

Lecture 16. scRNAseq and Gene Regulation

Michael Schatz

March 17, 2021

Applied Comparative Genomics



Project Proposal: Due March 15

A screenshot of a web browser window showing a GitHub project proposal page. The title bar says "github.com". The page content includes:

Project Proposal

Assignment Date: Monday March 8, 2021
Due Date: Monday, March 15 2021 @ 11:59pm

Review the [Project Ideas](#) page

Work solo or form a team for your class project of no more than 3 people.

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 2 to 3 relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)
- Please add a note if you need me to sponsor you for a MARCC account (high RAM, GPUs, many cores, etc)

Submit the proposal as a 1 to 2 page PDF on GradeScope (each team member should submit the same PDF). After submitting your proposal, we can 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 the project.

Later, you will present your project in class during the last week of class. You will also submit a written report (5-7 pages) of your project, formatting as a Bioinformatics article (Intro, Methods, Results, Discussion, References). Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online

Please use Piazza to coordinate proposal plans!

Assignment 5: Due Wed Mar 17

The screenshot shows a GitHub repository page for "Assignment 5: BWT and RNA-seq". The page has a light gray header with standard browser controls. Below the header, the repository details are shown: "148 Lines (89 sloc) 8.82 KB". On the right side of the header, there are buttons for "Raw", "Blame", "Copy", and "Edit". The main content area has a title "Assignment 5: BWT and RNA-seq" and a subtitle "Assignment Date: Wednesday, Mar. 3, 2021" and "Due Date: Wednesday, Mar. 17, 2021 @ 11:59pm". A section titled "Assignment Overview" contains text about the assignment requirements and a note to show work/code in the writeup. It also mentions Piazza for questions. A question section for BWT Encoding is described with pseudo code.

Assignment 5: BWT and RNA-seq

Assignment Date: Wednesday, Mar. 3, 2021
Due Date: Wednesday, Mar. 17, 2021 @ 11:59pm

Assignment Overview

In this assignment you will write a simple BWT encoder and decoder, and explore a couple of aspects of RNA-seq (with a small introduction to clustering). For this assignment, you will have to generate some visualizations - we recommend R or Python, but use a language you are comfortable with!

Make sure to show your work/code in your writeup!

As a reminder, any questions about the assignment should be posted to [Piazza](#).

Question 1. BWT Encoding [10 pts]

In the language of your choice, implement a BWT encoder and encode the string below. Linear time methods exist for computing the BWT, although for this assignment you can use the simple method based on standard sorting techniques. Your solution does not need to be an optimal algorithm and can use $O(n^2)$ space and $O(n^2 \lg n)$ time.

Here is the recommended pseudo code (make sure to submit your code as well as the encoded string):

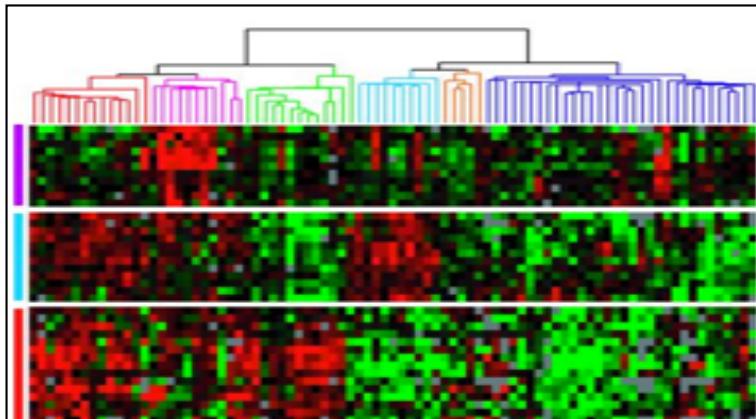
```
computeBwt(string s)
    ## add the magic end-of-string character
    s = s + "$"

    ## build up the BWT from the cyclic permutations
    ## note the i-th cyclic permutation is just "s[i..n] + s[0..i]"
    StringList rows = []
    for (i = 0; i < length(s); i++)
        rows.append(cyclic_permutation(s, i))

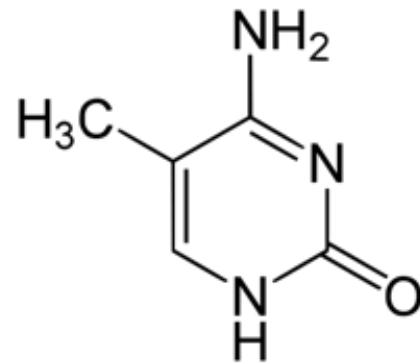
    ## last use the builtin sort command
```

*-seq in 4 short vignettes

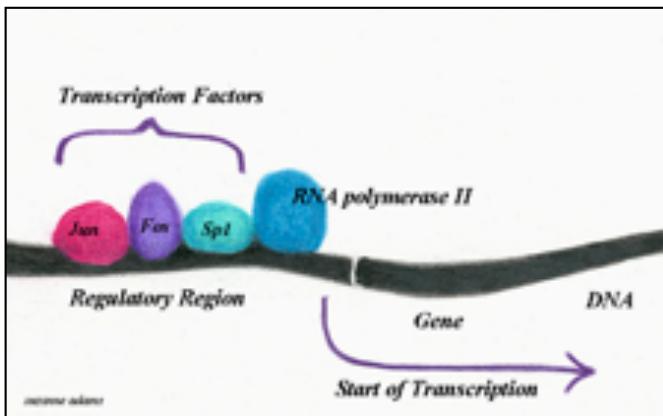
RNA-seq



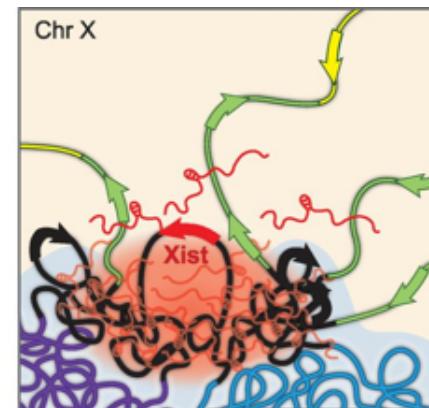
Methyl-seq



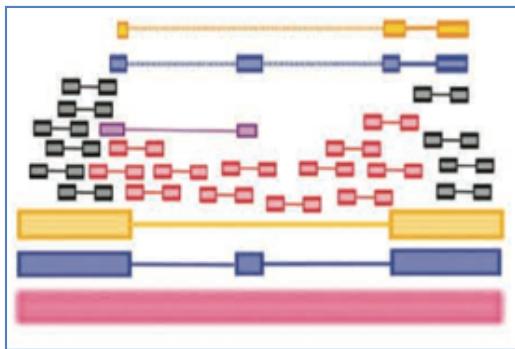
ChIP-seq



Hi-C



RNA-seq Challenges

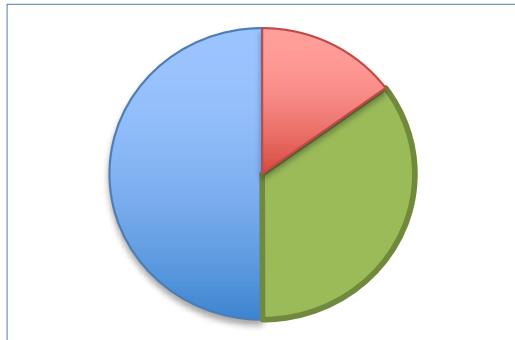


Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

TopHat: discovering spliced junctions with RNA-Seq.

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

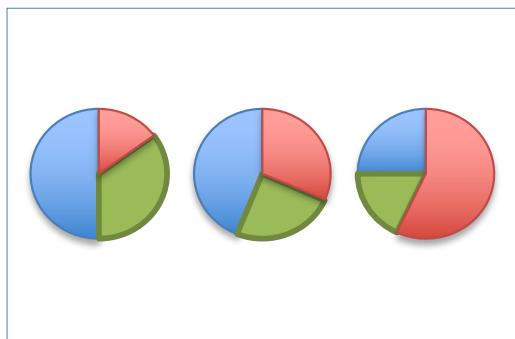


Challenge 2: Read Count != Transcript abundance

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

Transcript assembly and quantification by RNA-seq

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



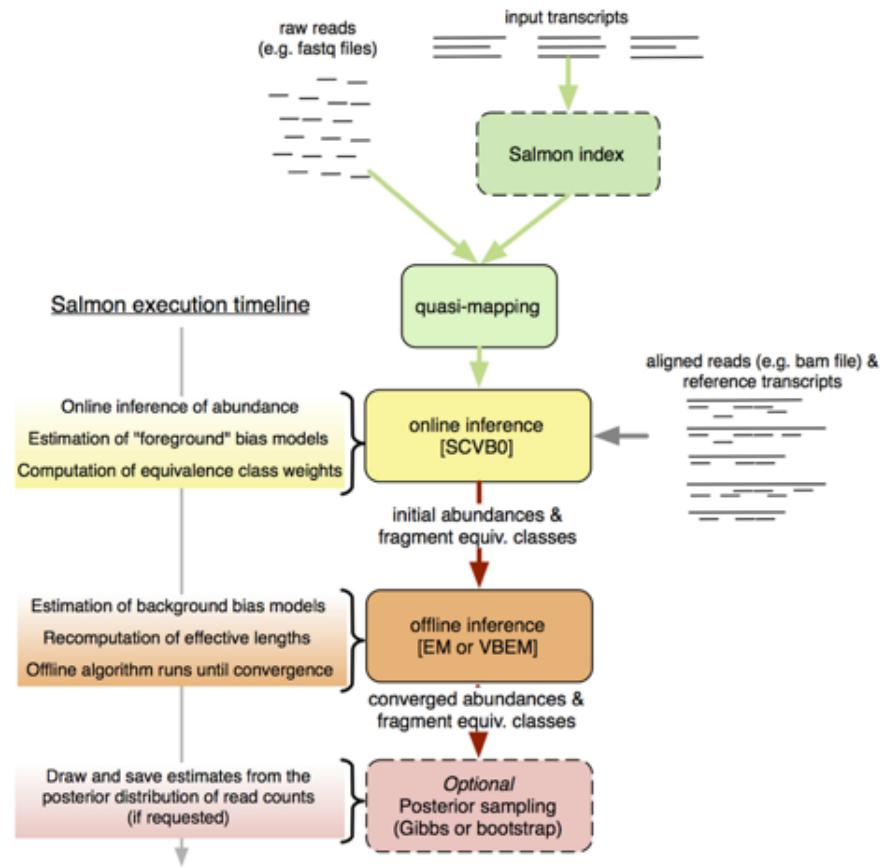
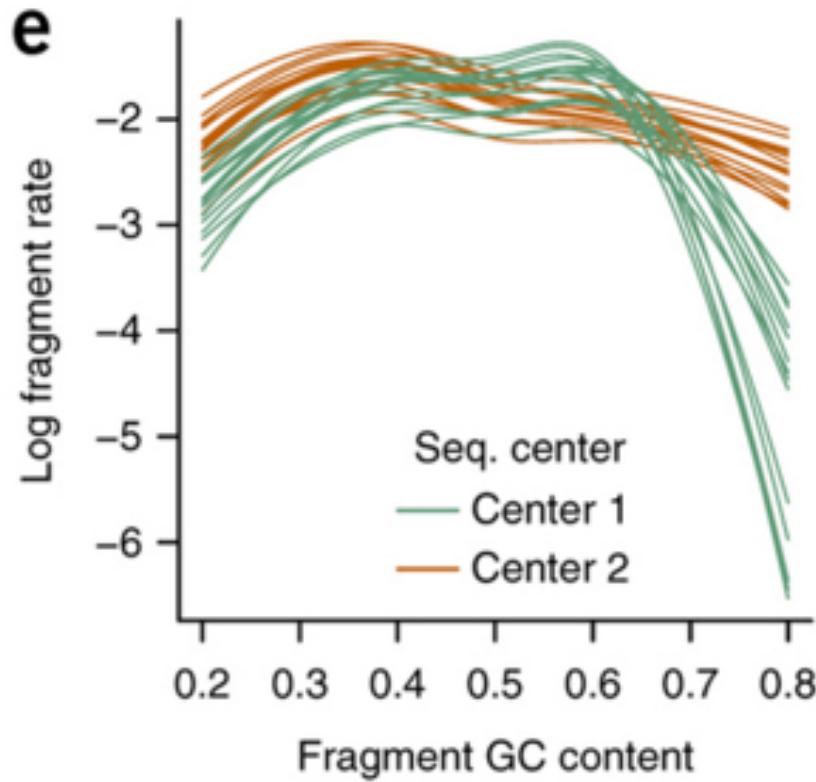
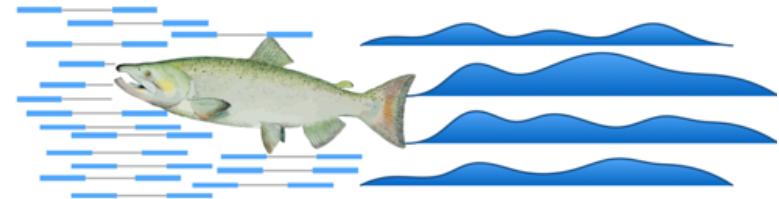
Challenge 3: Transcript abundances are stochastic

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

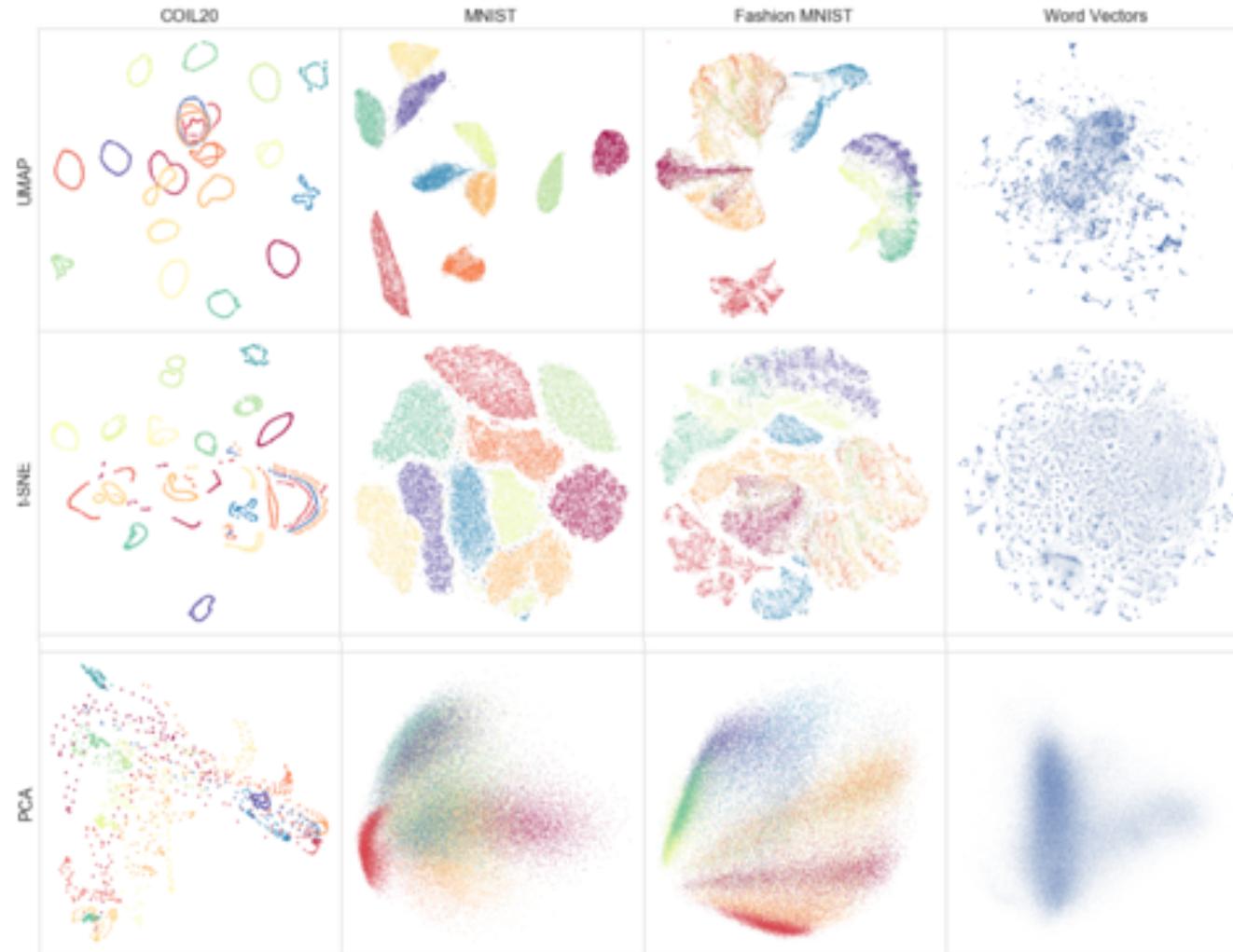
Salmon: The ultimate RNA-seq Pipeline?



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

UMAP



UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

McInnes et al (2018) arXiv. 1802.03426

<https://www.youtube.com/watch?v=nq6iPZVUxZU>

<https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668>



Single Cell Analysis

I. Why single cells?

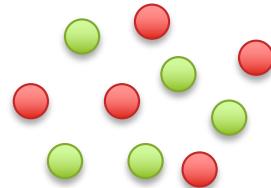
Bonus slides:

1. scDNA
2. scRNA and other assays

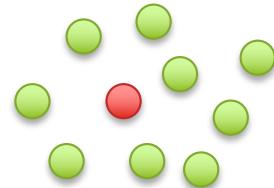
Population Heterogeneity

Red cells express twice the abundance of “brain” genes compared to green cells

Experiment 1: 50/50



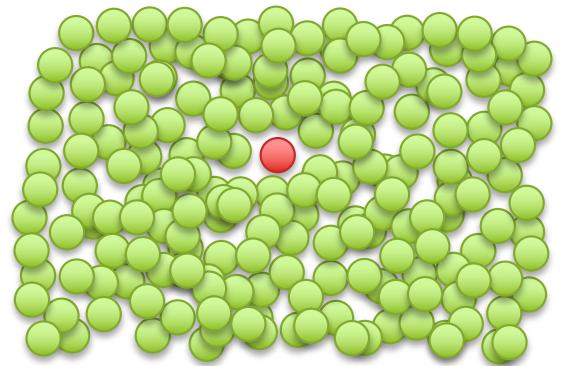
Experiment 2: 1/10



Compared to a control sample of pure green cells, this sample will show:

$$\begin{aligned} & 50\% 2x + 50\% 1x \\ & = 1.5x \text{ over expression of brain genes} \end{aligned}$$

Experiment 3: 1/1000



Compared to a control sample of pure green cells, this sample will show:

$$\begin{aligned} & 10\% 2x + 90\% 1x \\ & = 1.1x \text{ over expression of brain genes} \end{aligned}$$

Compared to a control sample of pure green cells, this sample will show:

$$\begin{aligned} & 0.1\% 2x + 99.1\% 1x \\ & = 1.001x \text{ over expression of brain genes} \end{aligned}$$

The limitations of averages

	Drug A	Drug B
Overall Response	78% (273/350)	83% (289/350)

The limitations of averages

	Drug A	Drug B
Overall Response	78% (273/350)	83% (289/350)
Male Response	93% (81/87)	87% (234/270)
Female Response	73% (192/263)	69% (55/80)

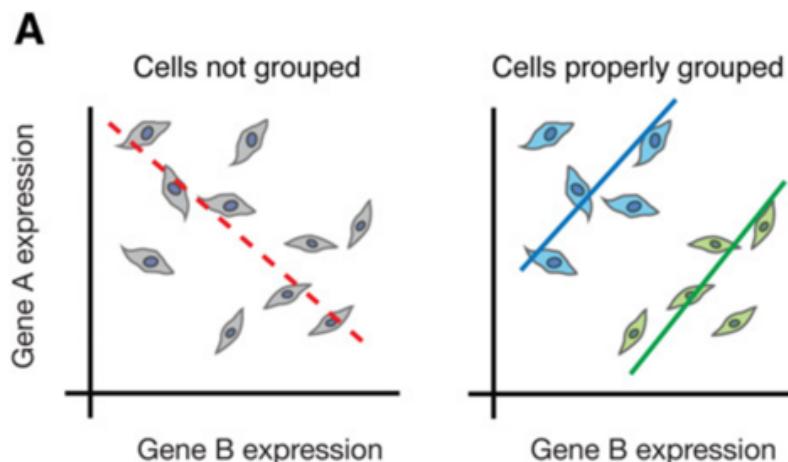
What??? How can the better performing drug depend on if you examine the overall response or separately examine by gender

Example of Simpson's paradox:

Trend of the overall average may reverse the trends of each constituent group

In this example, the "lurking" variable (or confounding variable) is the severity of the case (represented by the doctors' treatment decision trend of favoring B for less severe cases), which was not previously known to be important until its effects were included. (Based on real analysis of kidney stone treatments)

The paradox of averages



What??? How can the better performing drug depend on if you examine the overall response or separately examine by gender

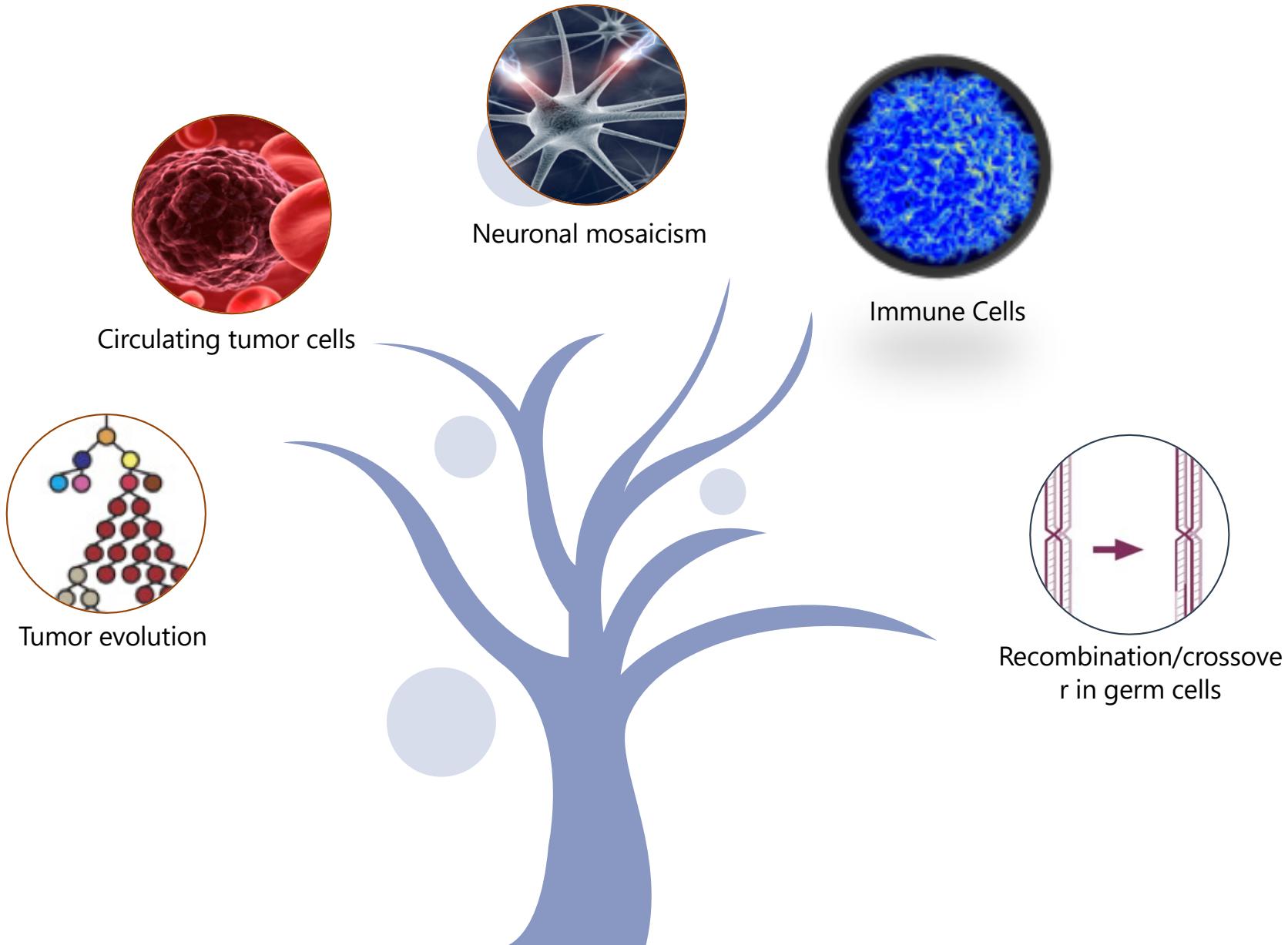
Example of Simpson's paradox:

Trend of the overall average may reverse the trends of each constituent group

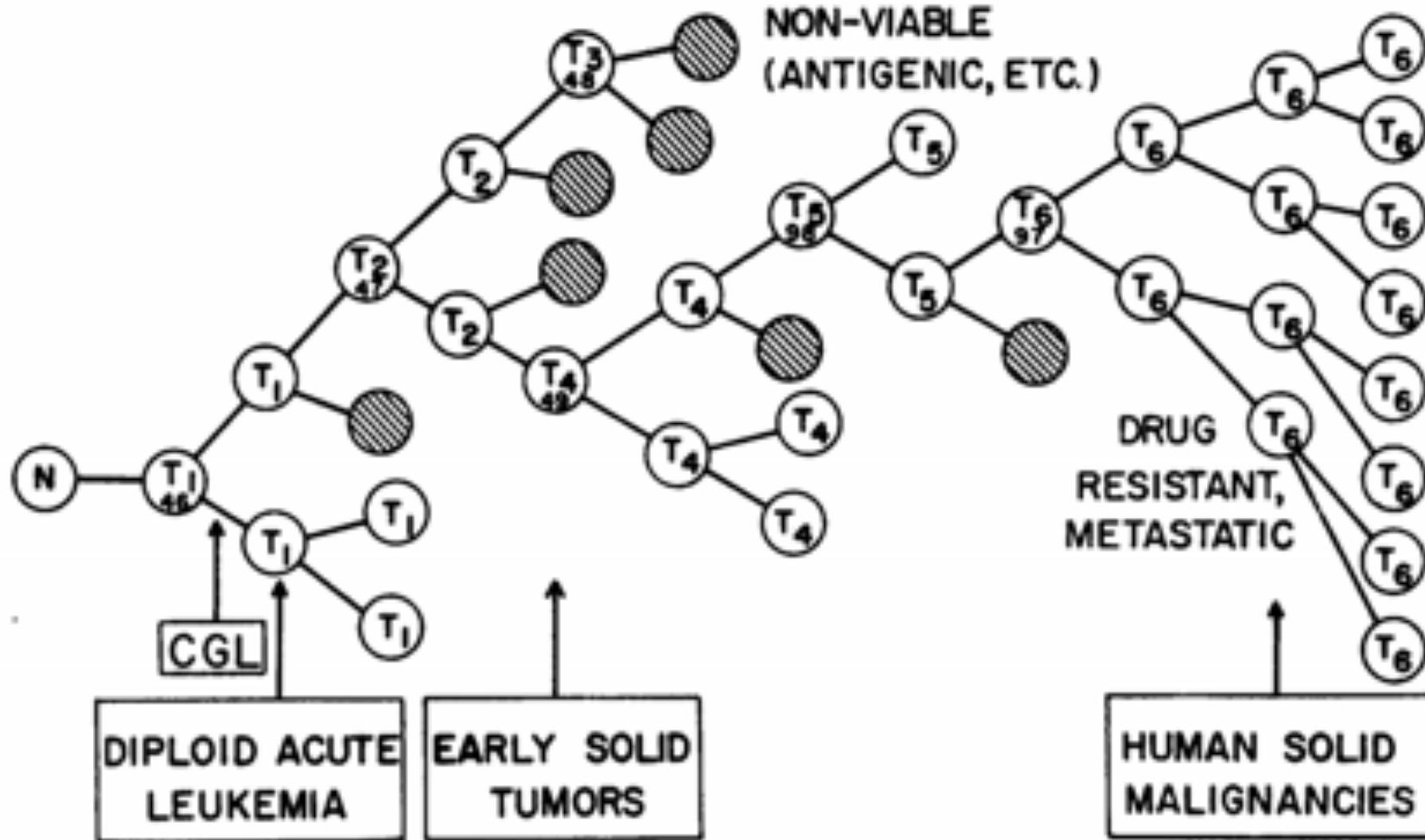
In this example, the "lurking" variable (or confounding variable) is the severity of the case (represented by the doctors' treatment decision trend of favoring B for less severe cases), which was not previously known to be important until its effects were included. (Based on real analysis of kidney stone treatments)

(Trapnell, 2015, Genome Research)

Sources of (Genomic) Heterogeneity

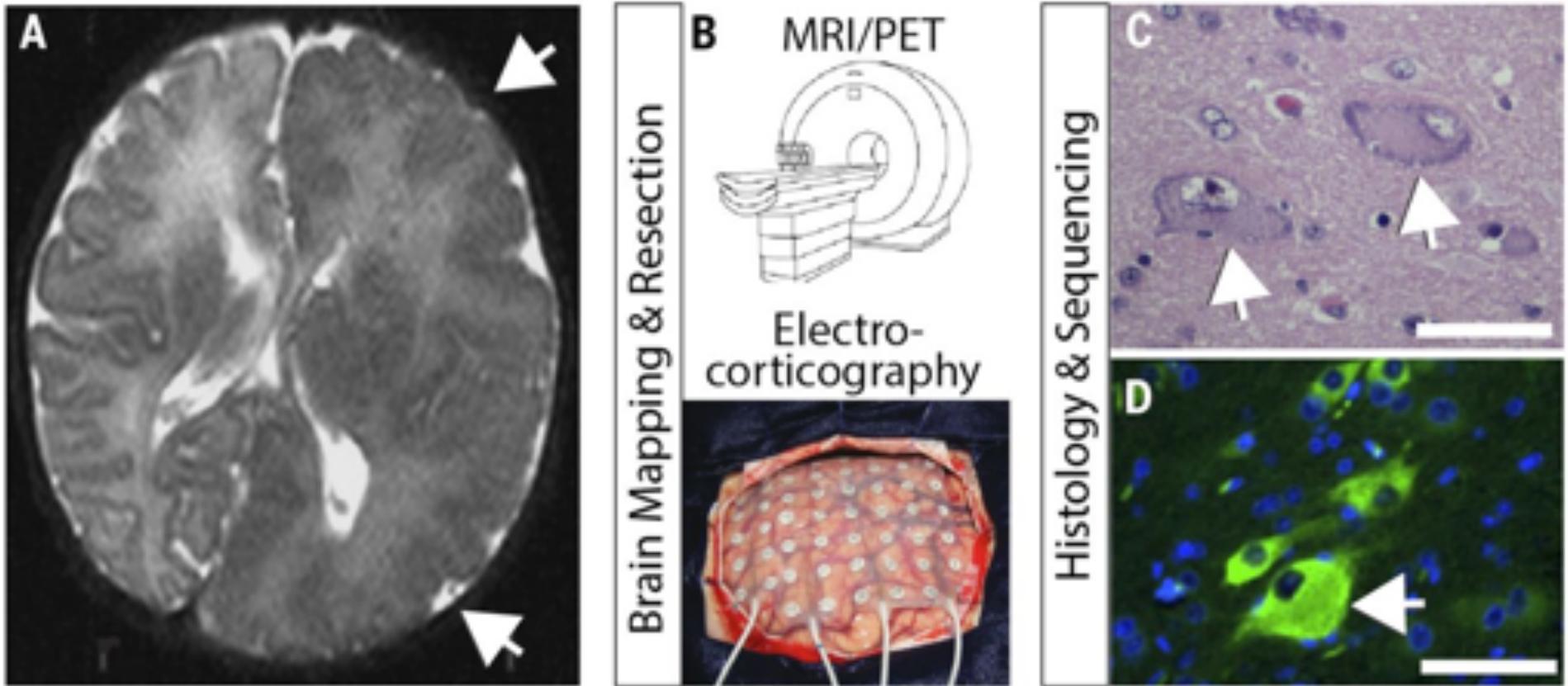


Tumor Evolution



The Clonal Evolution of Tumor Cell Populations

Peter C. Nowell (1976) Science. 194(4260):23-28 DOI: 10.1126/science.959840



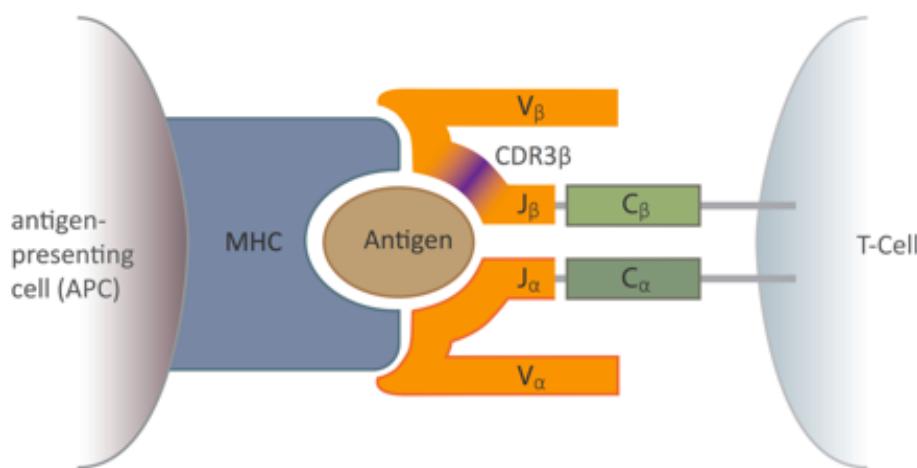
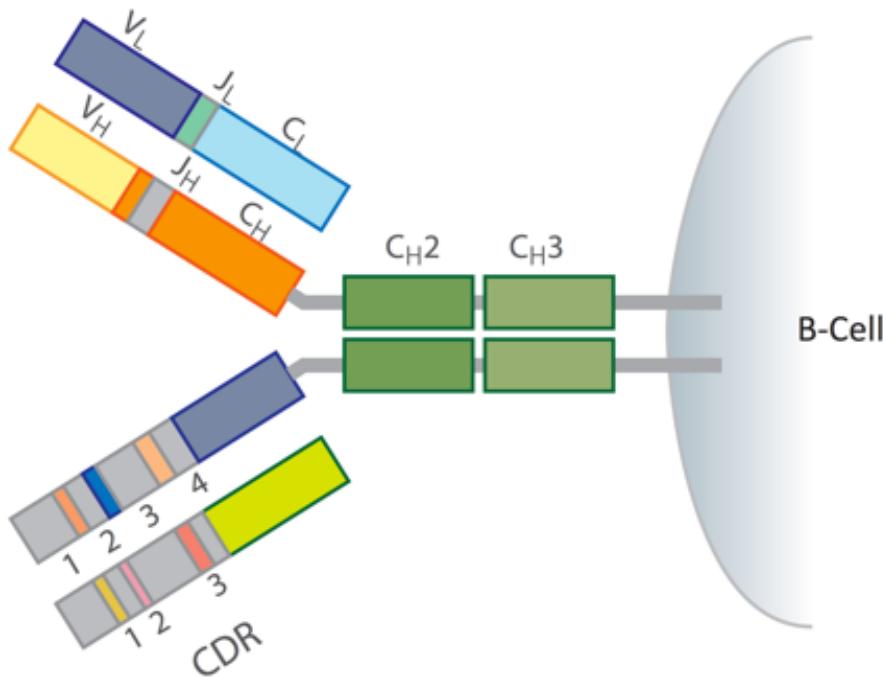
An example of brain somatic mosaicism that leads to a focal overgrowth condition.

(A) Axial brain MRI of focal overgrowth from a 2-month-old child with intractable epilepsy and intellectual disability. **(B)** Brain mapping using high-resolution MRI is followed by surgical resection of diseased brain tissue. **(C)** Histological analysis with hematoxylin/eosin showing characteristic balloon cells consisting of large nuclei, distinct nucleoli, and glassy eosinophilic cytoplasm. **(D)** After surgery, the patient showed clinical improvement.

Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaic Network. McConnell et al (2017) Science. doi: 10.1126/science.aal1641

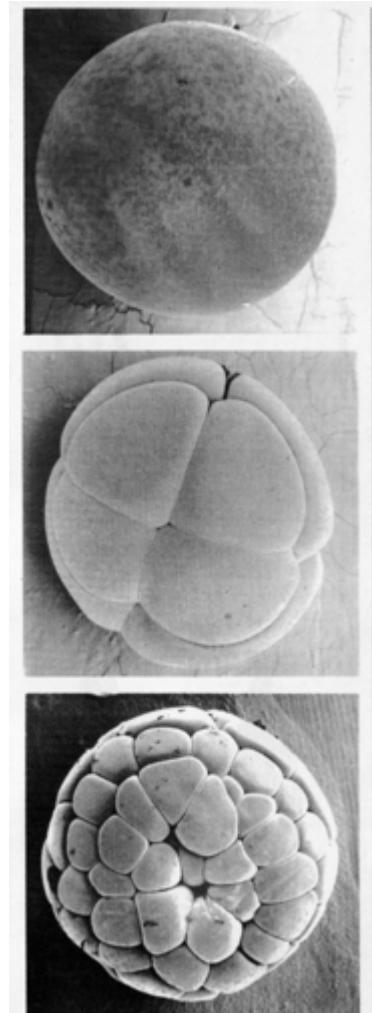
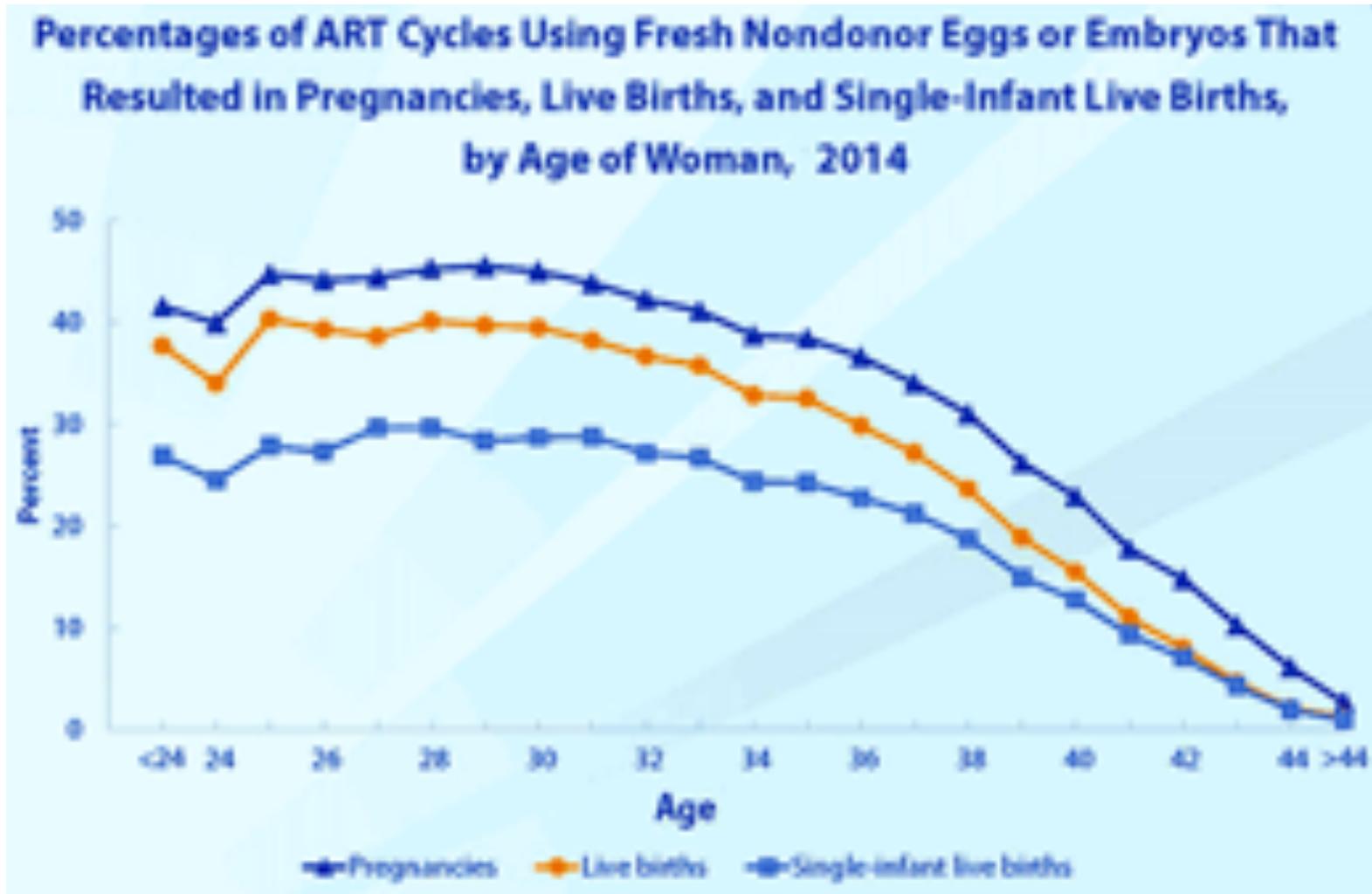
Immunology

- **Massive diversity rivaled only by germ cells**
- **Somatic recombination**

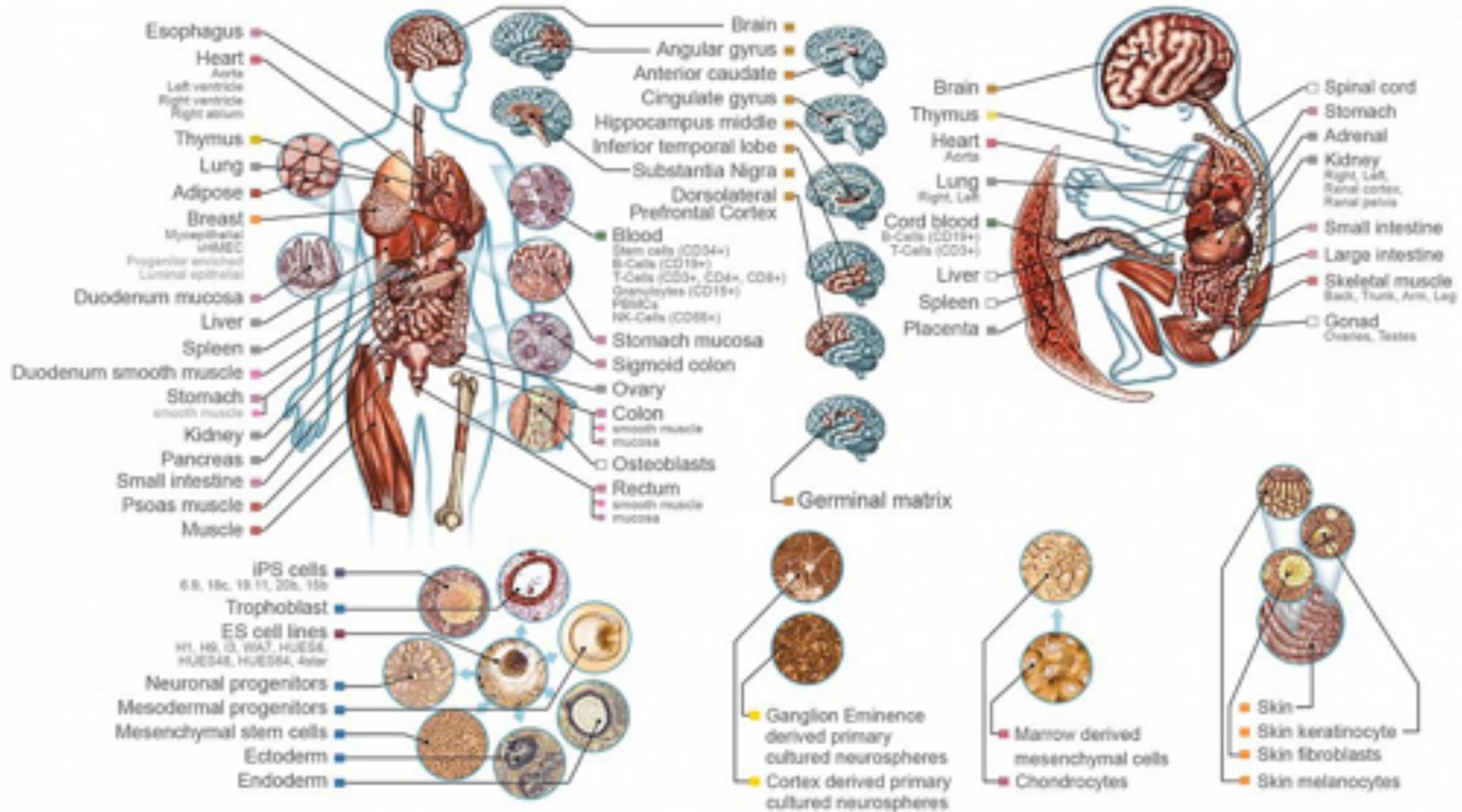


- **B cells – antibody generation**
- **T cells – antigen response**

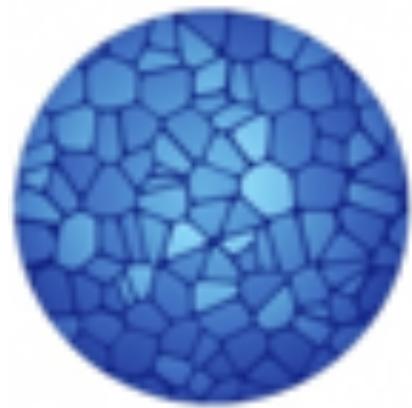
In-vitro Fertilization



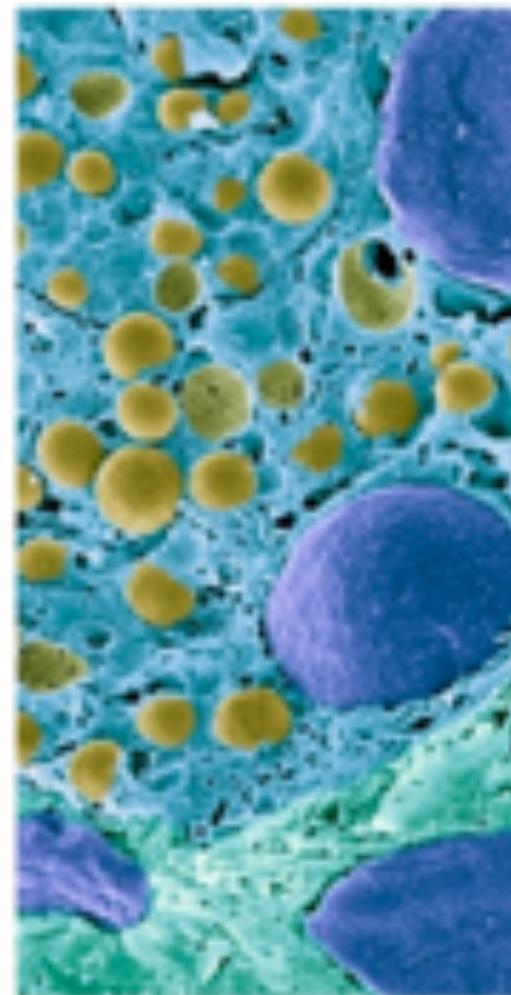
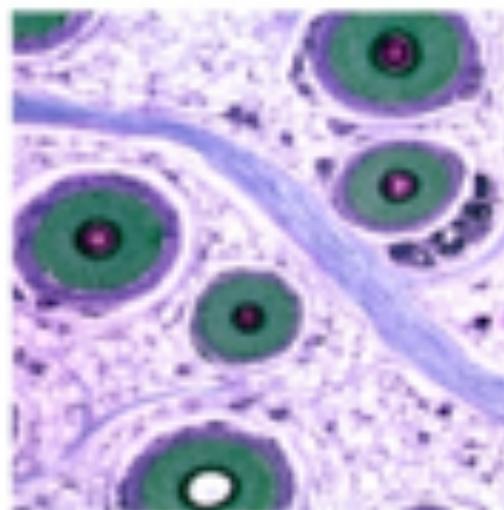
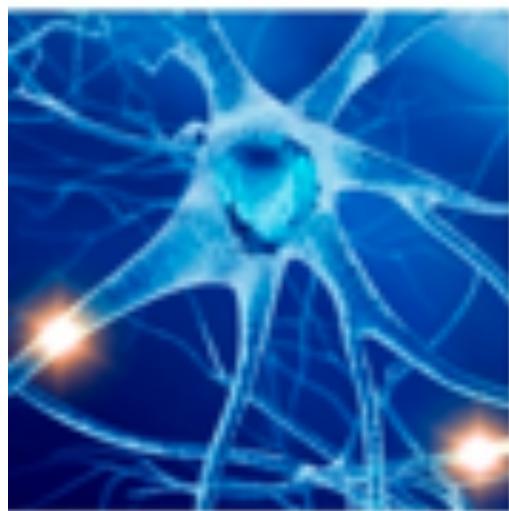
Sources of (Cellular) Heterogeneity



Roadmap Epigenomics Consortium

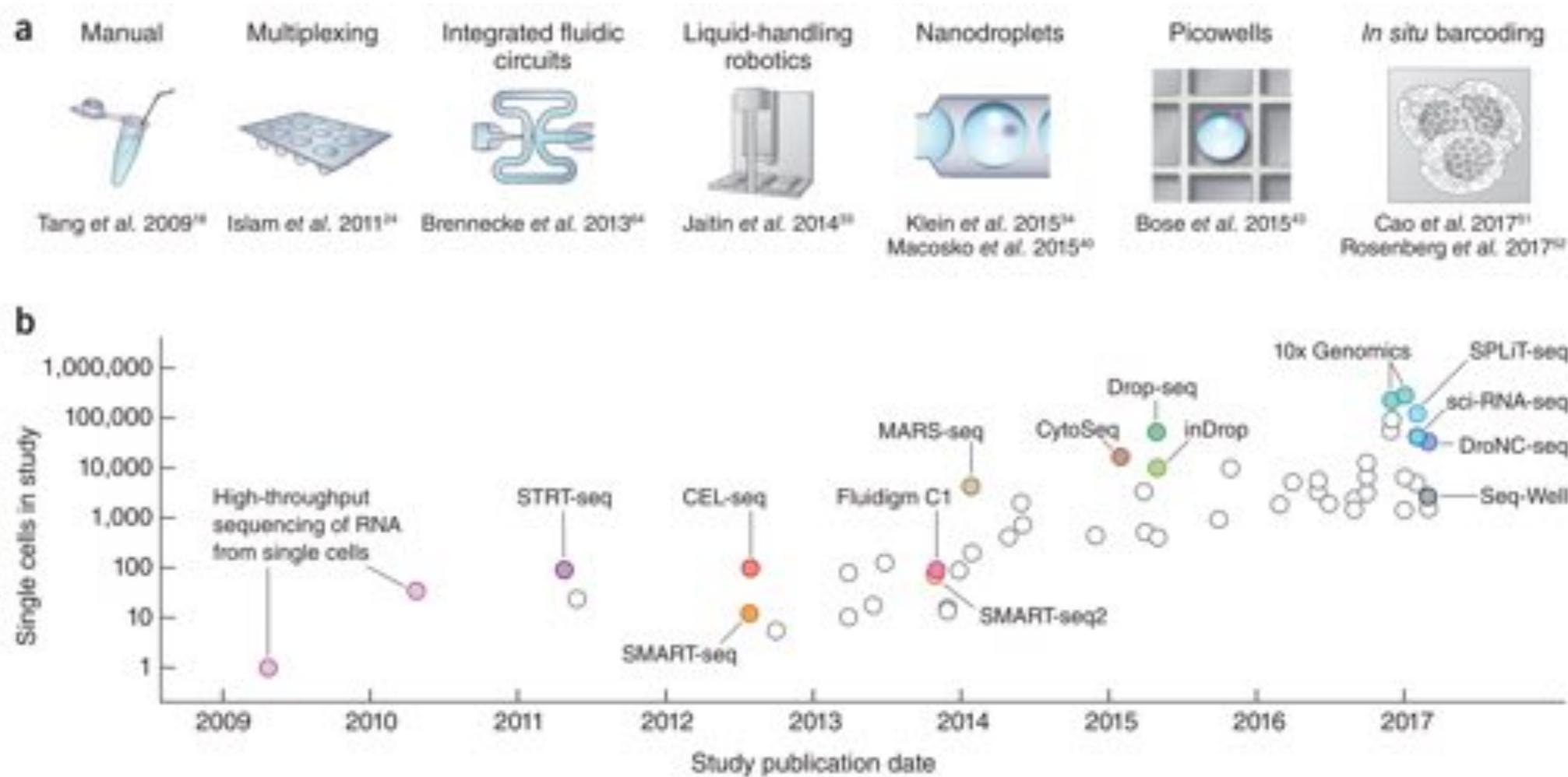


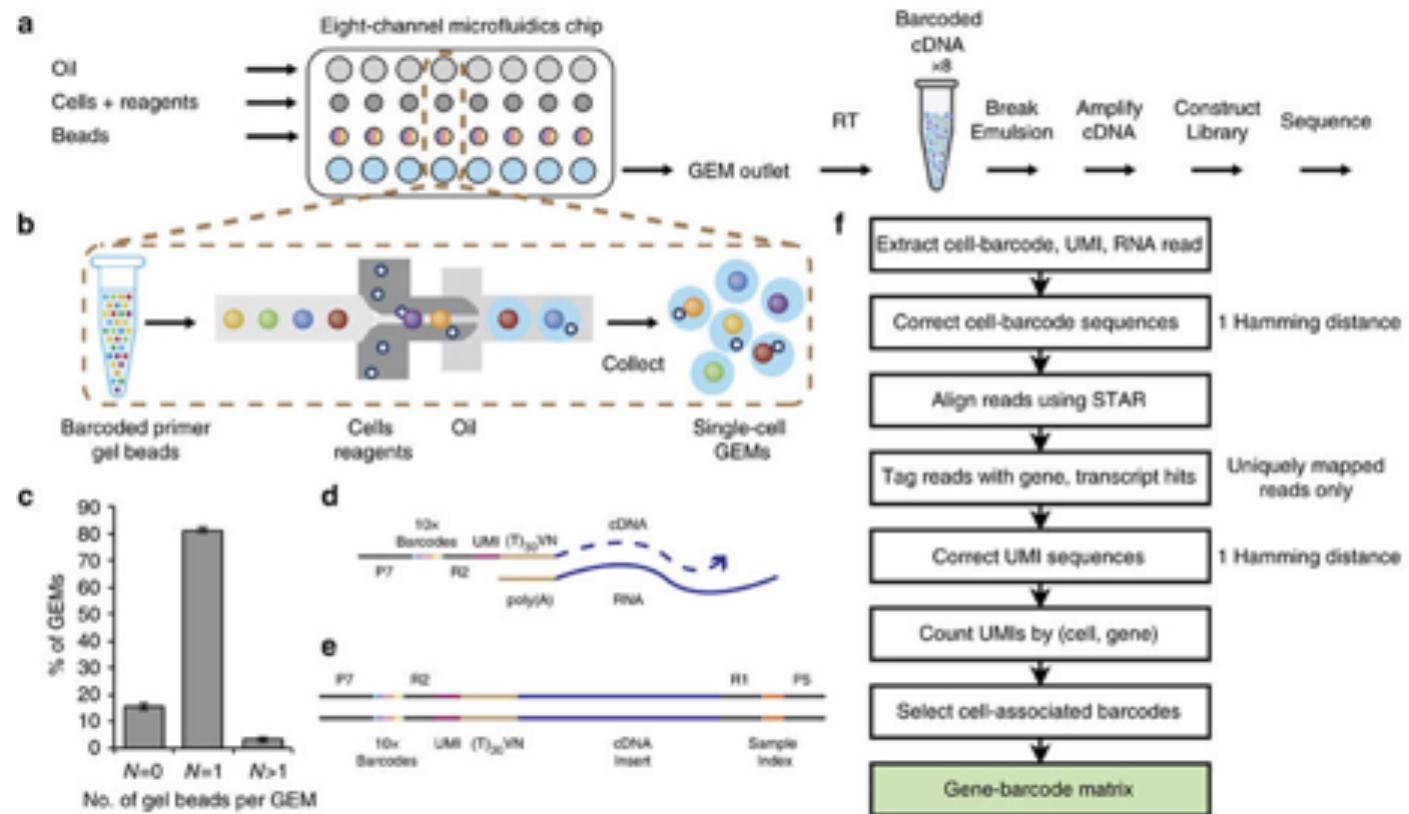
HUMAN CELL ATLAS



<https://www.humancellatlas.org/>

A decade of single-cell RNA-seq



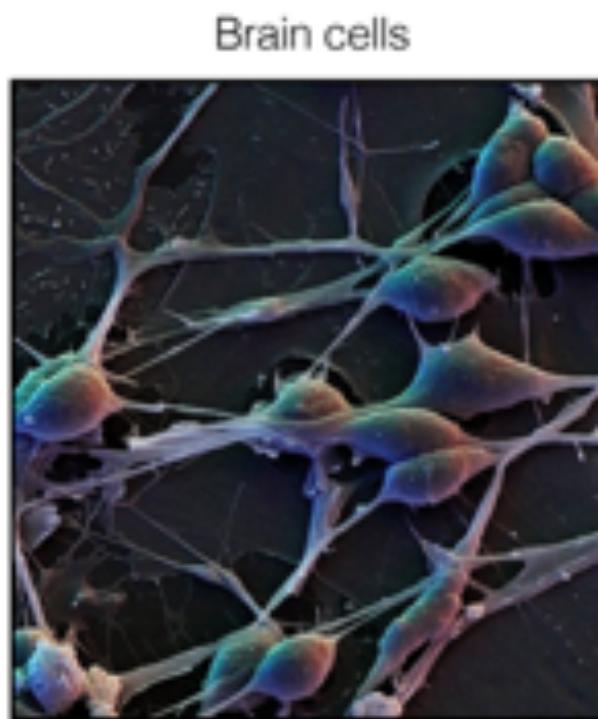


Up to 1M cells in a single analysis

Massively parallel digital transcriptional profiling of single cells
 Zheng et al (2017) Nature Communication. doi:10.1038/ncomms14049

Single-cell RNA sequencing, “the bioinformatician’s microscope”

— a snapshot of the underlying biology in a data matrix.



Biological sample

number of times gene i
was expressed in cell j

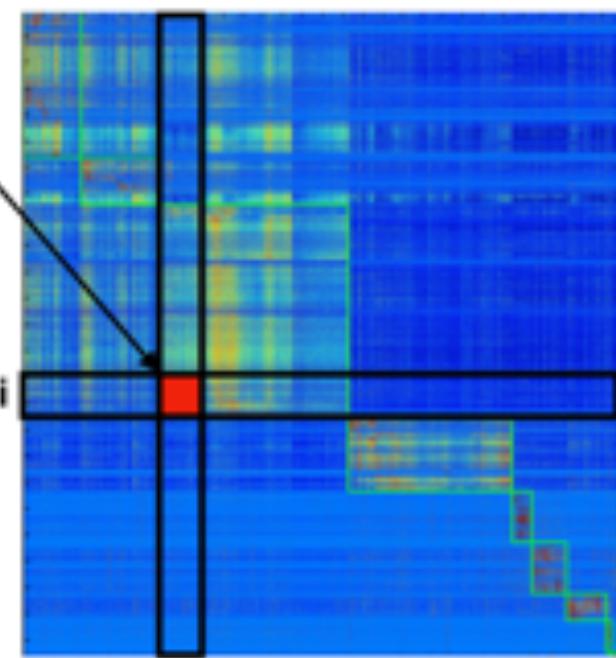


Single Cells (samples)

cell j

Genes (features)

gene i



Gene expression matrix

computationally explore complex biological systems

Martin Zhang

scSummary

Single cell analysis is a powerful tool to study heterogeneous tissues

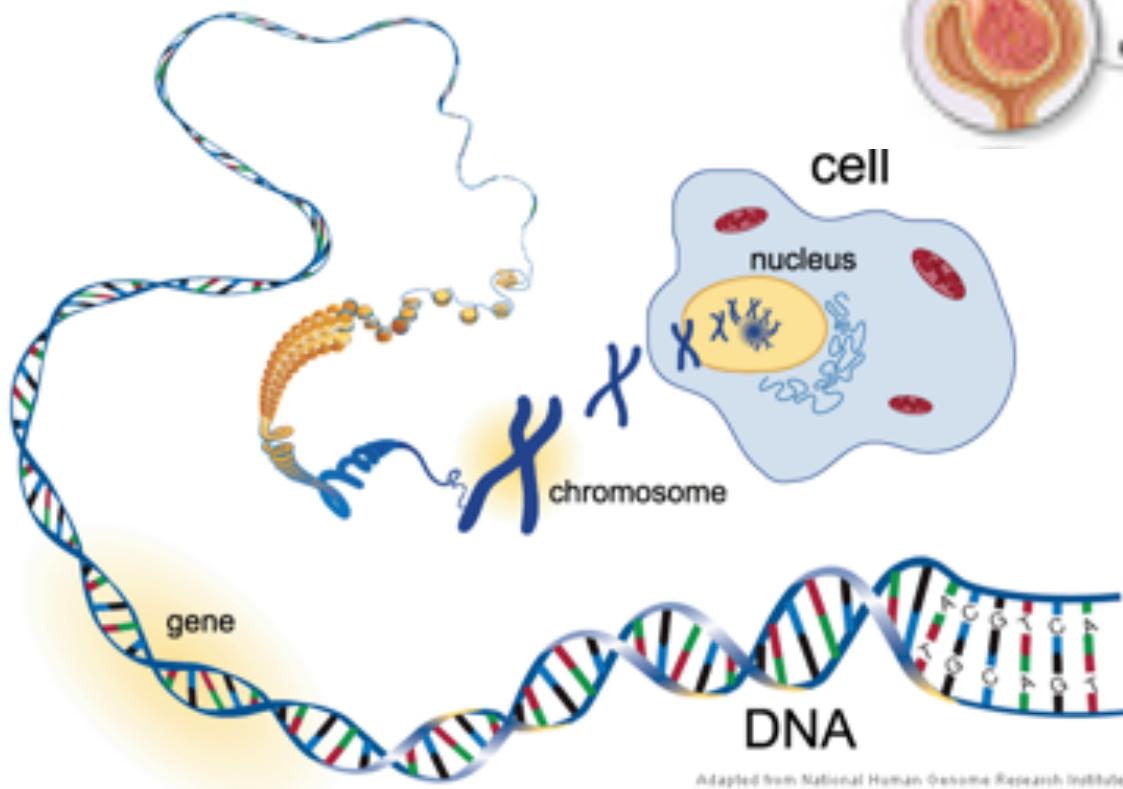
- Overcomes fundamental problems that can arise when averaging
- scRNA analysis used to identify novel cell types, understand the progression from one cell type to another across development or disease
- Many other sc-assays in development, expect 1000s to 1Ms of cells in essentially any assay

Major challenges

- Very sparse amplification and few reads per cell
- Find large CNVs, identify major cell types; hard to find small variants or perform differential expression
- Allelic-dropout and unbalanced amplification hides or distorts information
- Use statistical approaches to smooth results based on prior information or other cells from the same cell type
- Need new ways to process and analyze millions of cells at a time

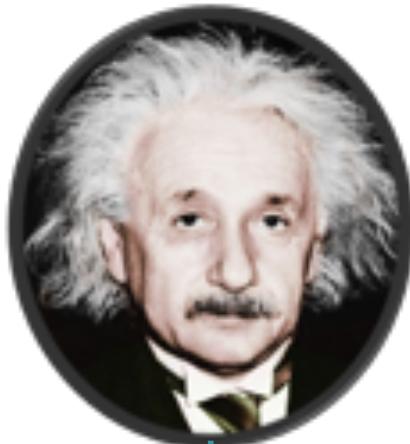
Why Genes?

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



~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



~5 Mya

~75 Mya

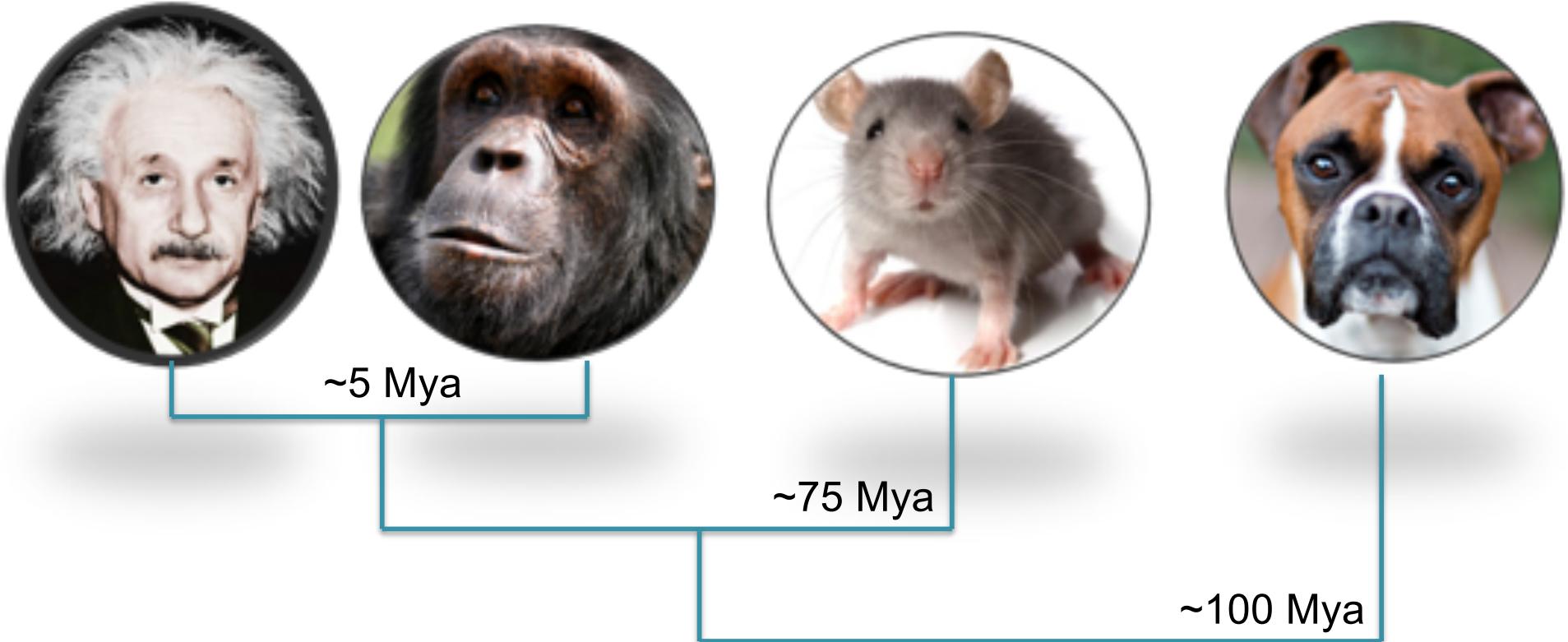
"In the roughly 75 million years since the divergence of the human and mouse lineages, the process of evolution has altered their genome sequences and caused them to diverge by ***nearly one substitution for every two nucleotides***"

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

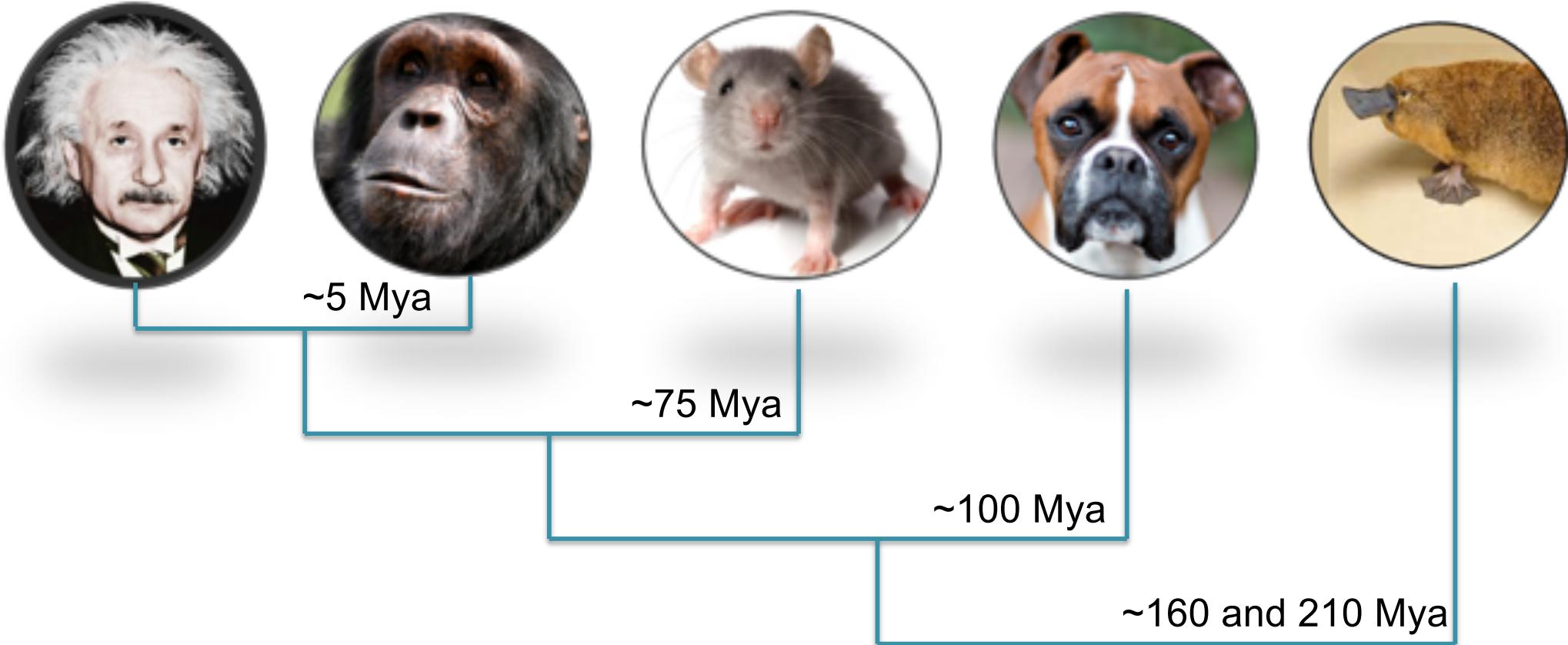
Human Evolution



"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

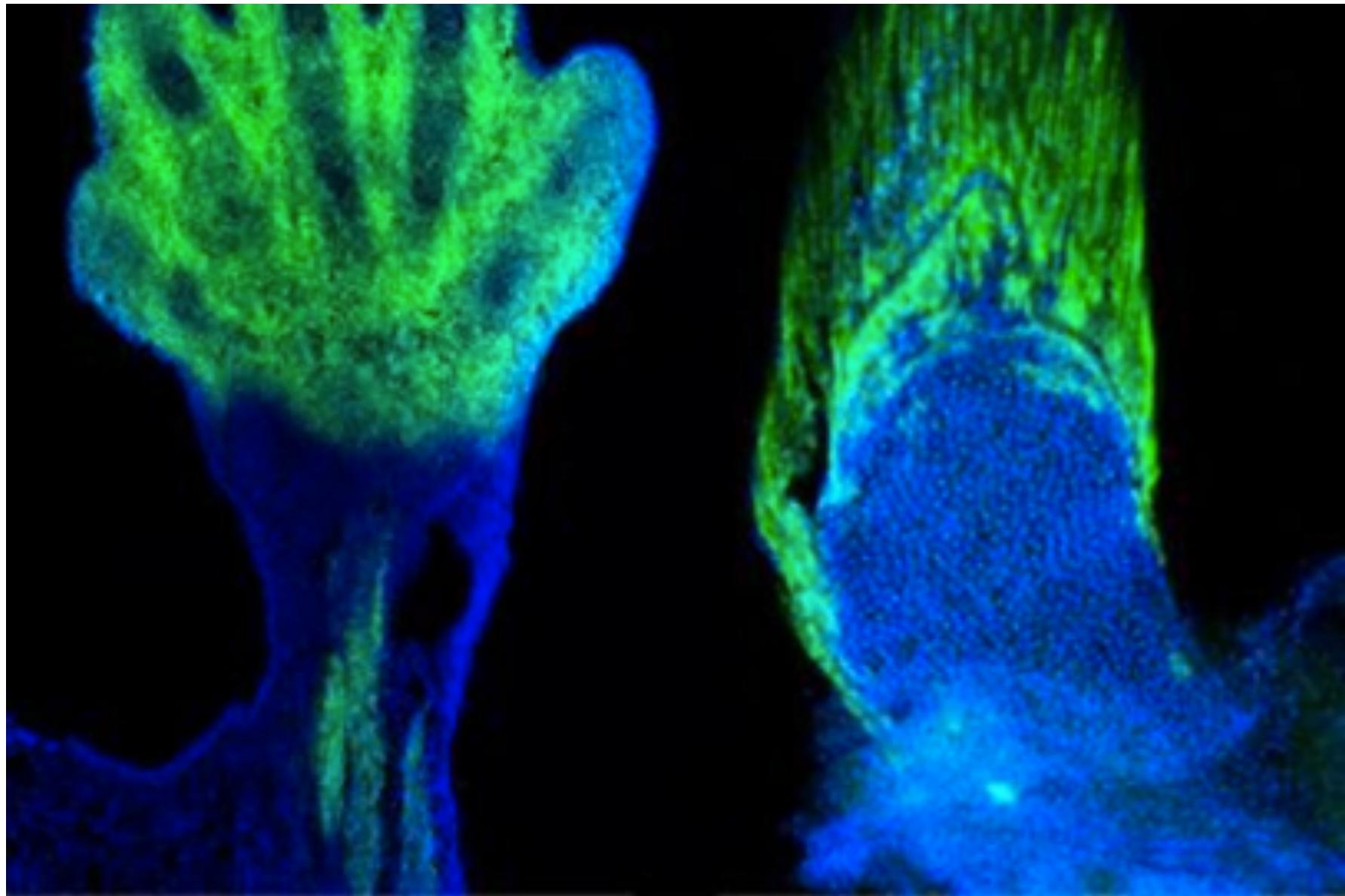
Human Evolution



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

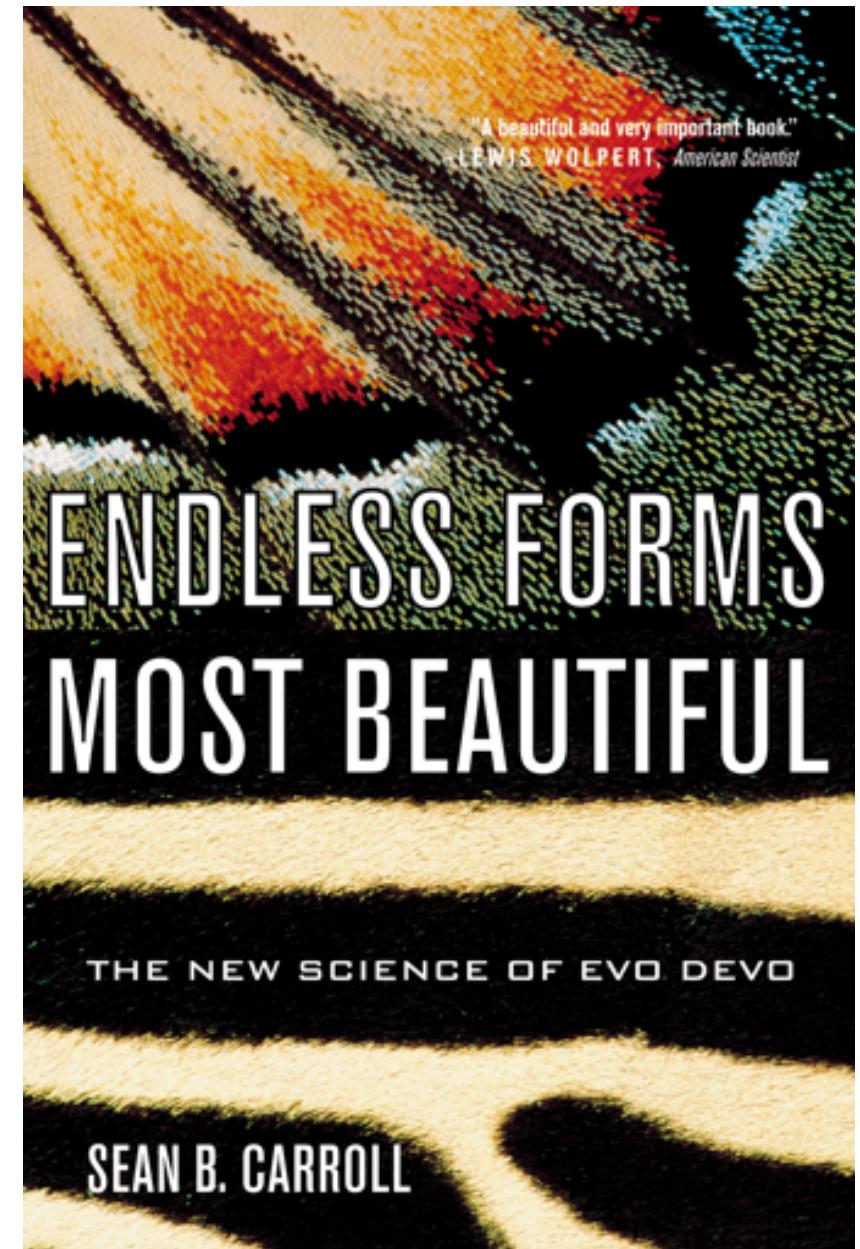
Nakamura et al (2016) *Nature*. 537, 225–228. doi:10.1038/nature19322

More Information



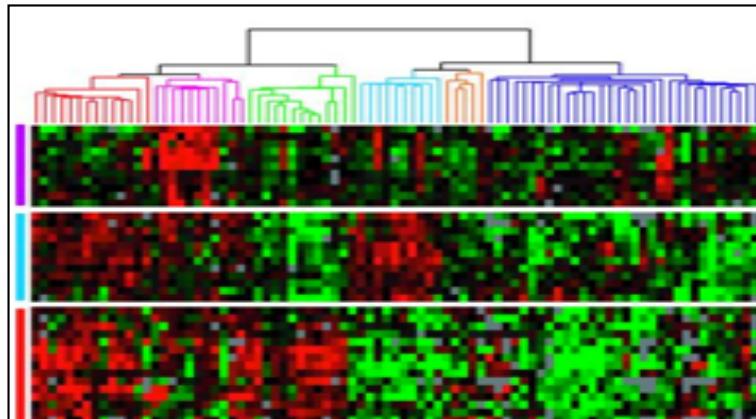
*“Anything found to be true of
E. coli must also be true of
elephants”*

-Jacques Monod

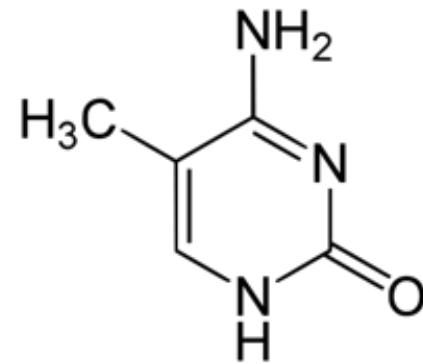


*-seq in 4 short vignettes

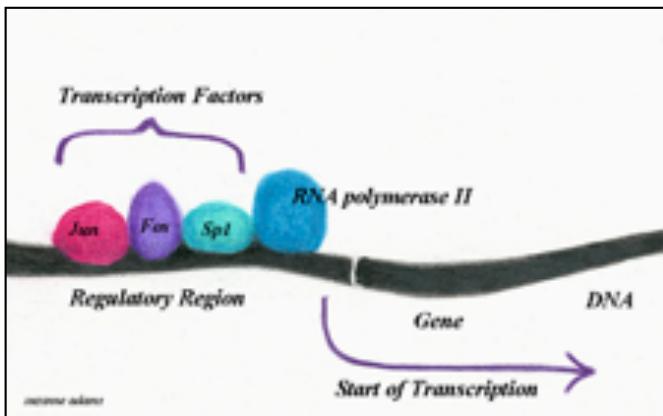
RNA-seq



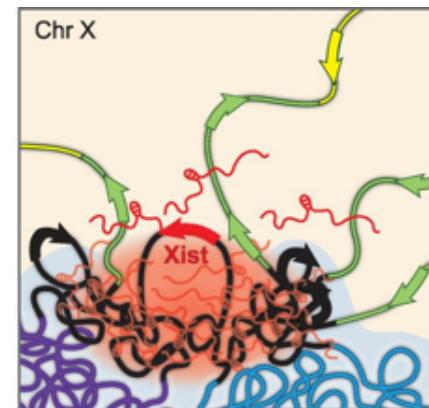
Methyl-seq



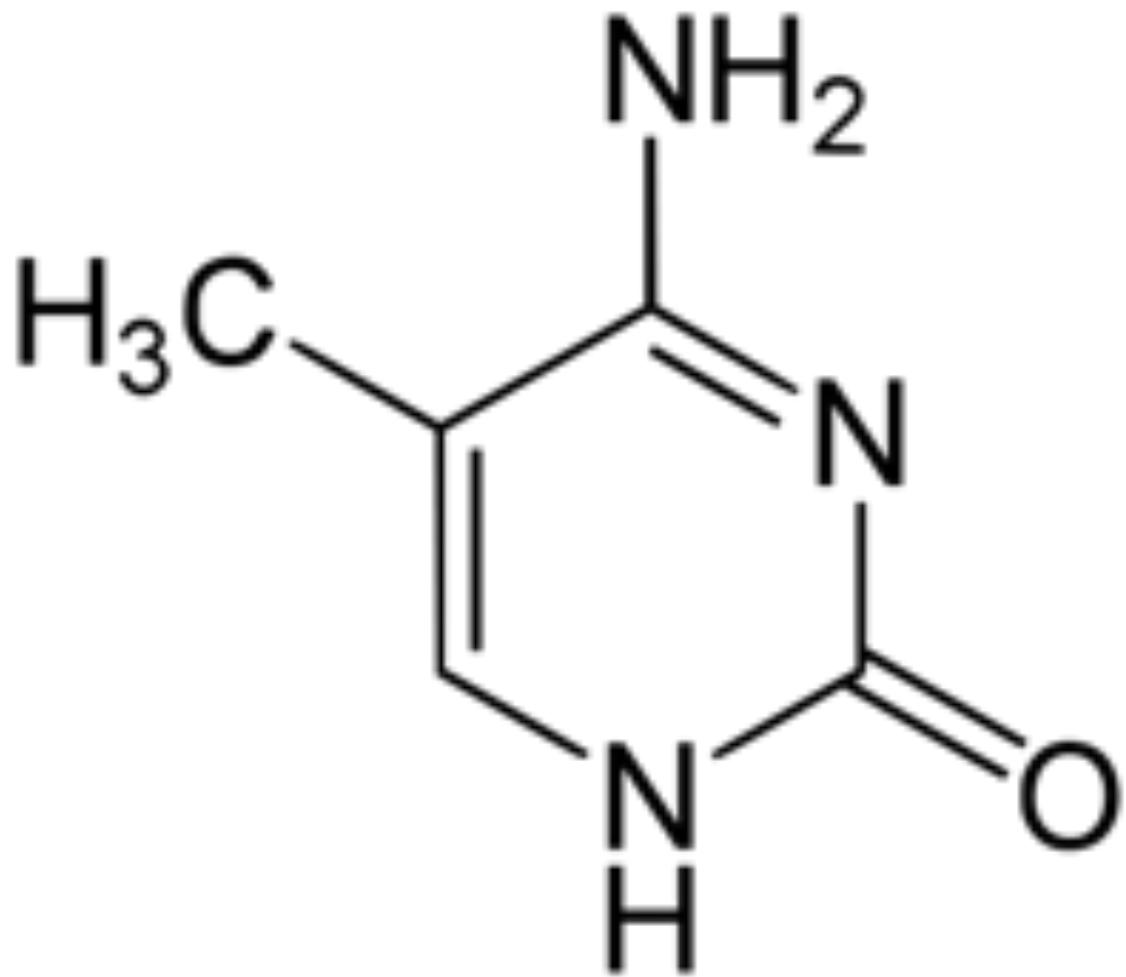
ChIP-seq



Hi-C

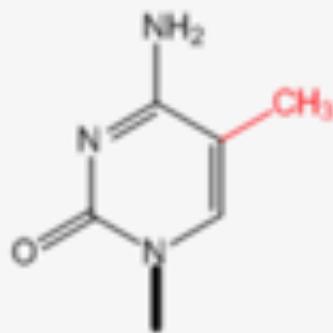


Methyl-seq

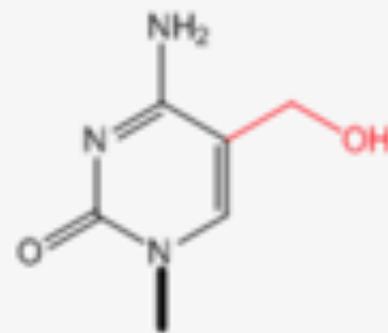


Finding the fifth base: Genome-wide sequencing of cytosine methylation
Lister and Ecker (2009) *Genome Research*. 19: 959-966

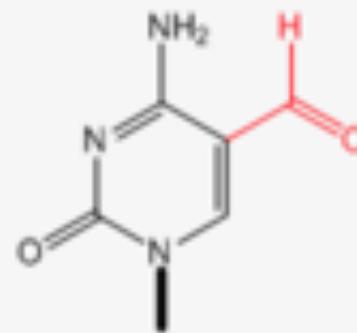
Epigenetic Modifications to DNA



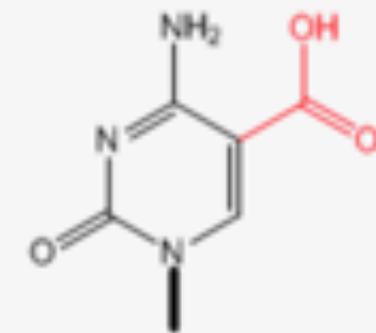
5-mC



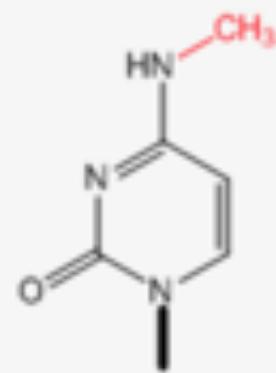
5-hmC



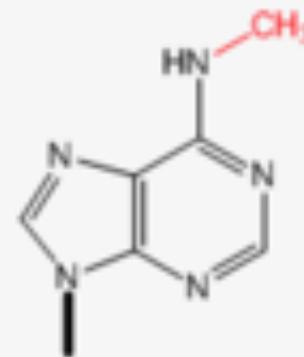
5-fC



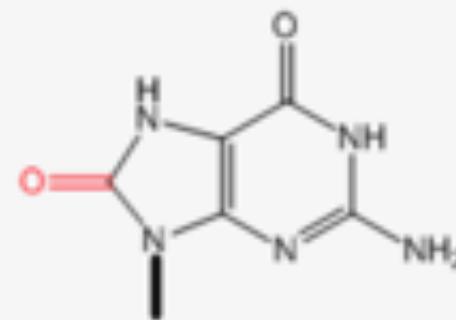
5-caC



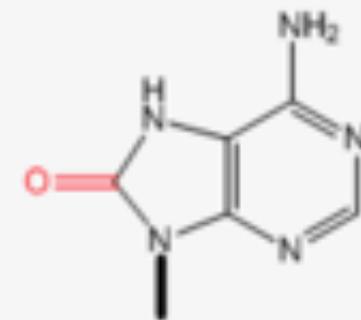
4-mC



6-mA



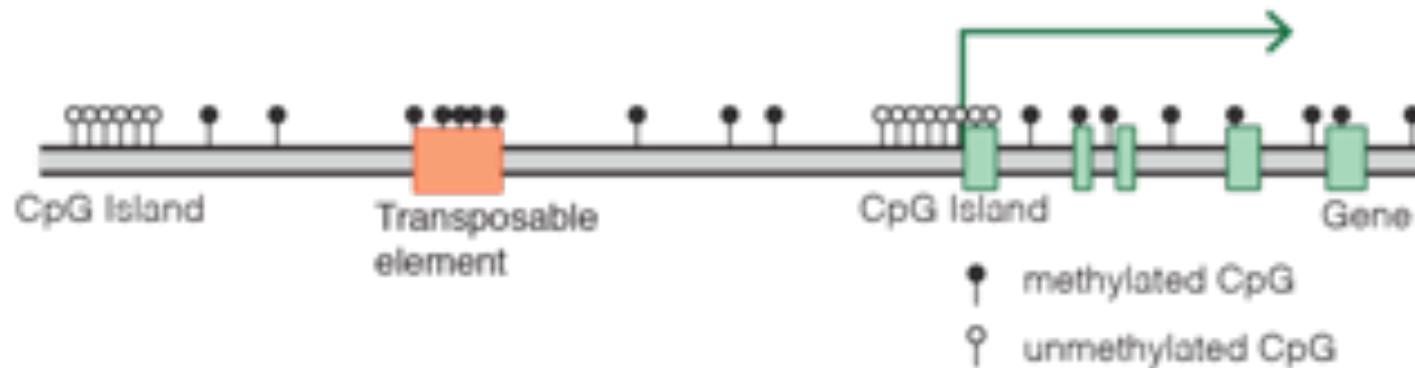
8-oxoG



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 Lyko^{1*}, 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^{8,12-20}. 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

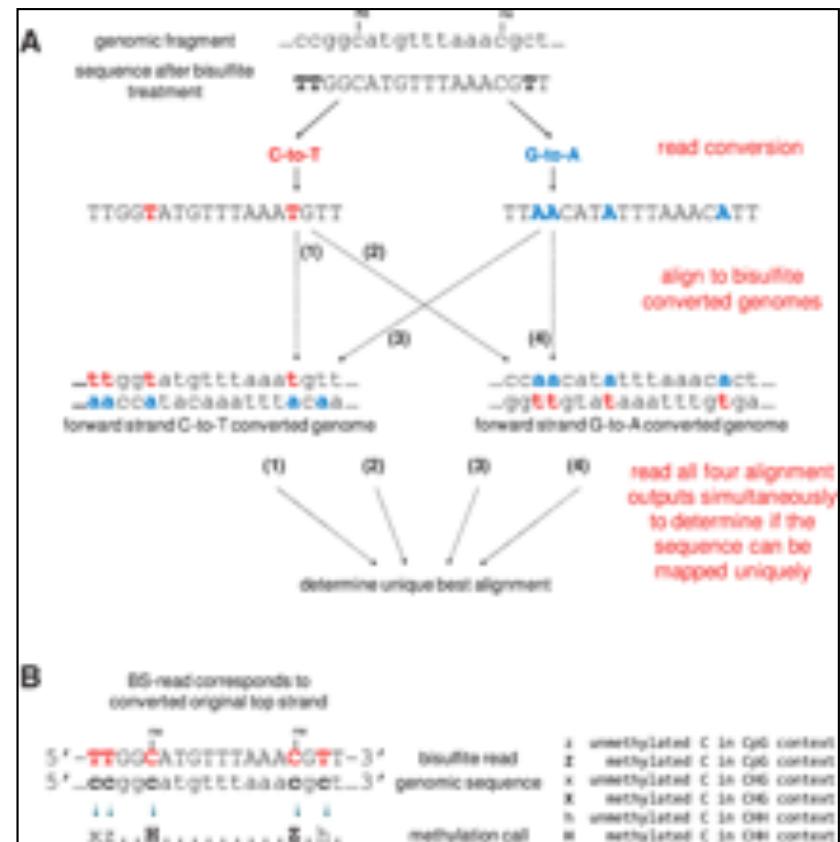
Table 1 Quantitation of methylation of specific genes in human cancers and adjacent analogous normal tissues

Patient	Carcinoma	Probe	Enzyme	% Hypomethylated fragments		
				N	C	M
1	Colon	HGH	{ <i>Hpa</i> II	<10	35	—
			{ <i>Hha</i> I	<10	39	—
			{ <i>Hpa</i> II	<10	52	—
		{ <i>Hha</i> I	{ <i>Hpa</i> II	<10	39	—
			{ <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	—
			{ <i>Hpa</i> II	<10	58	—
		{ <i>Hha</i> I	{ <i>Hpa</i> II	<10	23	—
			{ <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	—
			{ <i>Hpa</i> II	<10	50	—
		γ-Globin	{ <i>Hpa</i> II	<10	99	—

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.

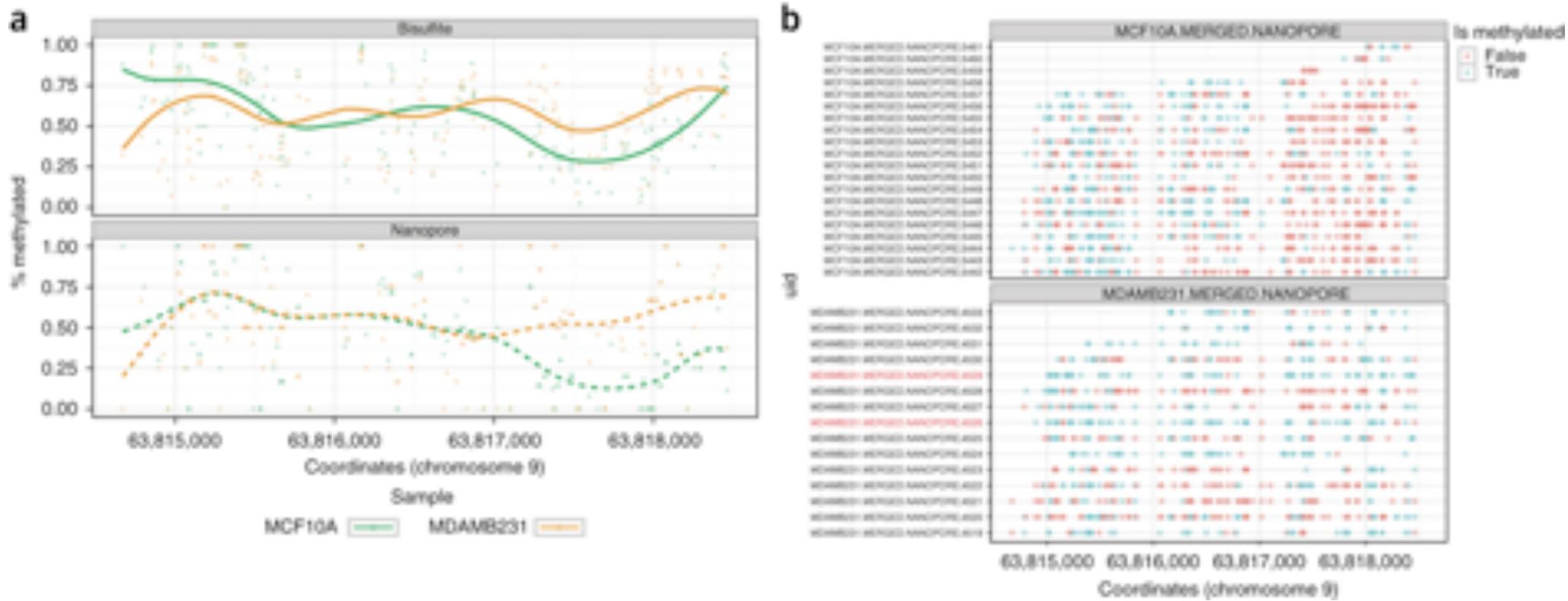
Bisulfite Conversion

To
w
•
•
•
•
•



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

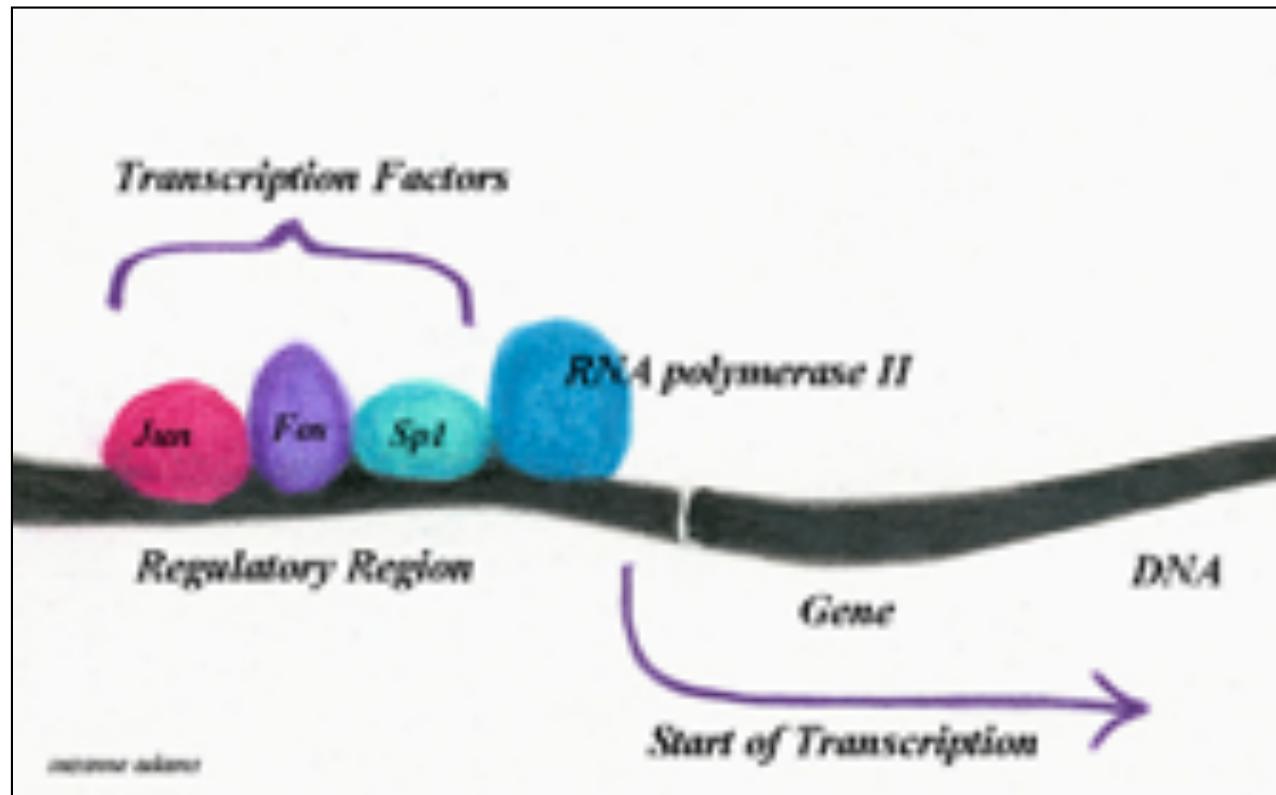


Comparison of bisulfite sequencing and nanopore-based R7.3 data in reduced representation data sets from cancer and normal cells. (a) Raw data (points) and smoothed data (lines) for methylation, as determined by bisulfite sequencing (top) and nanopore-based sequencing using an R7.3 pore (bottom), in 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

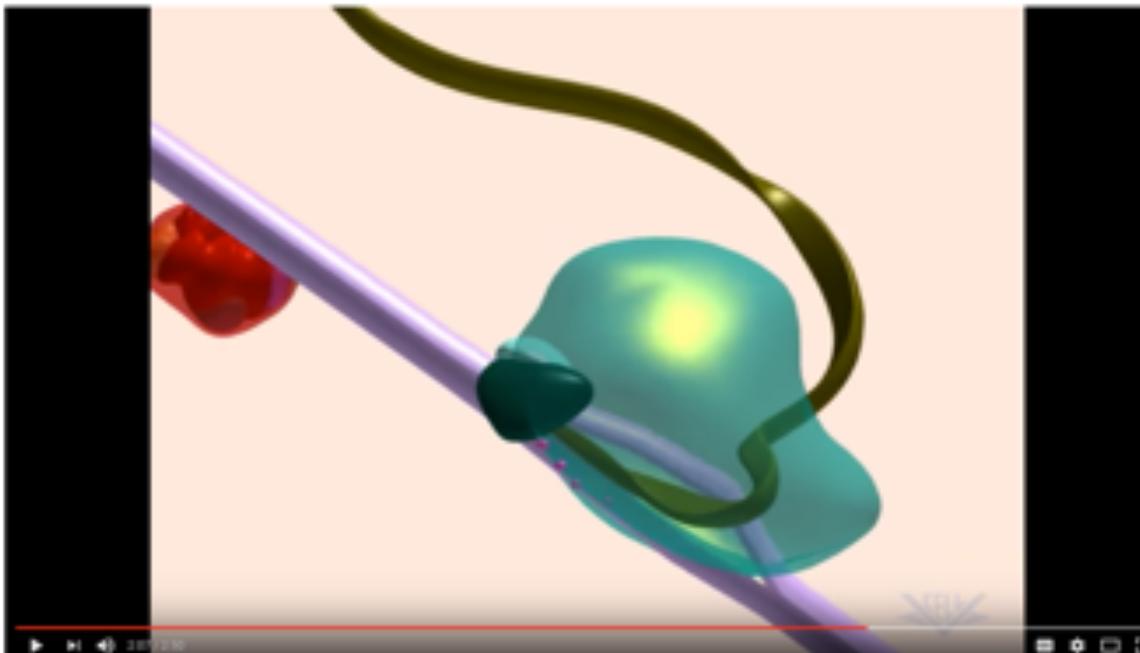
ChIP-seq



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

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

Transcription



The image shows a YouTube video player window. The main video frame displays a 3D molecular model of transcription. A purple RNA polymerase enzyme is shown binding to a segment of a double-stranded DNA molecule. One strand of the DNA is transcribing into a green RNA strand. The video has a play button at the bottom left and a progress bar indicating it is at 2:07 of 2:10. Below the video, the title "Transcription" and "2,018,430 views" are visible. On the right side of the player, there is a "SUBSCRIBE" button and a "4K" resolution indicator. To the right of the video frame, a sidebar titled "Up next" lists several other YouTube videos related to DNA, transcription, and translation.

Up next

- Transcription and Translation: From DNA to Protein
Professor Dave Explains
101K views
- DNA - transcription and translation
Wilson Kubista
40K views
- Transcription and mRNA processing | Biomolecules | Khan Academy
100K views
- DNA transcription and translation Animation
Hader add
33K views
- Translation
bioumuahmed
2.1M views
- Transcription and Translation Overview
Amarando Hastings
57K views
- DNA, Histone, & The Longest Word Ever Crash Course
CrashCourse
2.2M views
- TRANSCRIPTION 1
theacademymedicine
25K views
- TRANSCRIPTION
conghuang
70K views
- Moana - Best Scenes (HD)

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

Transcription Factors

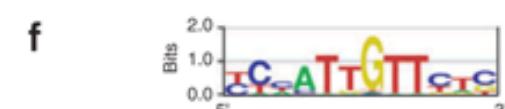
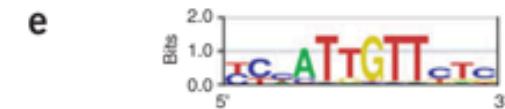
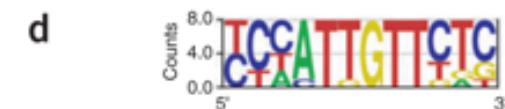
A transcription factor (or sequence-specific DNA-binding factor) is a protein that controls the rate of transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence.

- 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*.

a HEM13 CCCATTGTTCTC
HEM13 TTTCTGGTTCTC
HEM13 TCAATTGTTTAG
ANB1 CTCATTGTTGTC
ANB1 TCCATTGTTCTC
ANB1 CCTATTGTTCTC
ANB1 TCCATTGTTCGT
ROX1 CCAATTGTTTTG

b YCHAATTGTTCTC

c
A 002700000010
C 464100000505
G 000001800112
T 422087088261

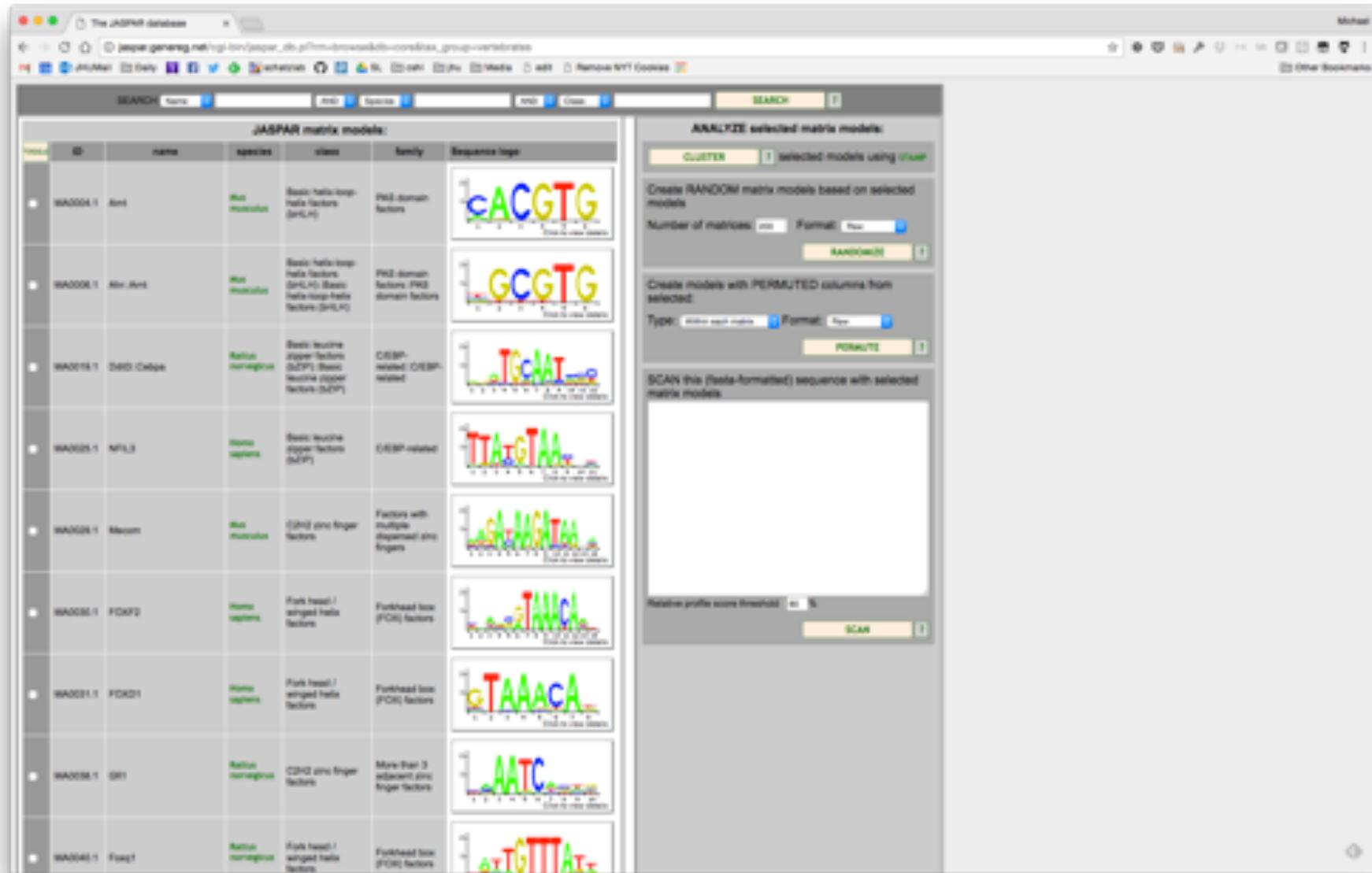


Bob Crimi

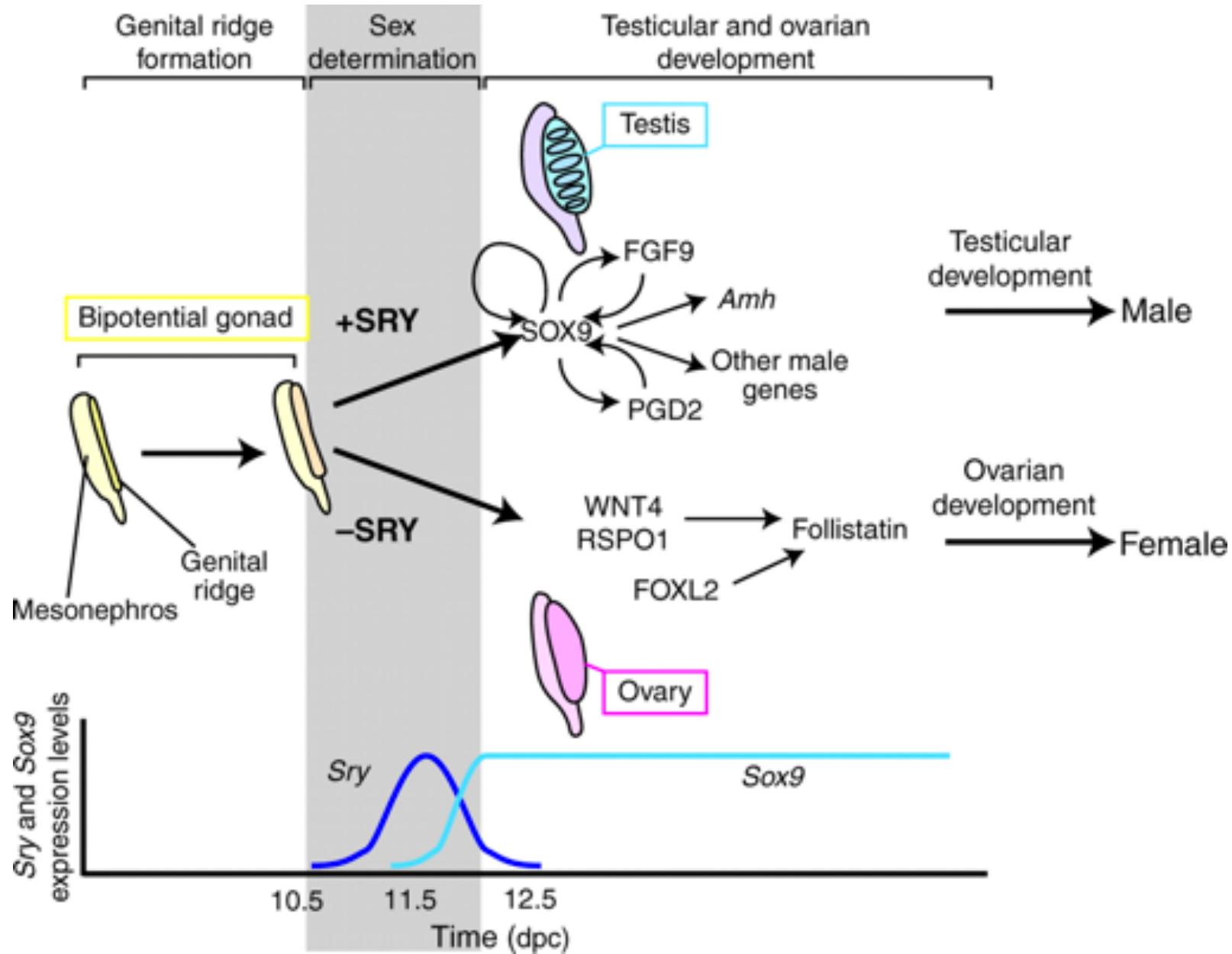
What are DNA sequence motifs?

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

Transcription Factors Database

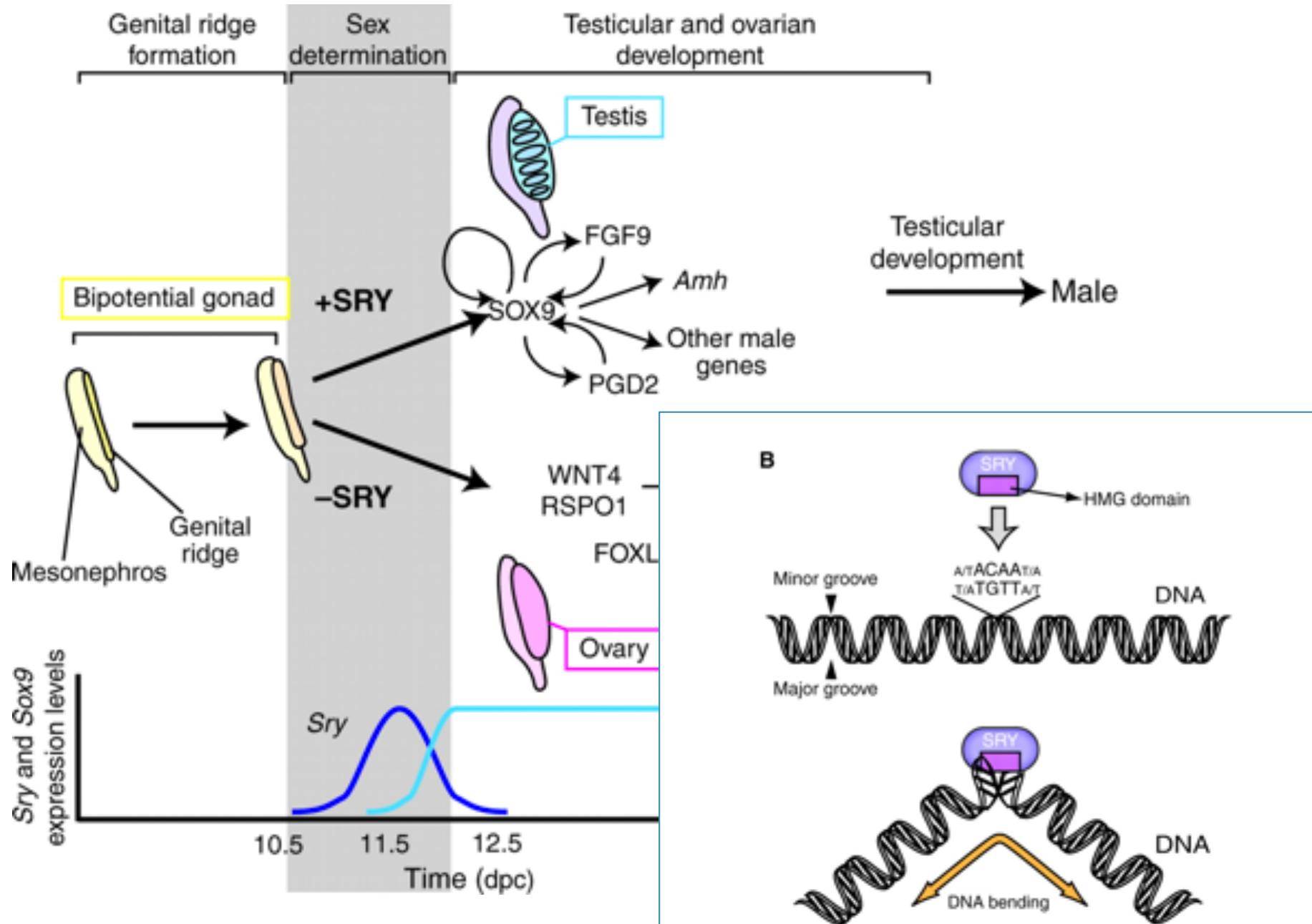


JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles
Anthony Mathelier (2014) Nucleic Acids Res. 42 (D1): D142-D147. DOI: <https://doi.org/10.1093/nar/gkt997>



SRY: The master switch in mammalian sex determination

Kashimada and Koopman (2010) Development 137: 3921-3930; doi: 10.1242/dev.048983



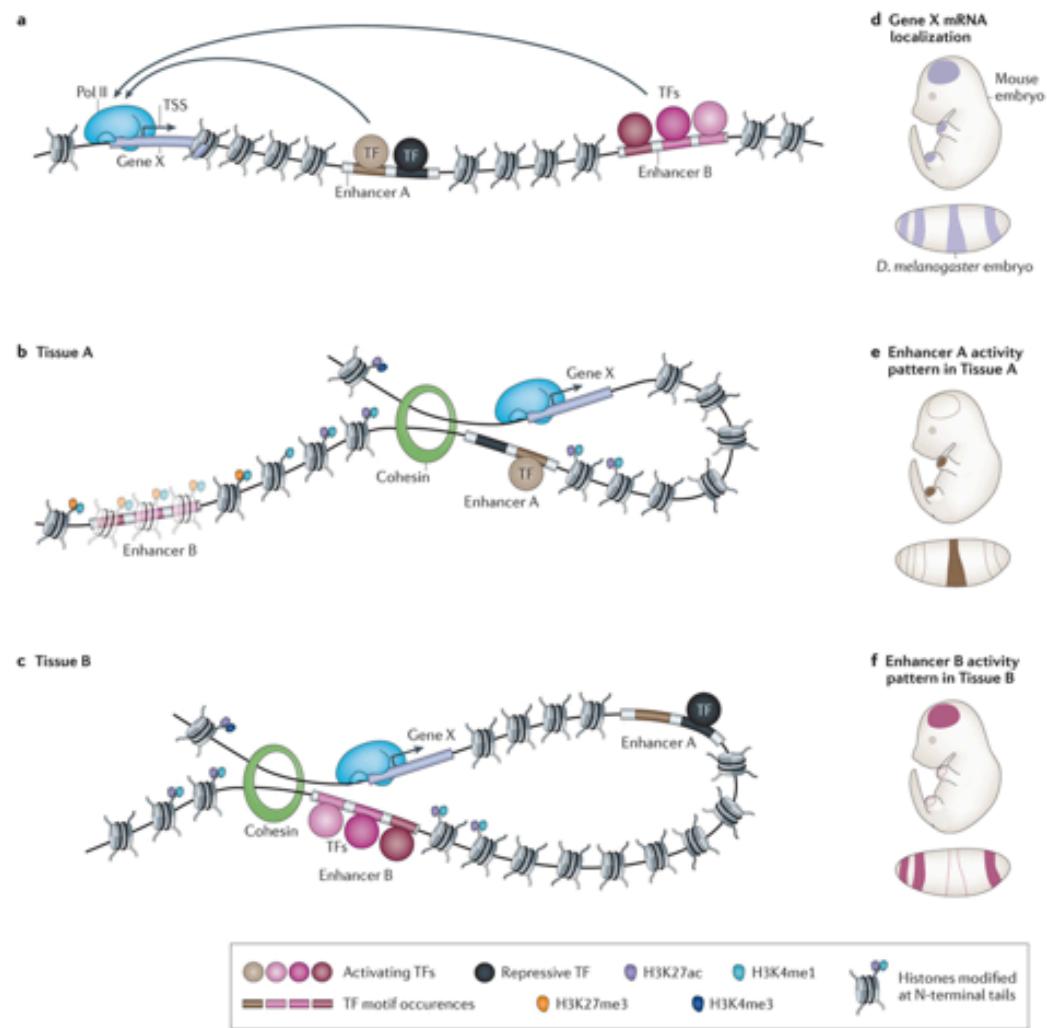
SRY: The master switch in mammalian sex determination

Kashimada and Koopman (2010) Development 137: 3921-3930; doi: 10.1242/dev.048983

Enhancers

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

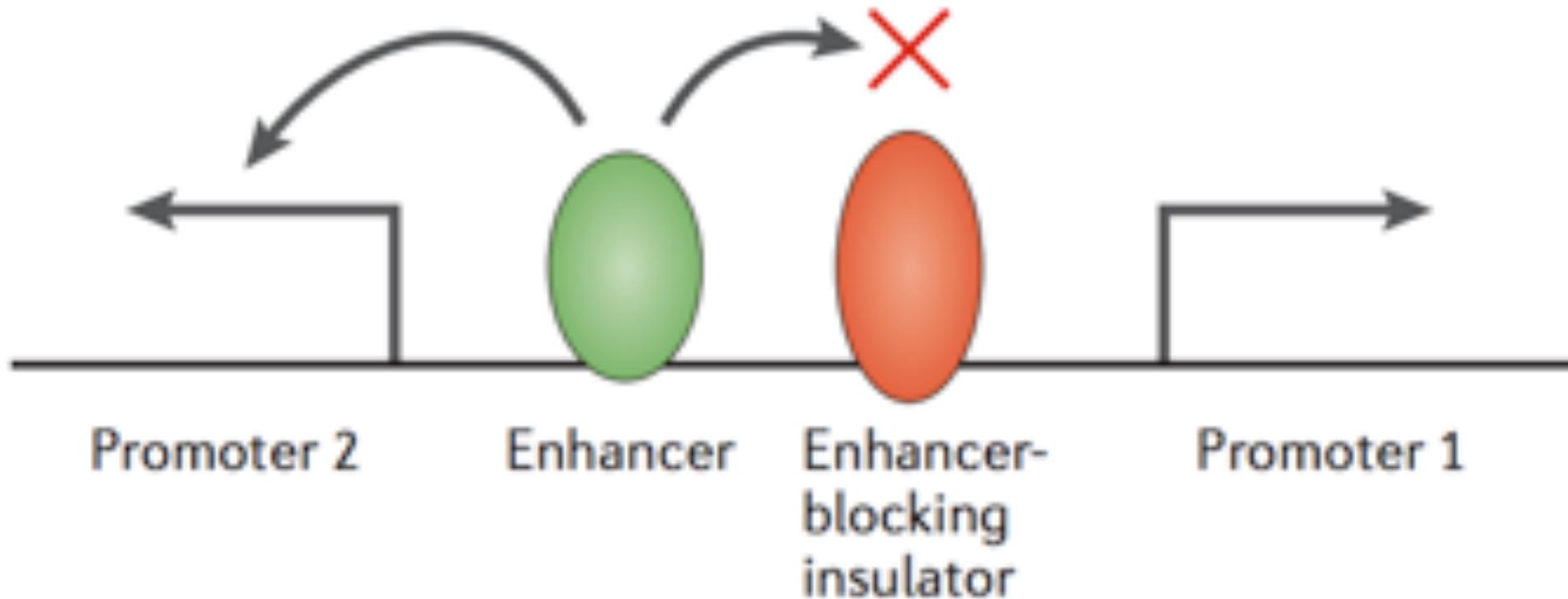
- Enhancers can be located at any distance from their target genes (up to ~1Mbp)
- In a given tissue, active enhancers (Enhancer A in part b or Enhancer B in part c) are bound by activating TFs and are brought into proximity of their respective target promoters by looping
- Active and inactive gene regulatory 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.

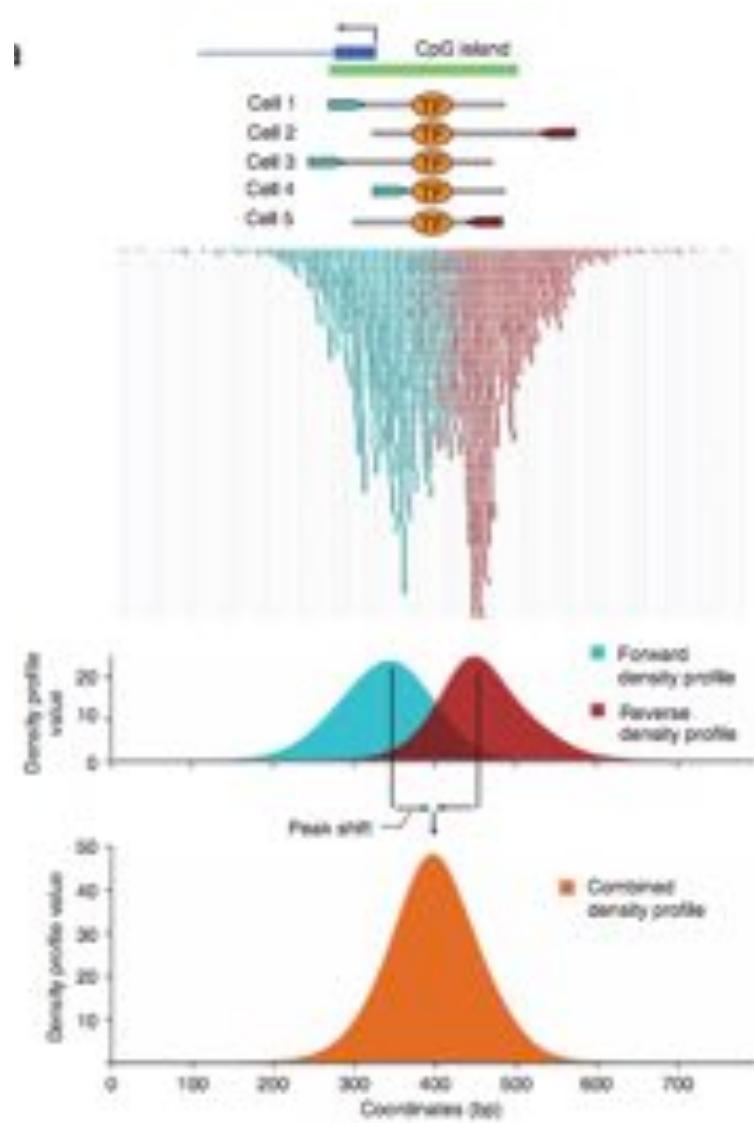
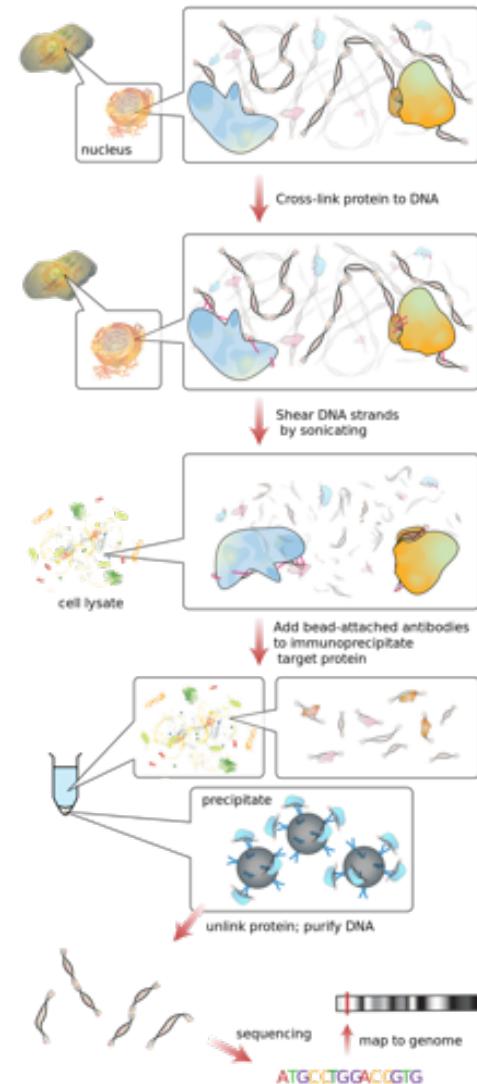
Insulators: exploiting transcriptional and epigenetic mechanisms

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

ChIP-seq: TF Binding

Goals:

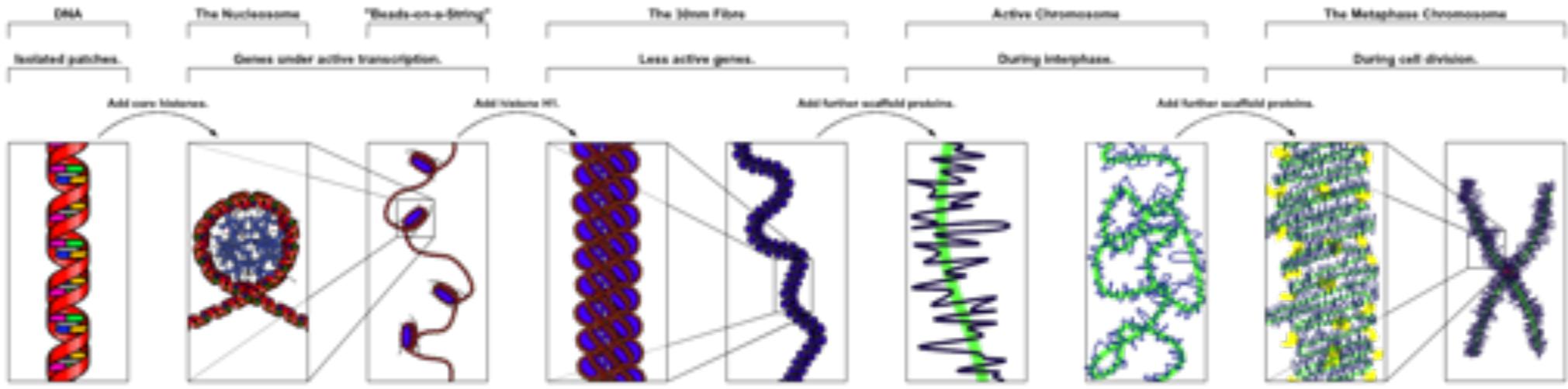
- Where are transcription factors and other proteins binding to the DNA?
- How strongly are they binding?
- Do the protein binding patterns change over developmental stages or when the cells are stressed?



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



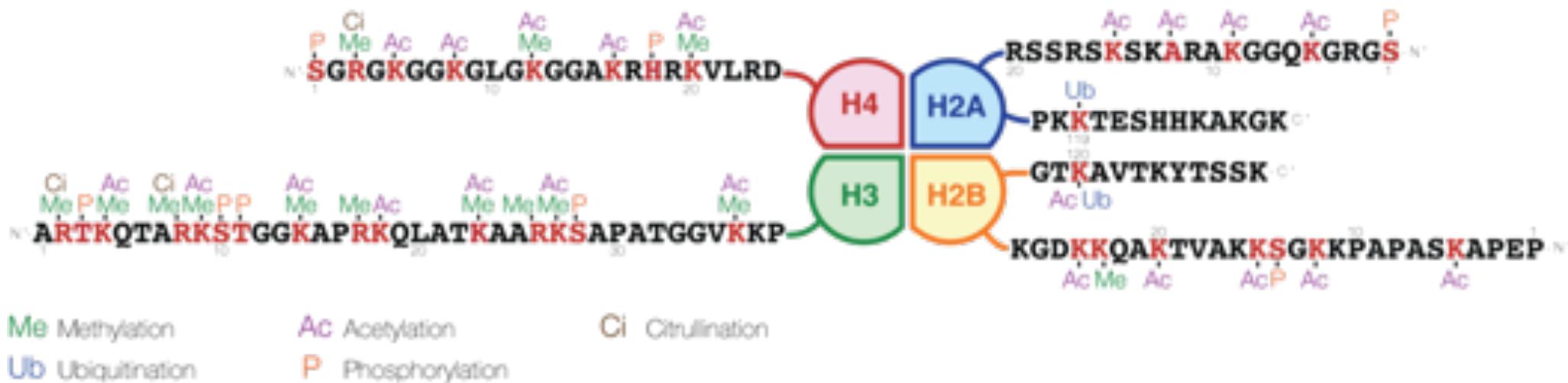
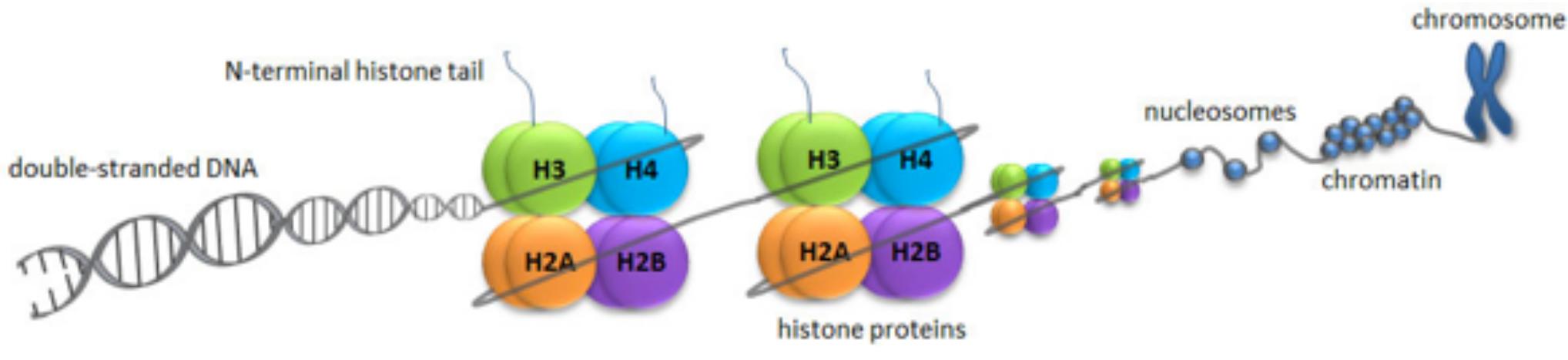
Nucleosome is a basic unit of DNA packaging in eukaryotes

- 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

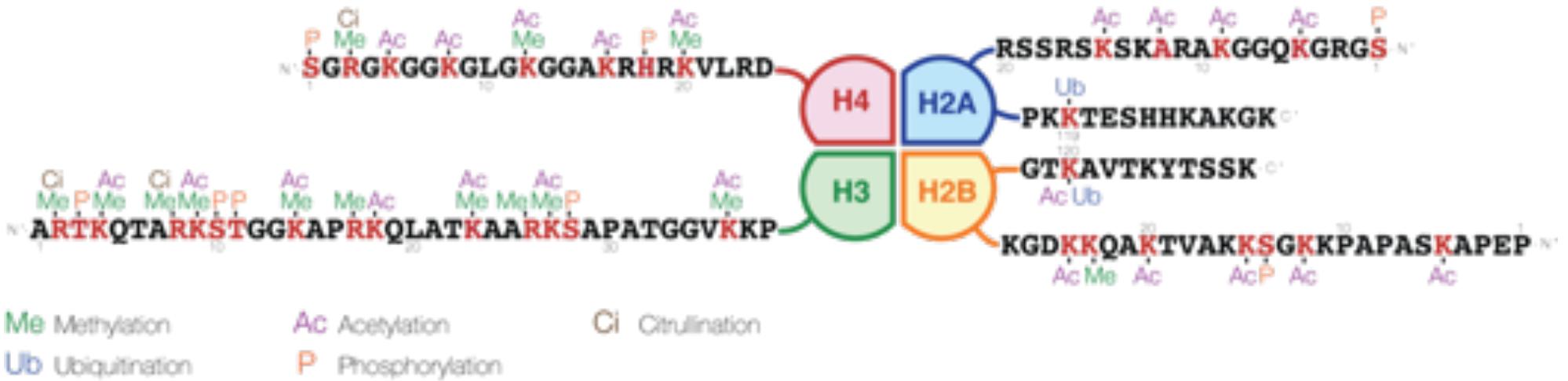
Nucleosomes form the fundamental repeating units of eukaryotic chromatin

- 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



ChIP-seq: Histone Modifications

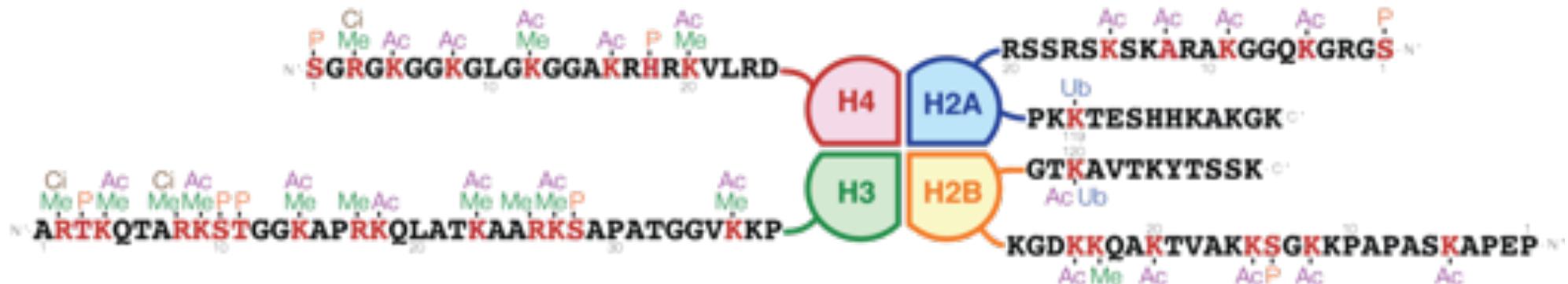


The common nomenclature of histone modifications is:

- 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.

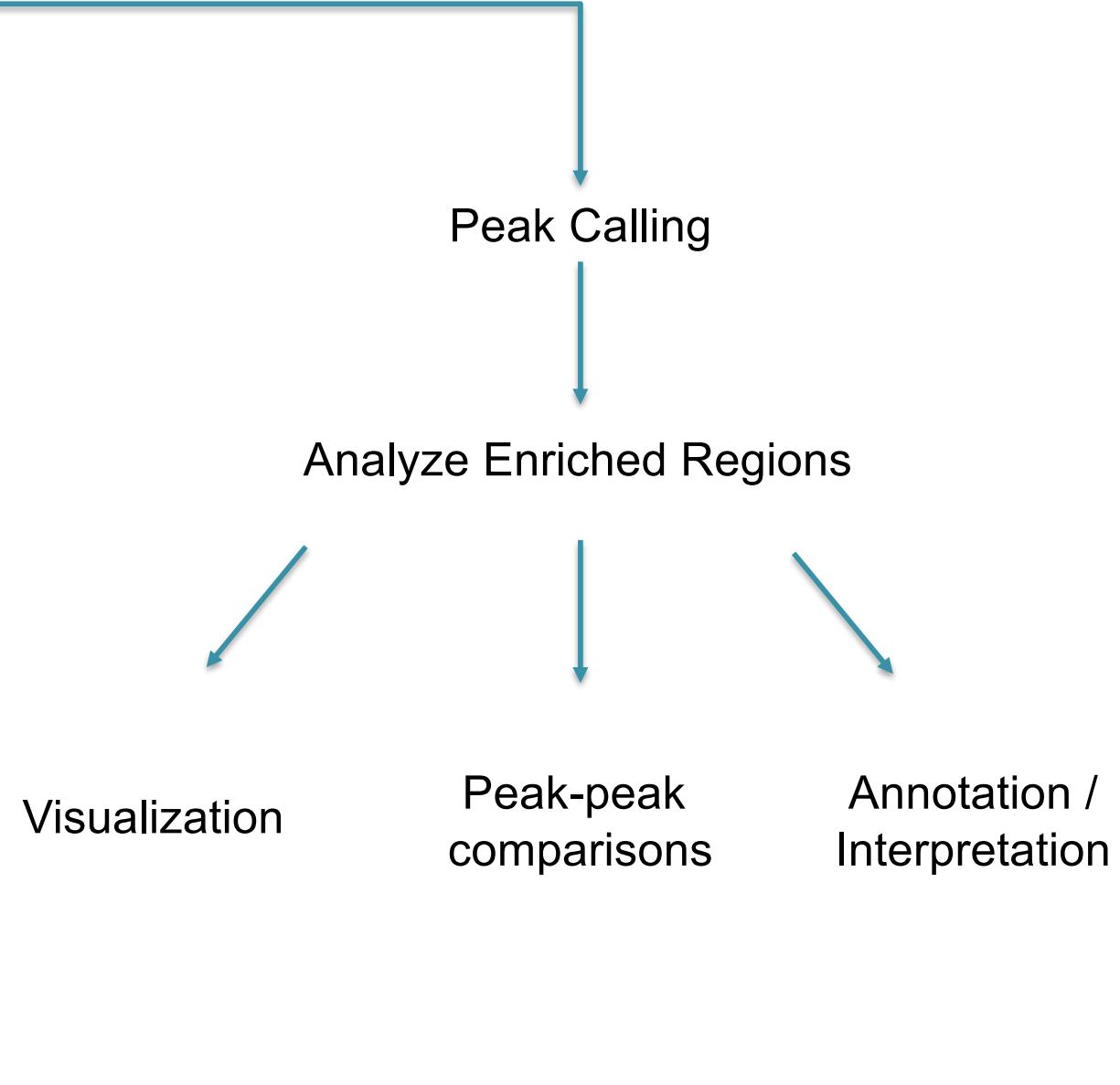
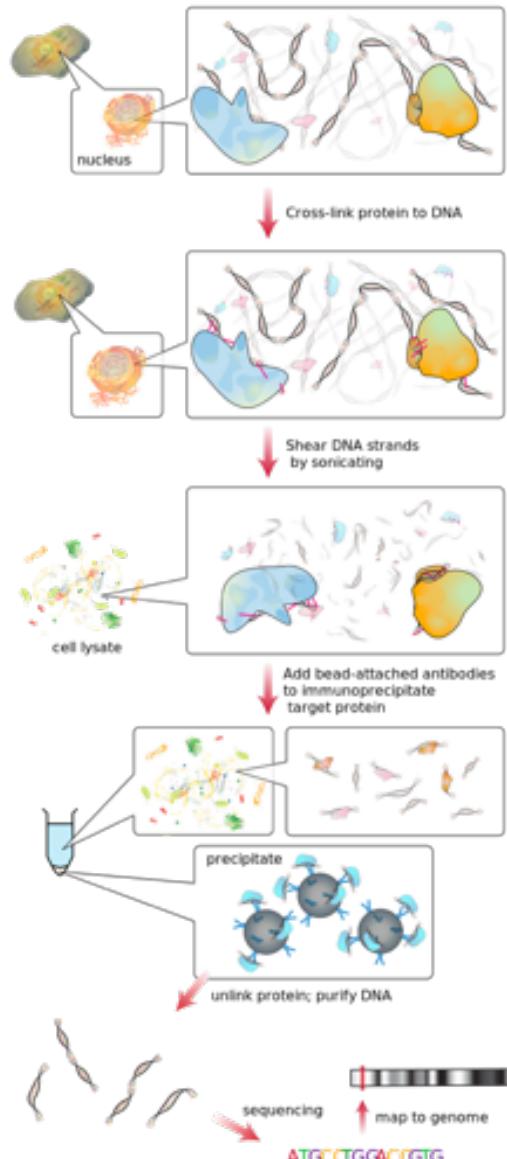
ChIP-seq: Histone Modifications



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 ^[8]	activation ^[9]	activation ^[10]		activation ^[11]		

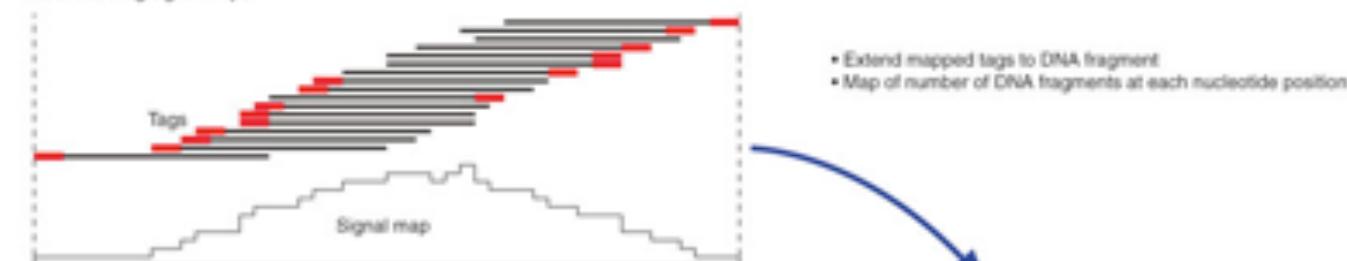
- H3K4me3 is enriched in transcriptionally active promoters.^[12]
- H3K9me3 is found in constitutively repressed genes.
- H3K27me is found in facultatively repressed genes.^[7]
- H3K36me3 is found in actively transcribed gene bodies.
- H3K9ac is found in actively transcribed promoters.
- H3K14ac is found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

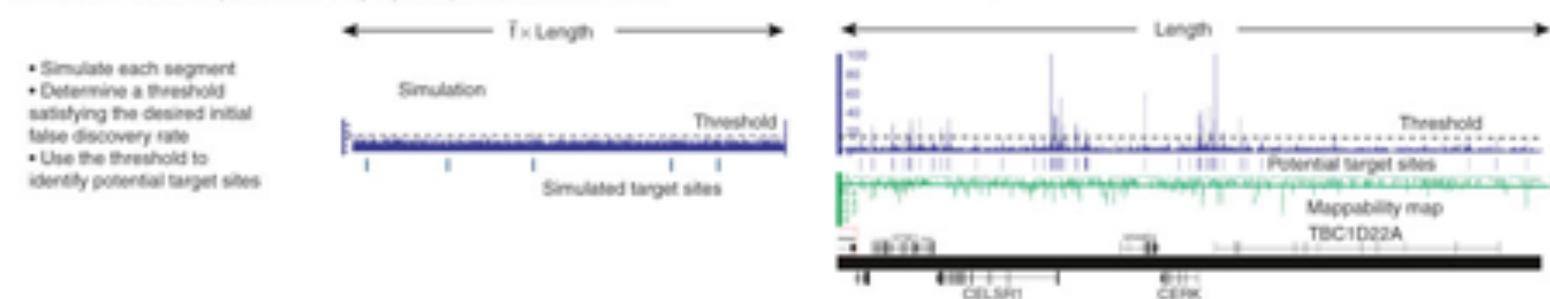


PeakSeq

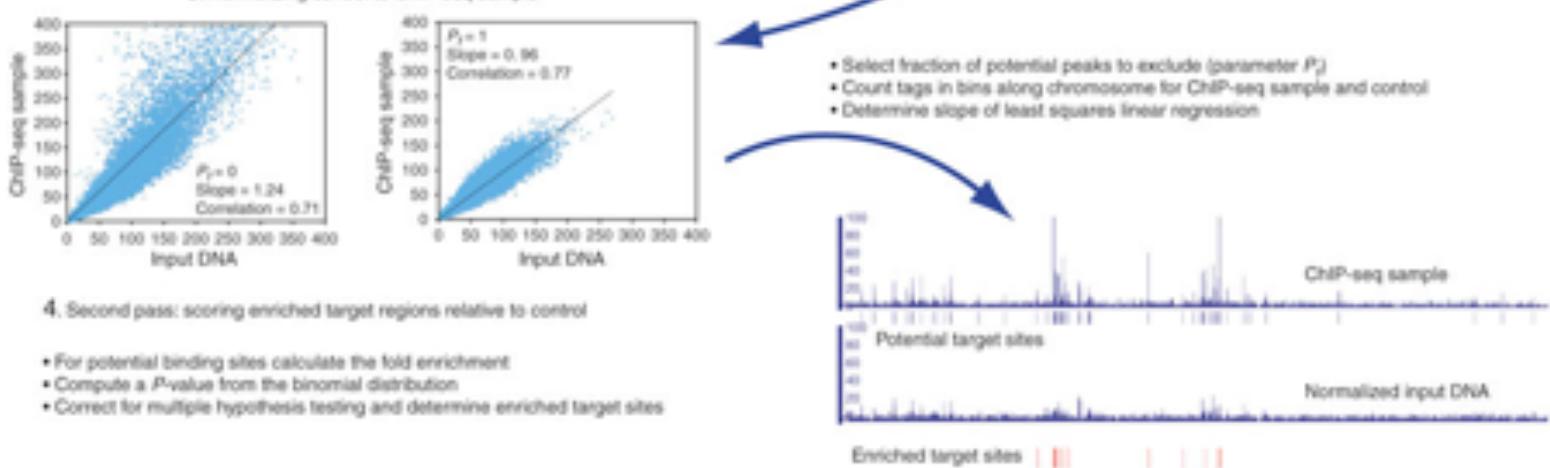
1. Constructing signal maps



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



3. Normalizing control to ChIP-seq sample



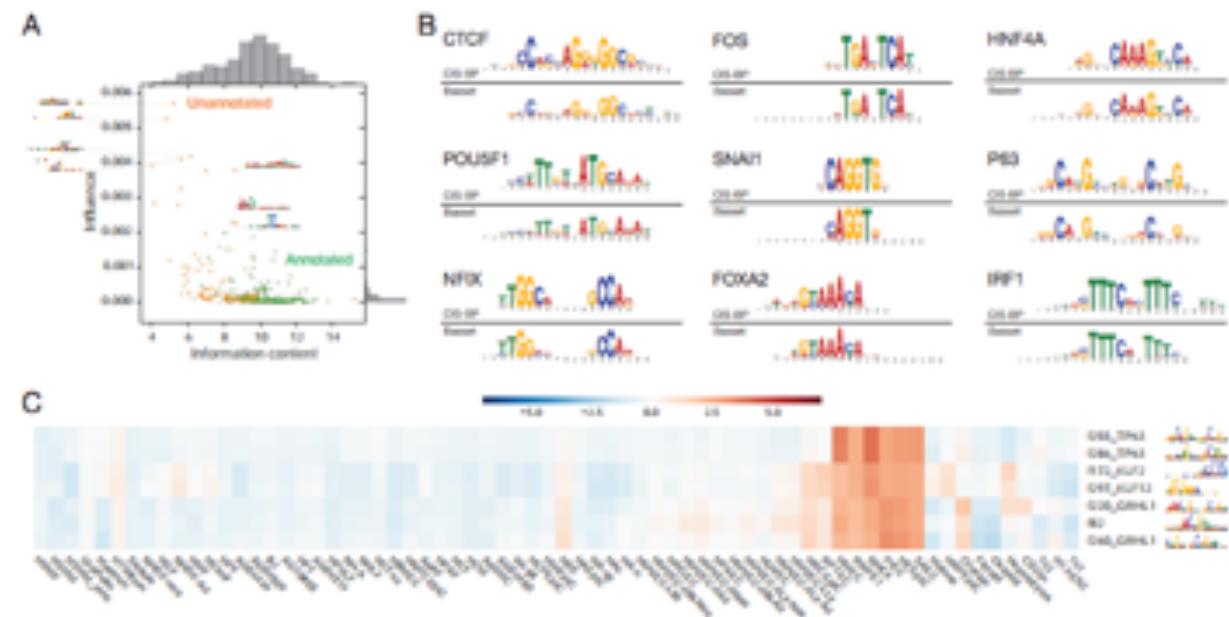
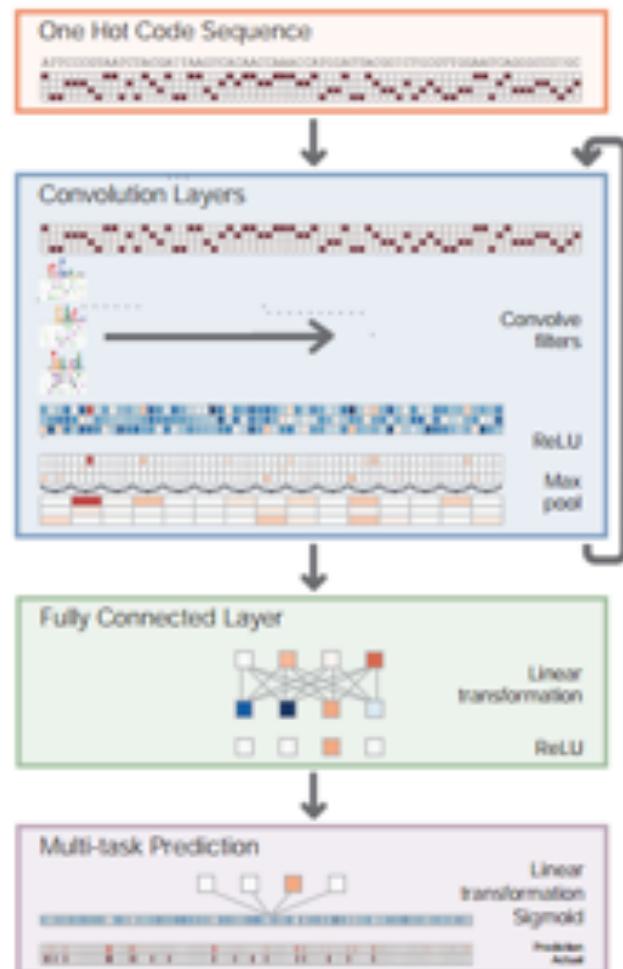
4. Second pass: scoring enriched target regions relative to control

- 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

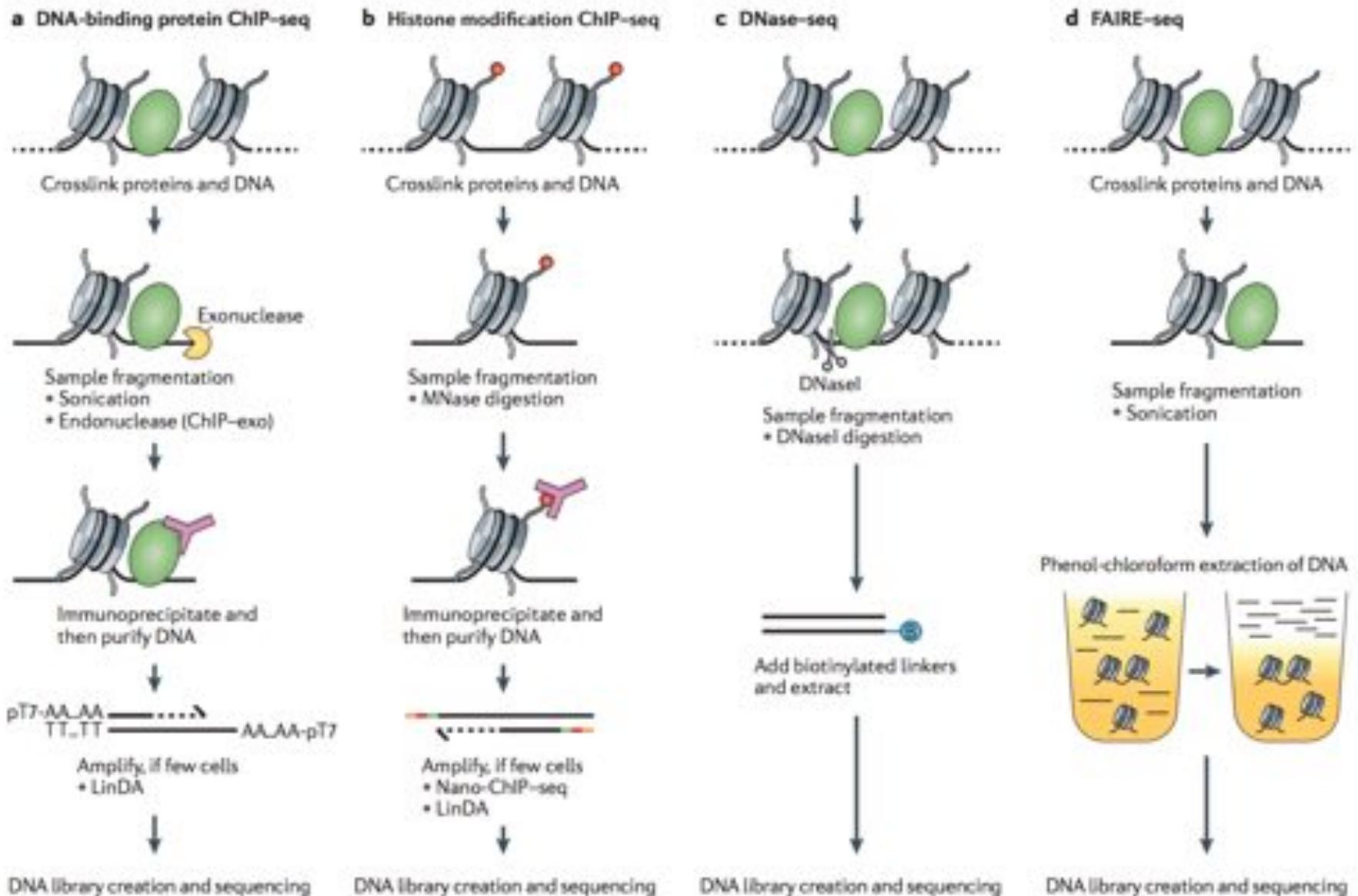
Rozowsky et al (2009) Nature Biotechnology 27, 66 - 75

Basset



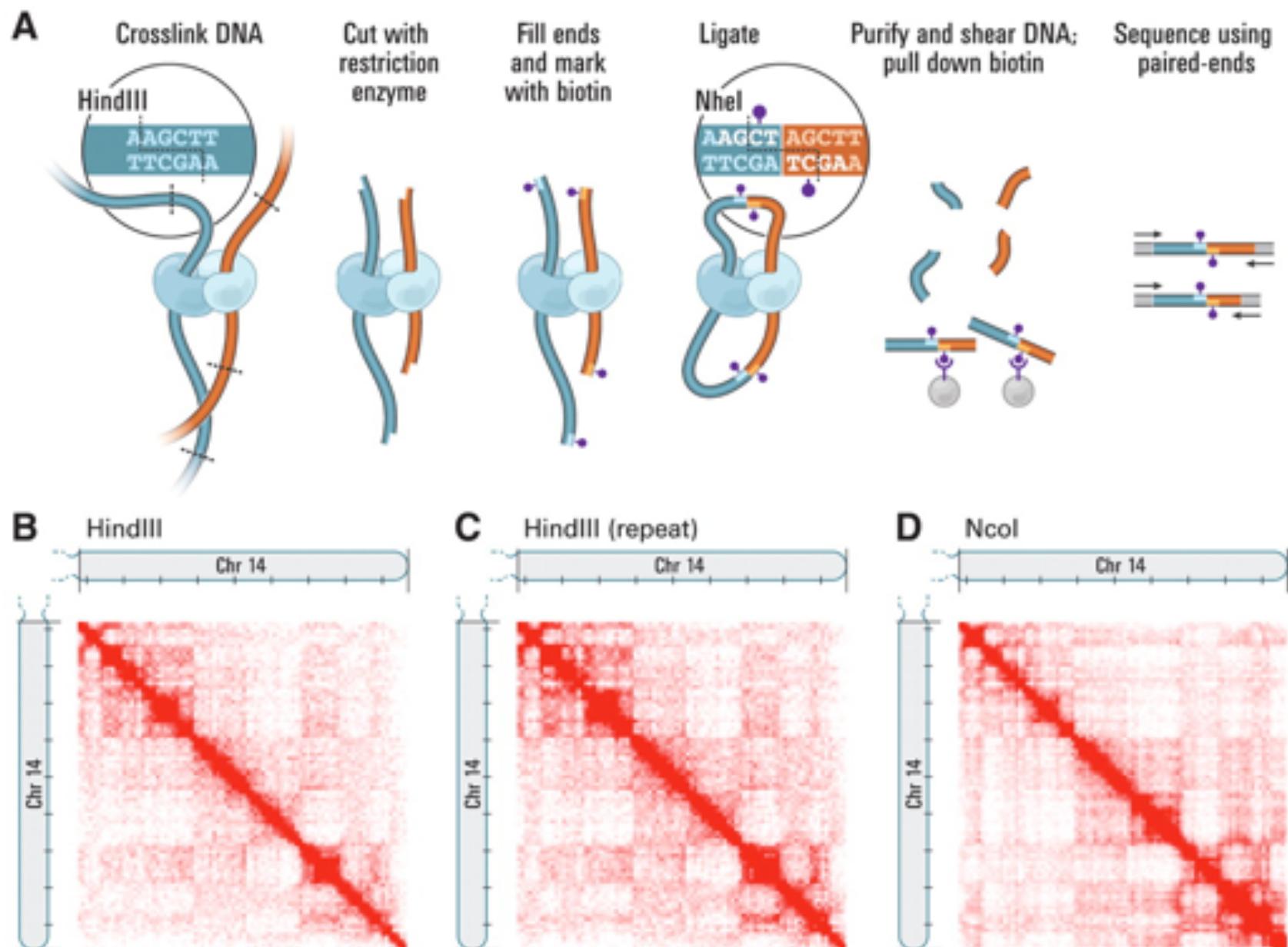
Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks
Kelley et al. (2016) Genome Research doi: 10.1101/gr.200535.115

Related Assays



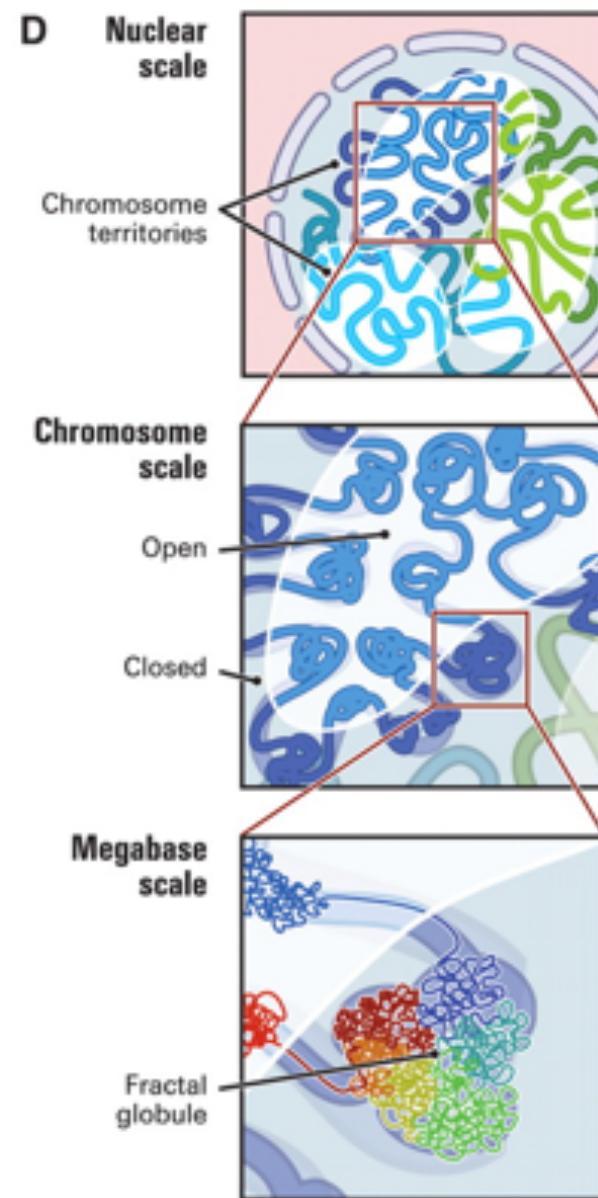
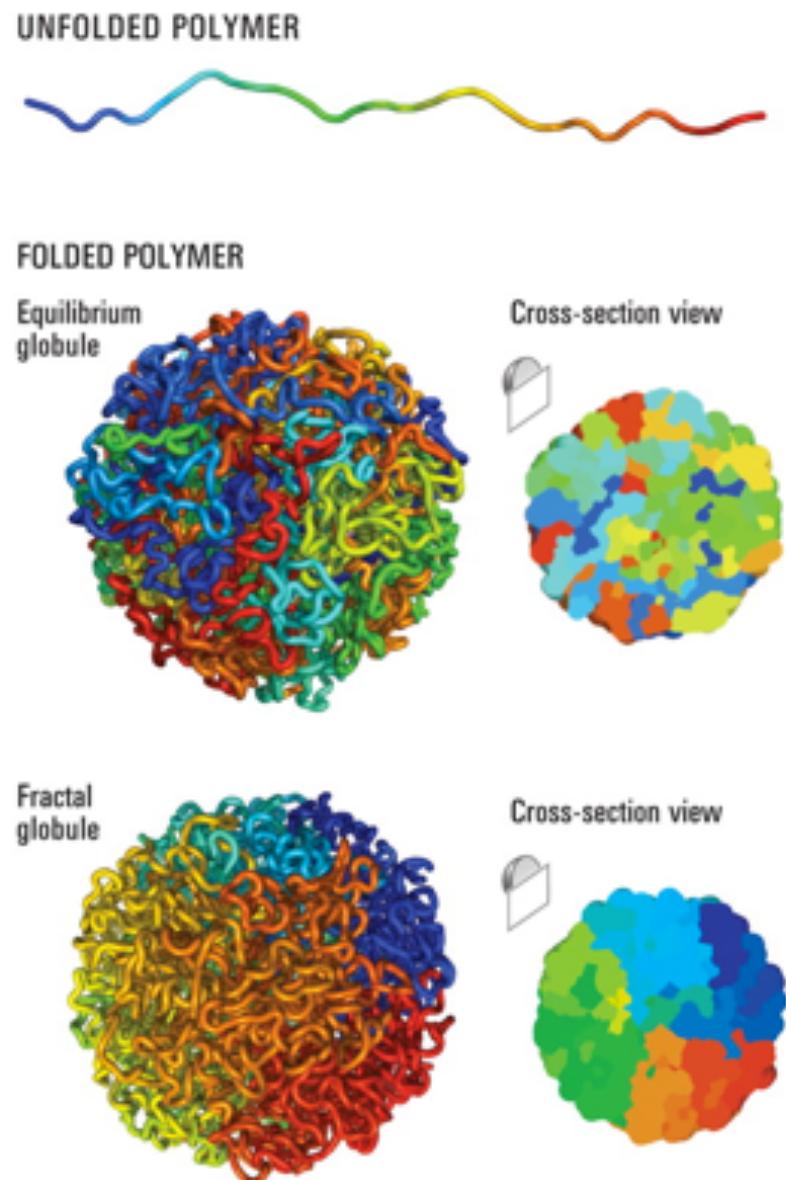
ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions
Furey (2012) *Nature Reviews Genetics*. 13, 840-852

HI-C: Mapping the folding of DNA



Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome
Lieberman-Aiden et al. (2009) Science. 326 (5950): 289-293

Hi-C: Mapping the folding of DNA



Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome
Lieberman-Aiden et al. (2009) Science. 326 (5950): 289-293

Gene Regulation in 3-dimensions

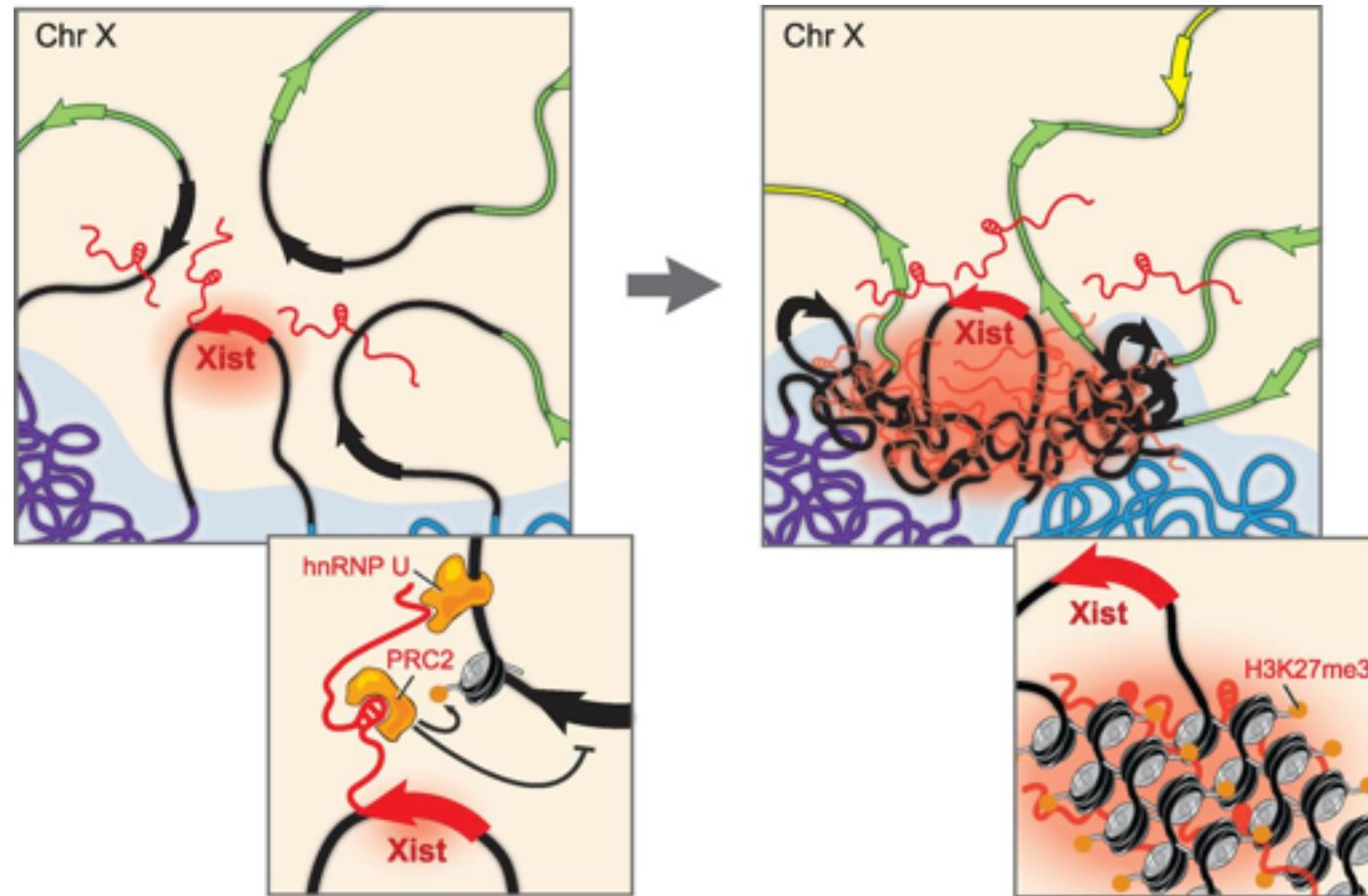
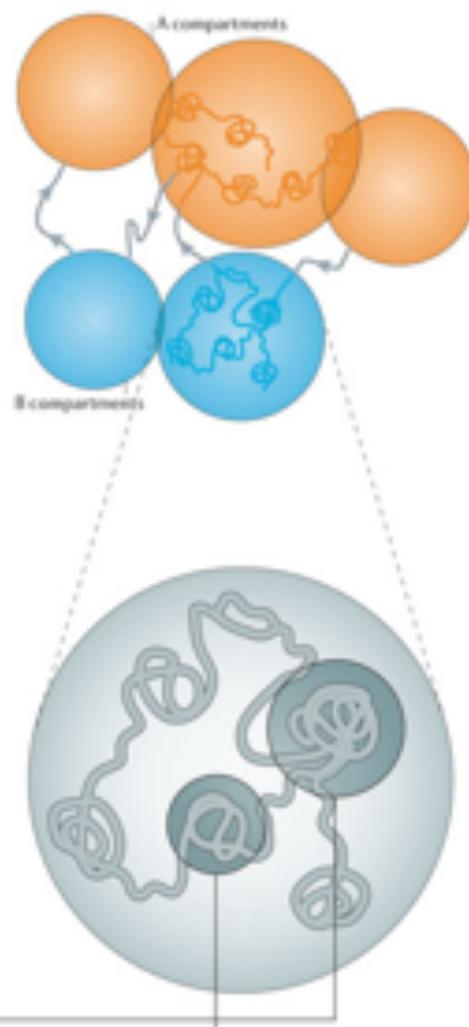
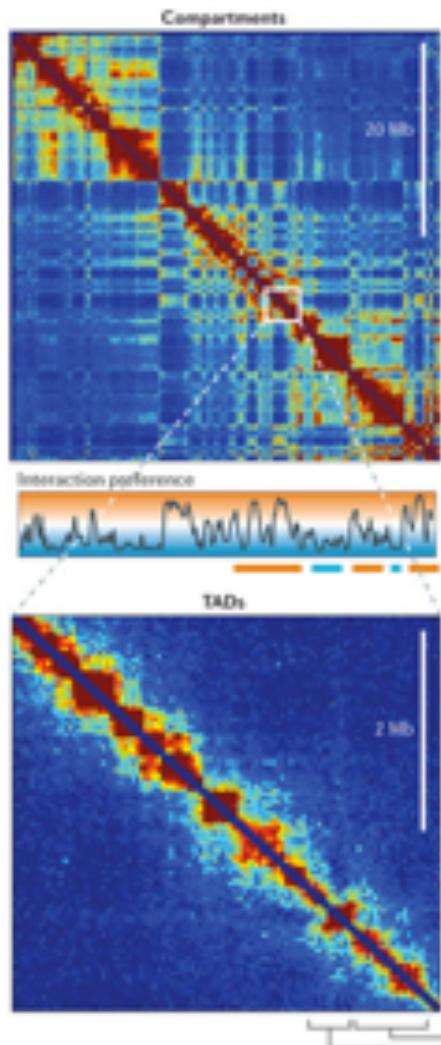


Fig 6. A model for how *Xist* exploits and alters three-dimensional genome architecture to spread across the *X* chromosome.

The *Xist* lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the *X* Chromosome
Engreitz et al. (2013) Science. 341 (6147)

Genome compartments & TADs



Nature Reviews | Genetics

Mammalian genomes have a pattern of interactions that can be approximated by two compartments called A and B

- alternate along chromosomes and have a characteristic size of ~5 Mb each.
- A compartments (orange) preferentially interact with other A compartments; B compartments (blue) associate with other B compartments.
- A compartments are largely euchromatic, transcriptionally active regions.

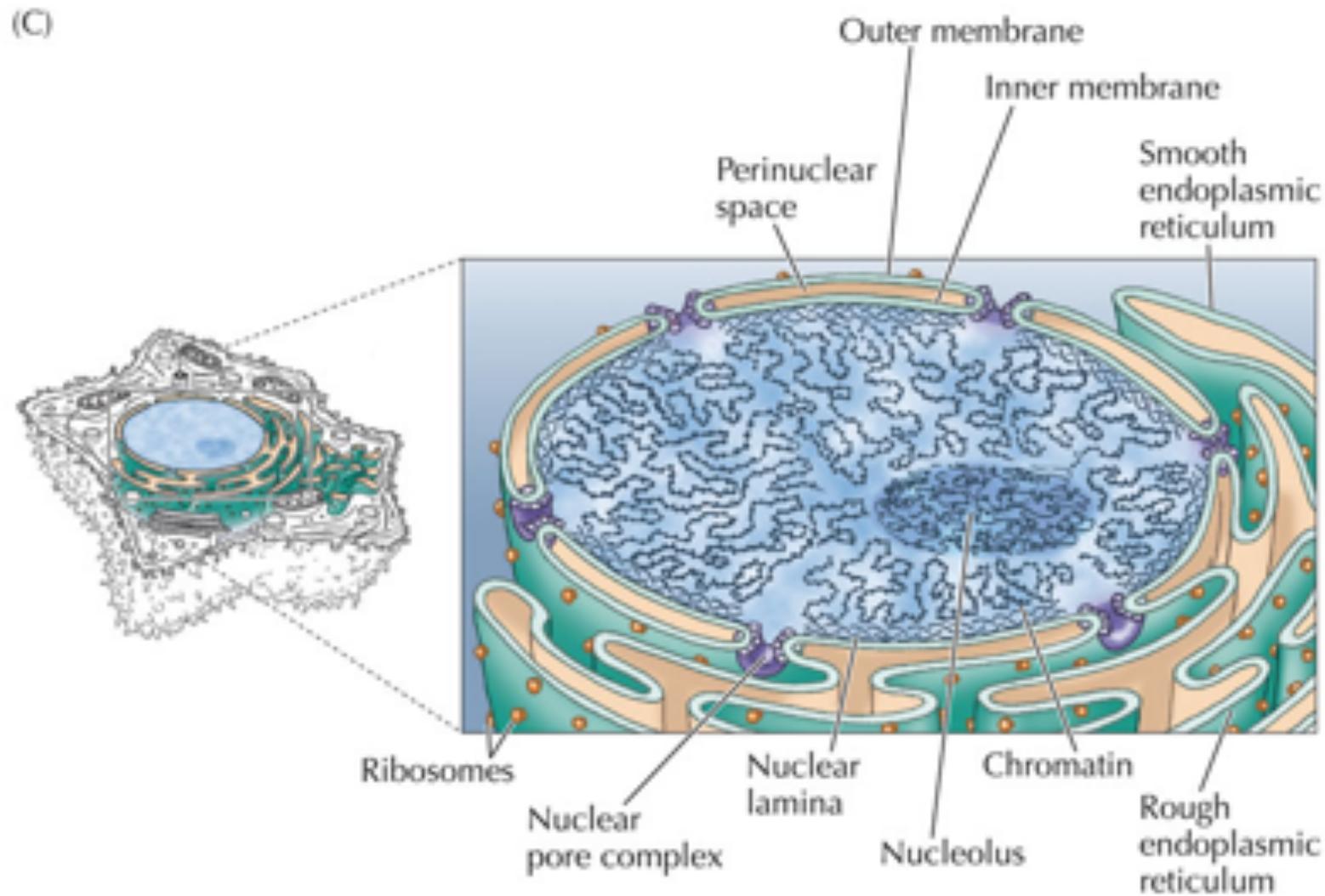
Topologically associating domains (TADs)

- TADs are smaller (~400–500 kb)
- Can be active or inactive, and adjacent TADs are not necessarily of opposite chromatin status.
- TADs are hard-wired features of chromosomes, and groups of adjacent TADs can organize in A and B compartments

Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data

Dekker et al. (2013) *Nature Reviews Genetics* 14, 390–403

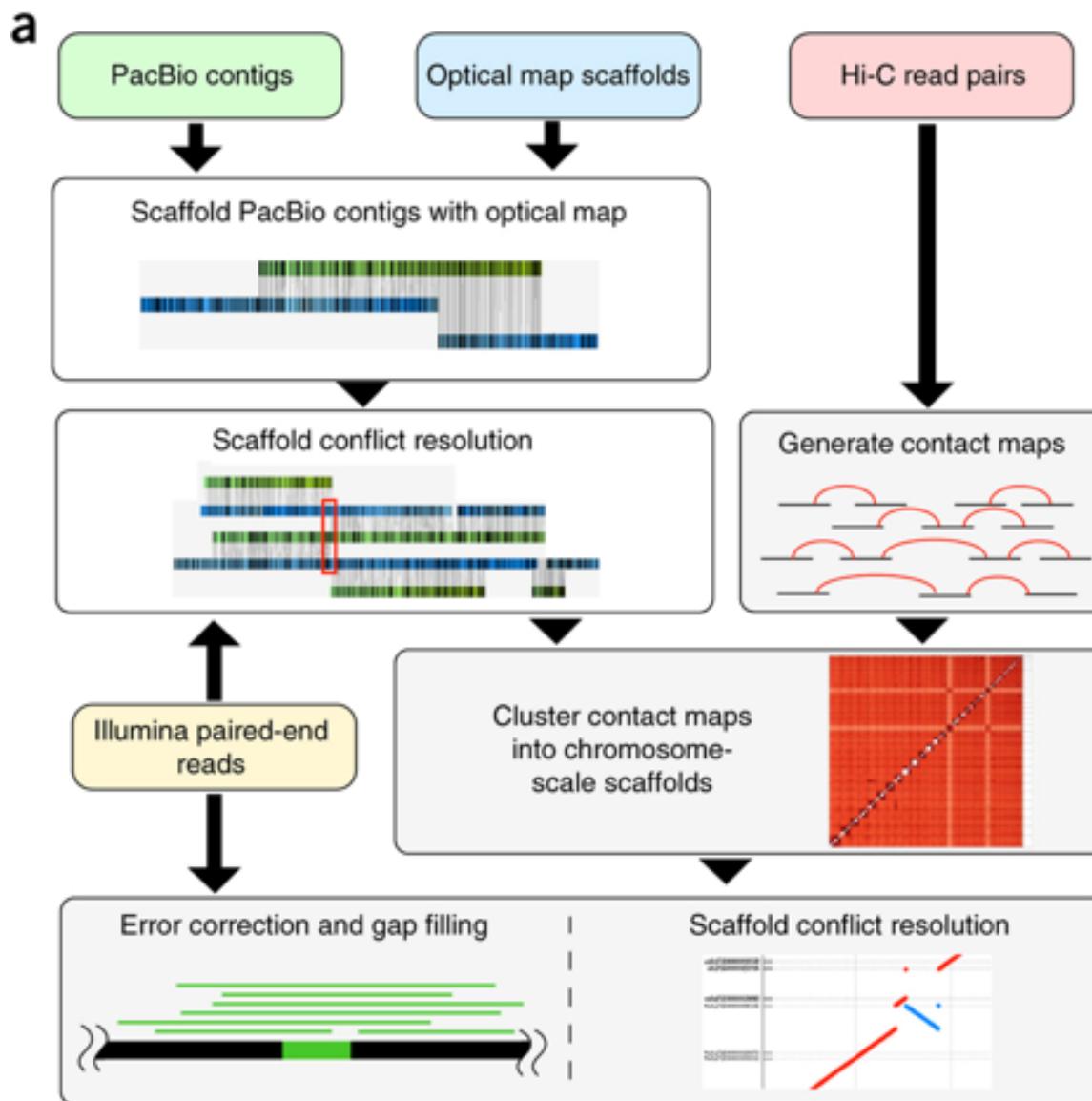
“Lamina-Associated Domains are the B compartment”



THE CELL, Fourth Edition, Figure 9.1 (Part B) © 2006 ASM Press and Sinauer Associates, Inc.

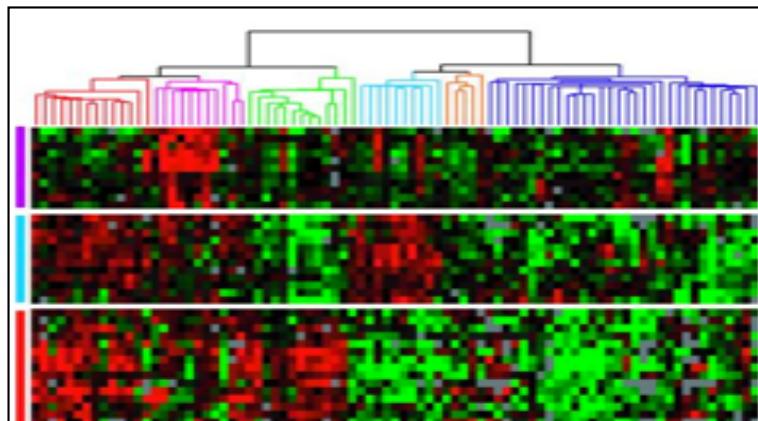
Chromosome Conformation Paints Reveal the Role of Lamina Association in Genome Organization and Regulation
Luperchio et al. (2017) bioRxiv. doi: <https://doi.org/10.1101/122226>

Scaffolding with Hi-C

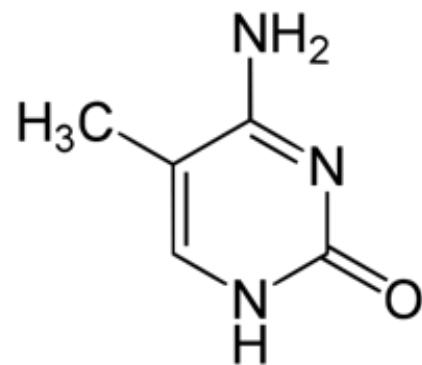


Putting it all together!

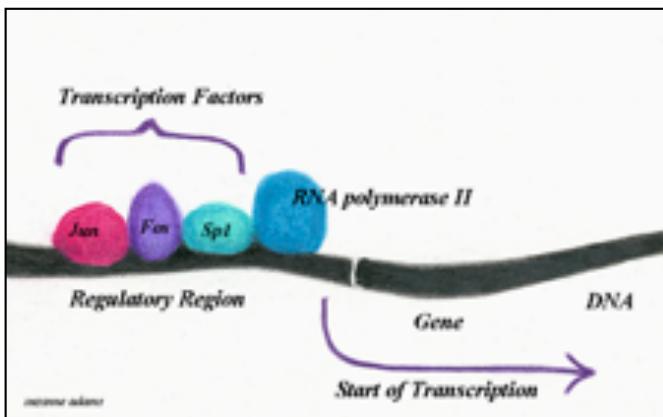
RNA-seq



Methyl-seq



ChIP-seq



Hi-C

