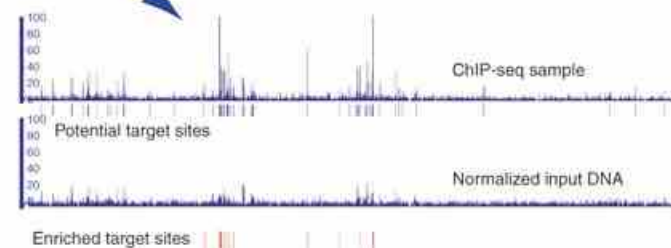
3. Normalizing control to ChIP-seq sample



- Select fraction of potential peaks to exclude (parameter P_l)
- Count tags in bins along chromosome for ChIP-seq sample and control
- Determine slope of least squares linear regression

4. Second pass: scoring enriched target regions relative to control

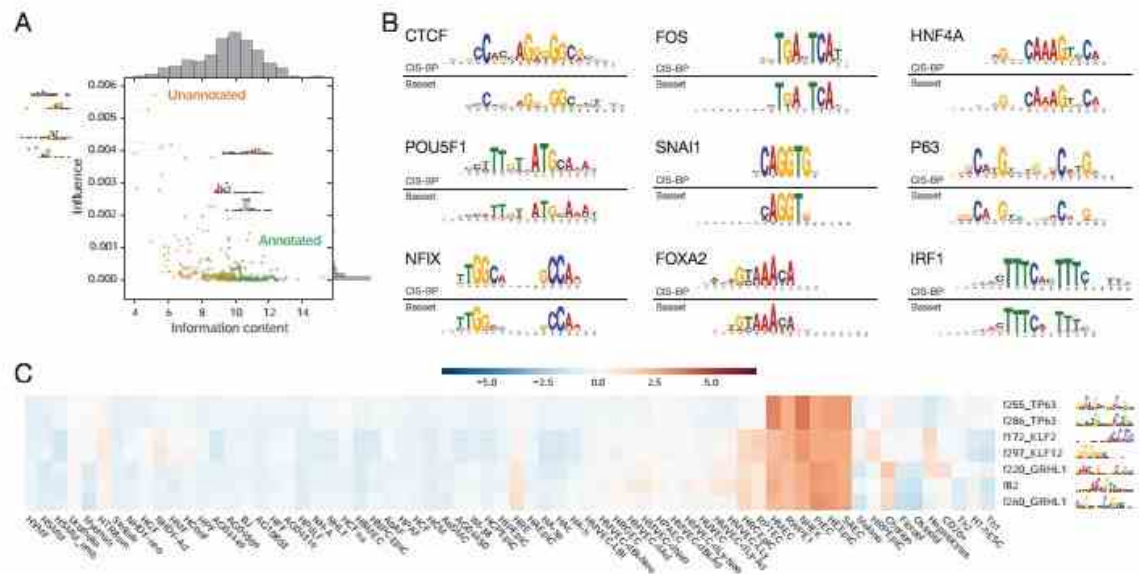
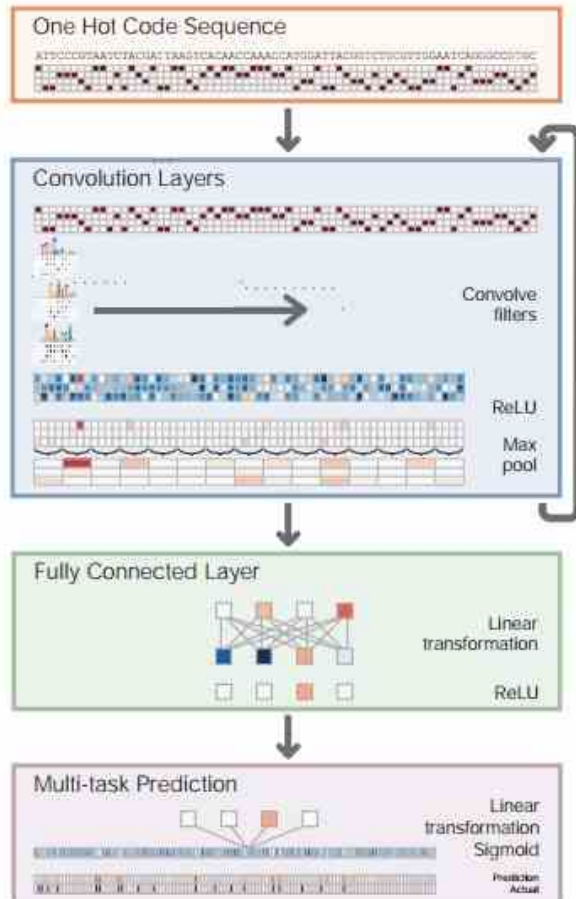
- For potential binding sites calculate the fold enrichment
- Compute a P -value from the binomial distribution
- Correct for multiple hypothesis testing and determine enriched target sites



PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls

Rozowsky et al (2009) Nature Biotechnology 27, 66 - 75

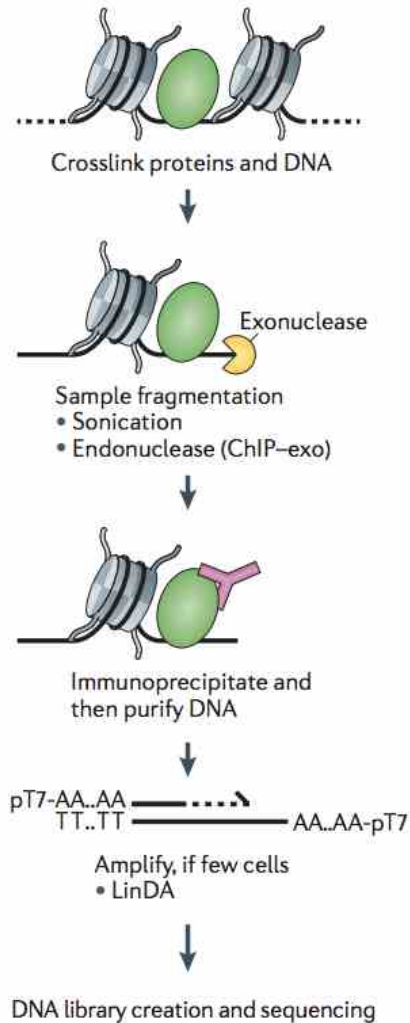
Basset



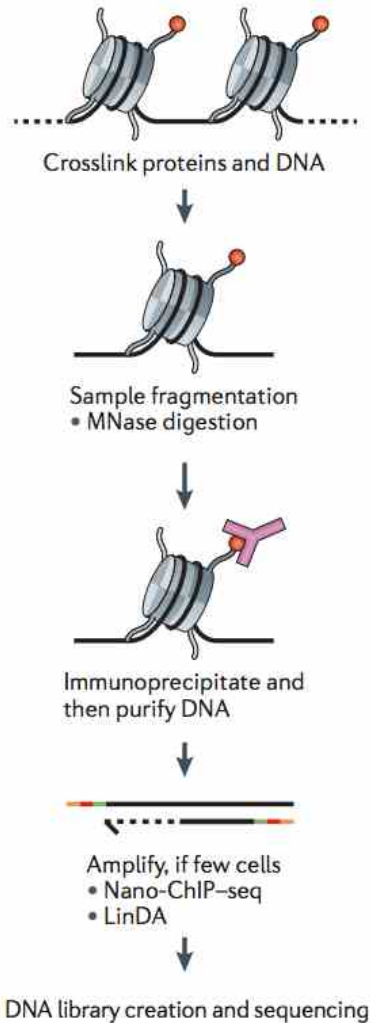
Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks
 Kelley et al. (2016) Genome Research doi: 10.1101/gr.200535.115

Related Assays

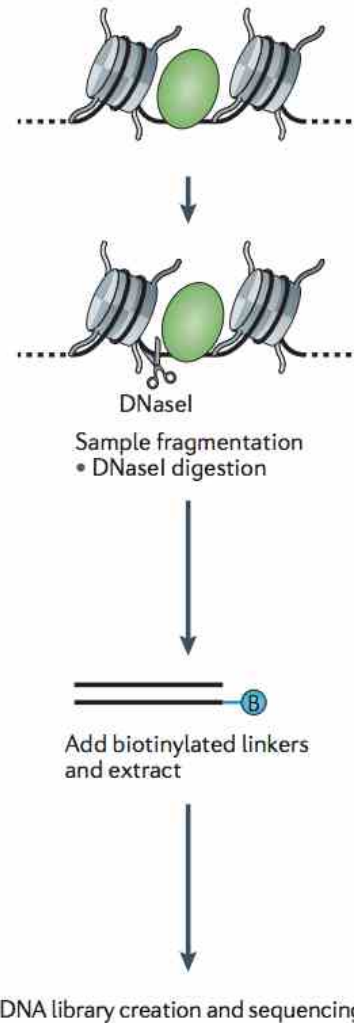
a DNA-binding protein ChIP-seq



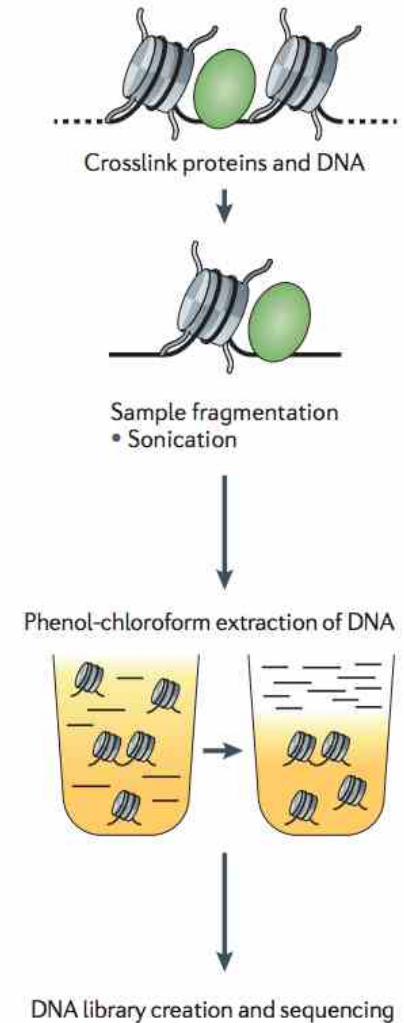
b Histone modification ChIP-seq



c DNase-seq



d FAIRE-seq



ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions

Furey (2012) *Nature Reviews Genetics*. 13, 840-852