SUNY Korea
The State University of New York
한국뉴욕주립대학교

Stony Brook University

Instructor: Sael Lee

CS549 Spring – Computational Biology

LECTURE 6:
FINDING NUCLEOSOME POSITIONS

Reference:
C. Jiang and B. F. Pugh. Nucleosome positioning and gene regulation: advances through genomics.
Nature Reviews Genetics 10 161-172 (2009)

YOUTUBE VIDEOS

How DNA is Packaged

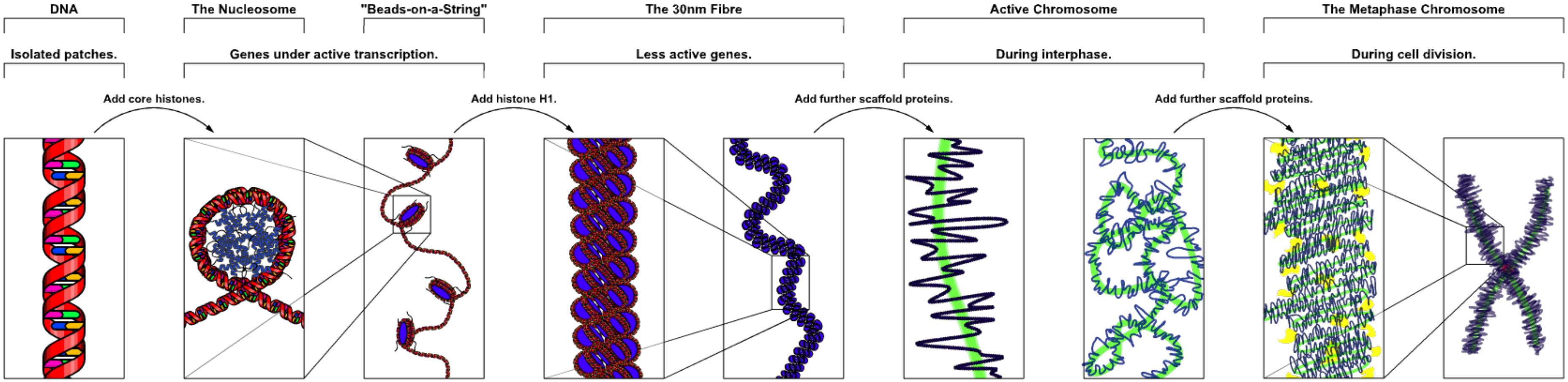
Chromatin, Histones and Modifications, Rate My Science

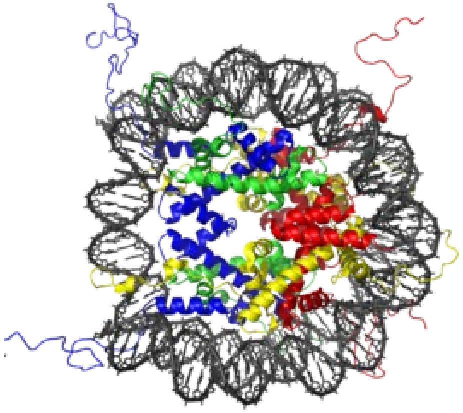
Epigenetics Tutorial - advanced

WHY NUCLEOSOMES POSITION?

- Knowing the precise locations of nucleosomes in a genome is key to understanding how genes are regulated.
- Nucleosome positions can tell us about
  - How nucleosome positioning distinguish promoter regions and transcriptional start sites, and
  - How the composition and structure of promoter nucleosomes facilitate or inhibit transcription.
  - How diverse factors, including underlying DNA sequences and chromatin remodeling complexes, influence nucleosome positioning

CHROMATIN STRUCTURES

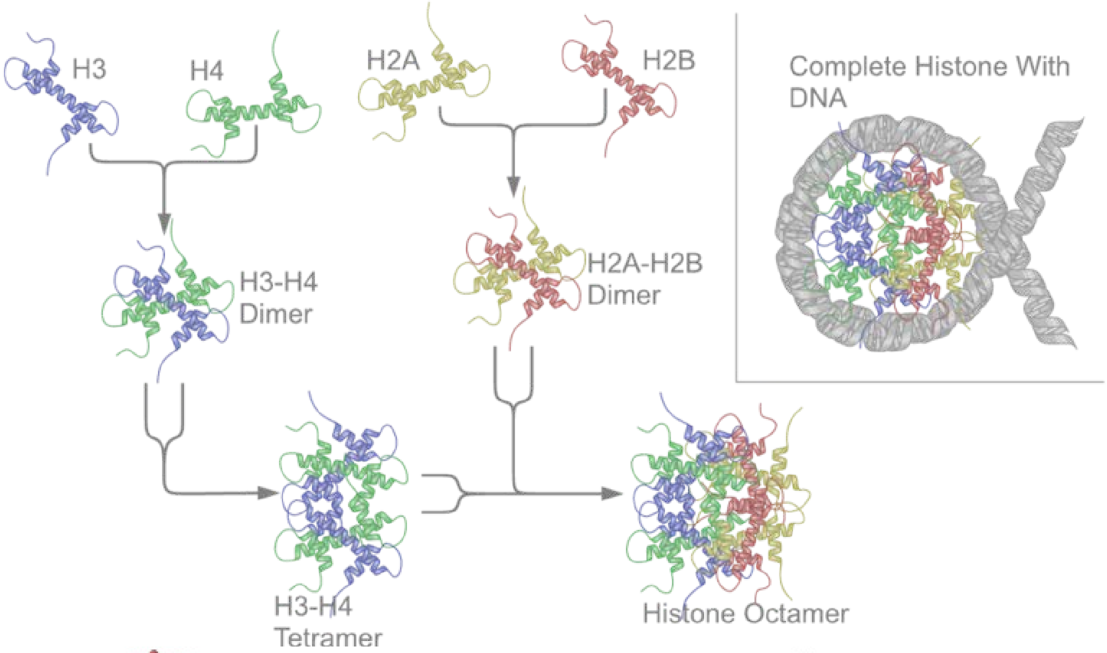


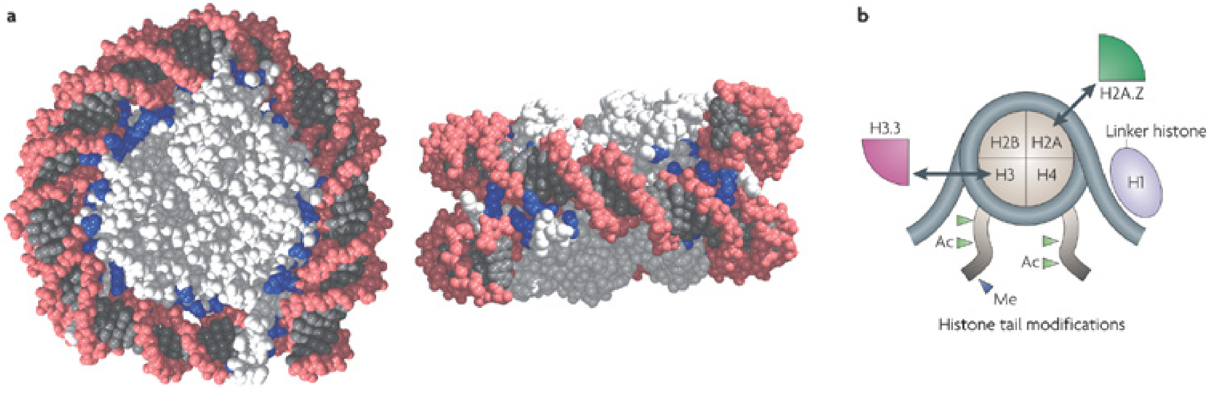


The packaging of DNA creates both a problem and an opportunity:

- Wrapping DNA around histones may be a obstacle in accessing the genetic code;
- Can be exploited so that enzymes that read, replicate and repair DNA can be directed to the appropriate entry sites

NUCLEOSOME STRUCTURE





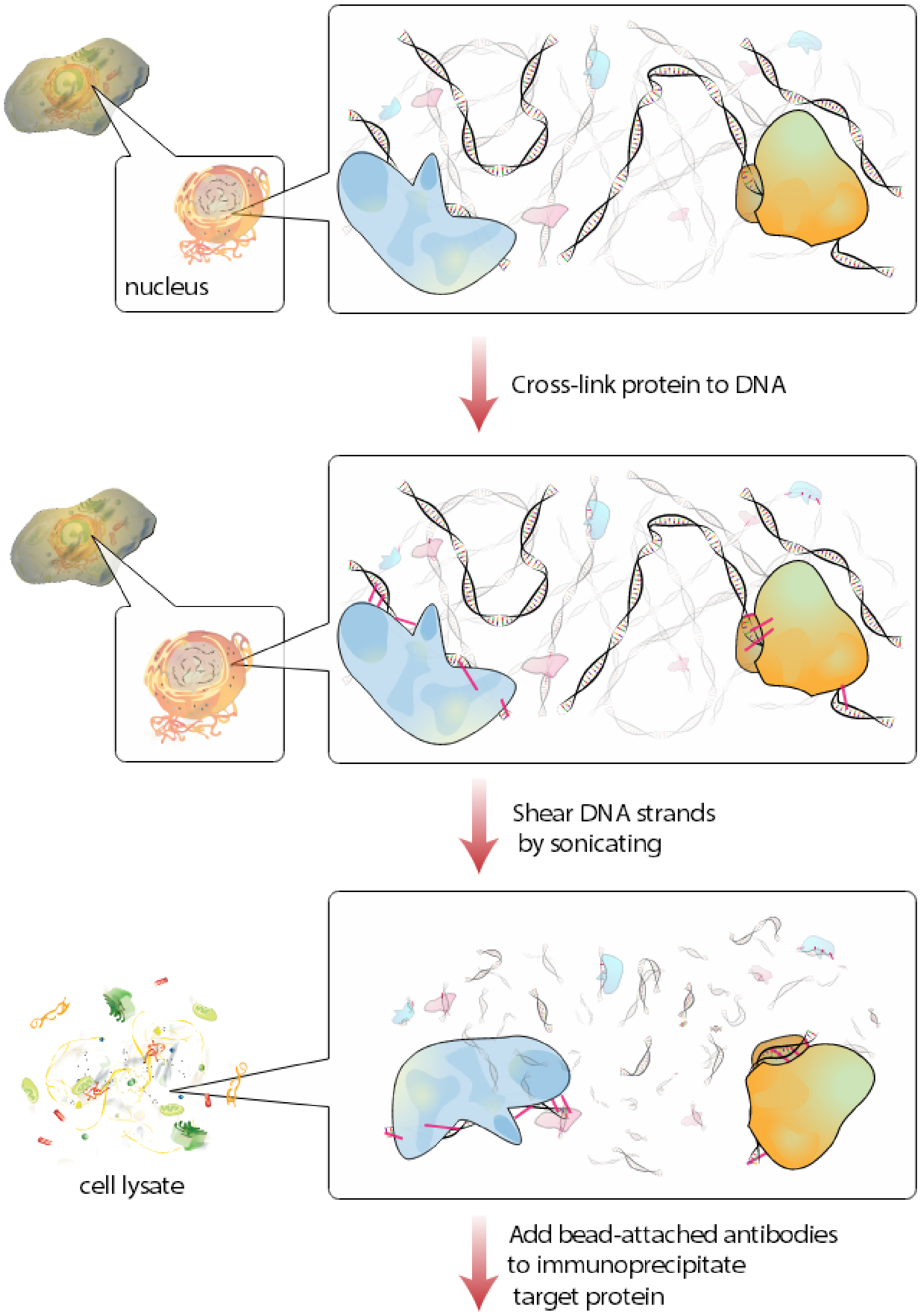
Nature Reviews | Genetics

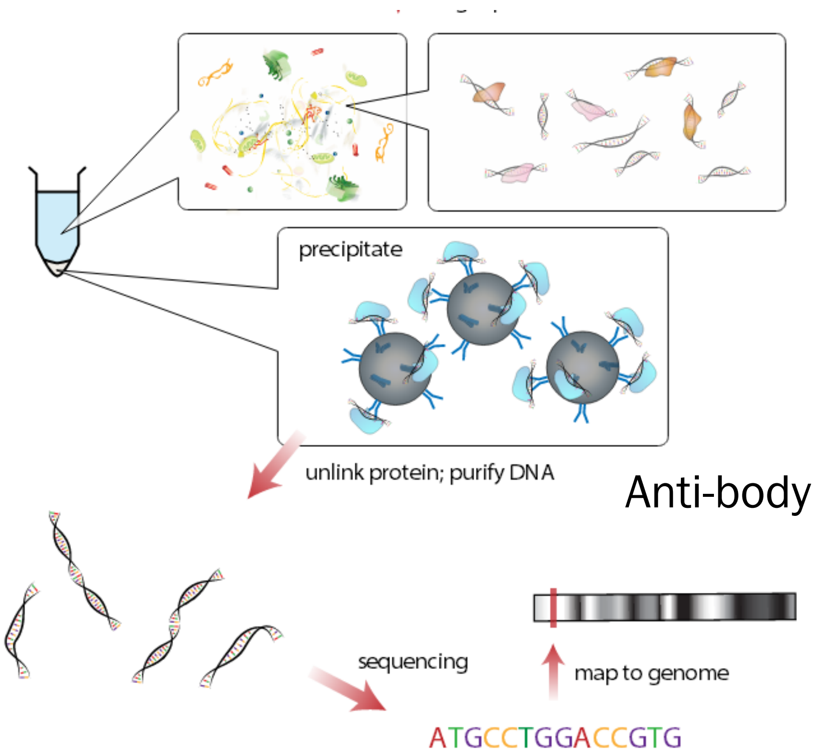
The **nucleosome** is the basic unit of eukaryotic chromatin, consisting of a **histone** core around DNA.

Each histone core is composed of two copies of each of the histone proteins H2A, H2B, H3 and H4. Approximately **147 bp** of DNA coils **1.65 times** around the histone octamer in a left-handed toroid.

CHROMATIN IMMUNOPRECIPITATION SEQUENCING (CHIP-SEQ) WORK FLOW

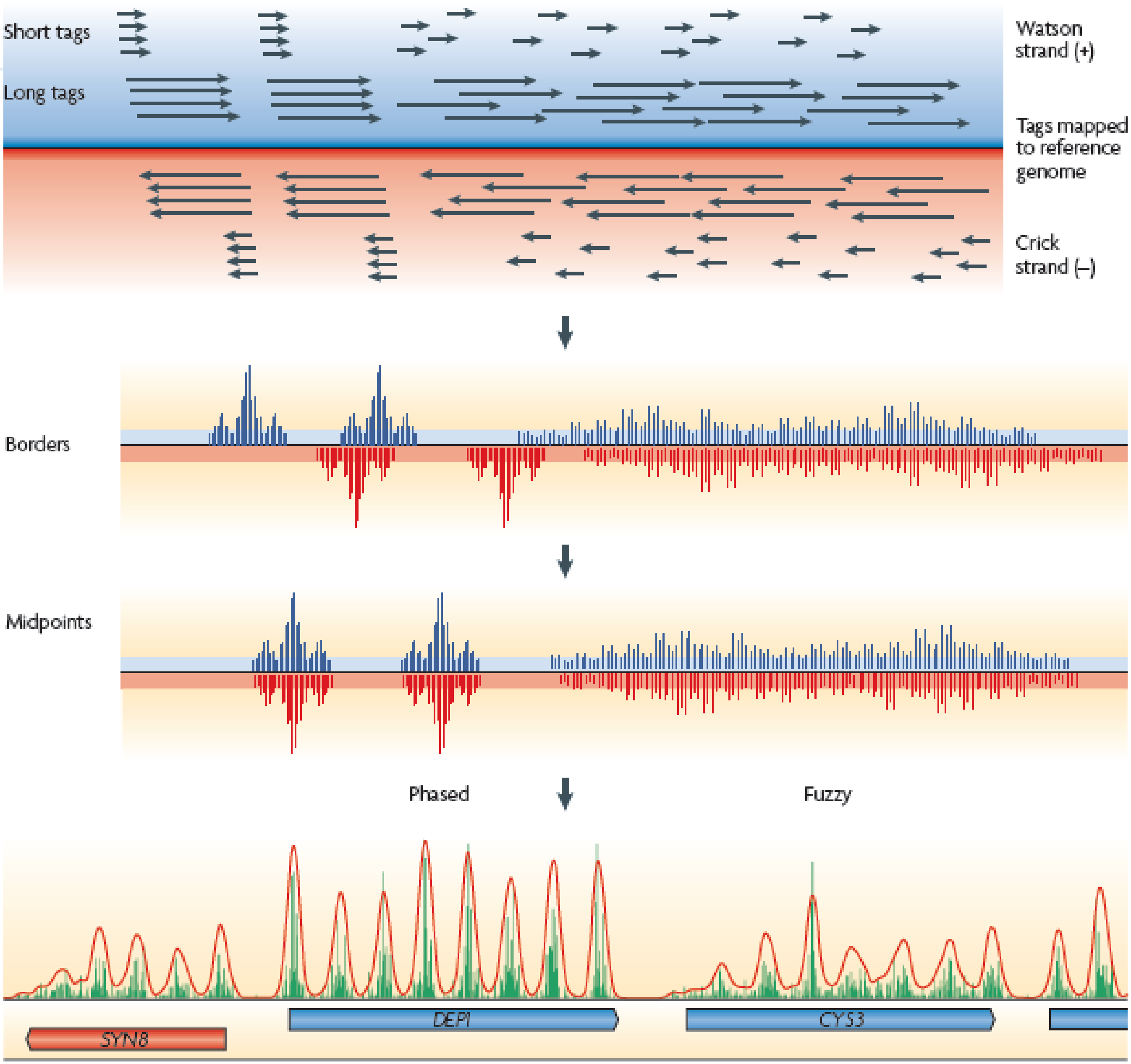
ChIP-seq experiments introduced in 2007


nucleus
Cross-link protein to DNA
Shear DNA strands by sonicating
cell lysate
Add bead-attached antibodies to immunoprecipitate target protein



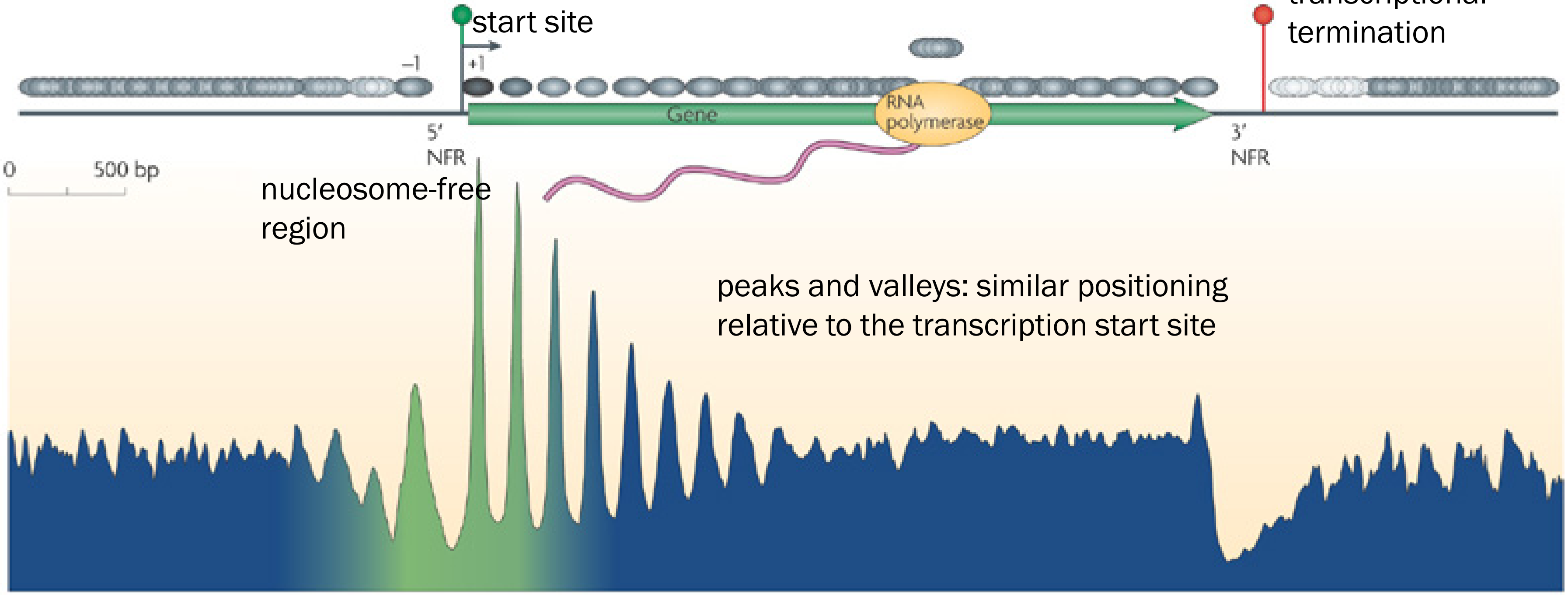
Anti-body

Box 1 | **ChIP–Seq nucleosome mapping technology**



NUCLEOSOMAL LANDSCAPE OF YEAST GENES.

Nucleosome maps of a similar resolution in yeast, worms, flies, humans, etc. have been published



transcription start site

transcriptional termination

nucleosome-free region

peaks and valleys: similar positioning relative to the transcription start site

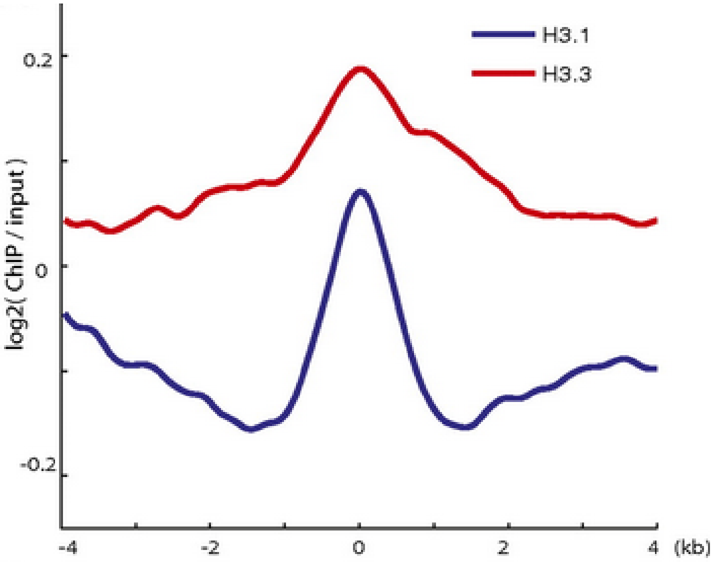
Nature Reviews | Genetics

Nucleosomes generally adopt canonical positions around promoter regions and more random positions in the interior of genes.

GENOMEWIDE NUCLEOSOME MAPS

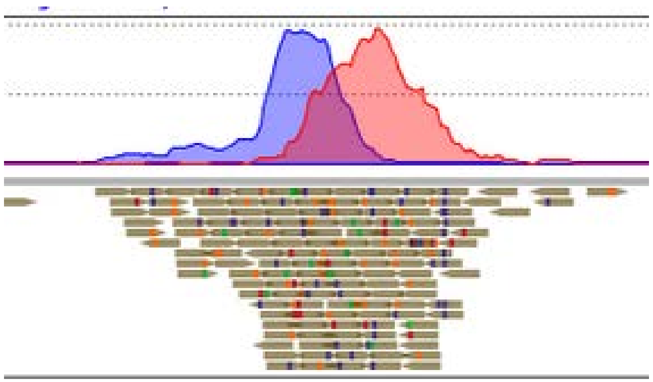
Allow us to explore the genomic properties of chromatin

At most loci, there is an approximately **Gaussian (normal) distribution** of nucleosome positions around particular genomic coordinates, ranging from ~30 bp for highly phased nucleosomes to a random continuous distribution throughout an array.



Cause of variation:

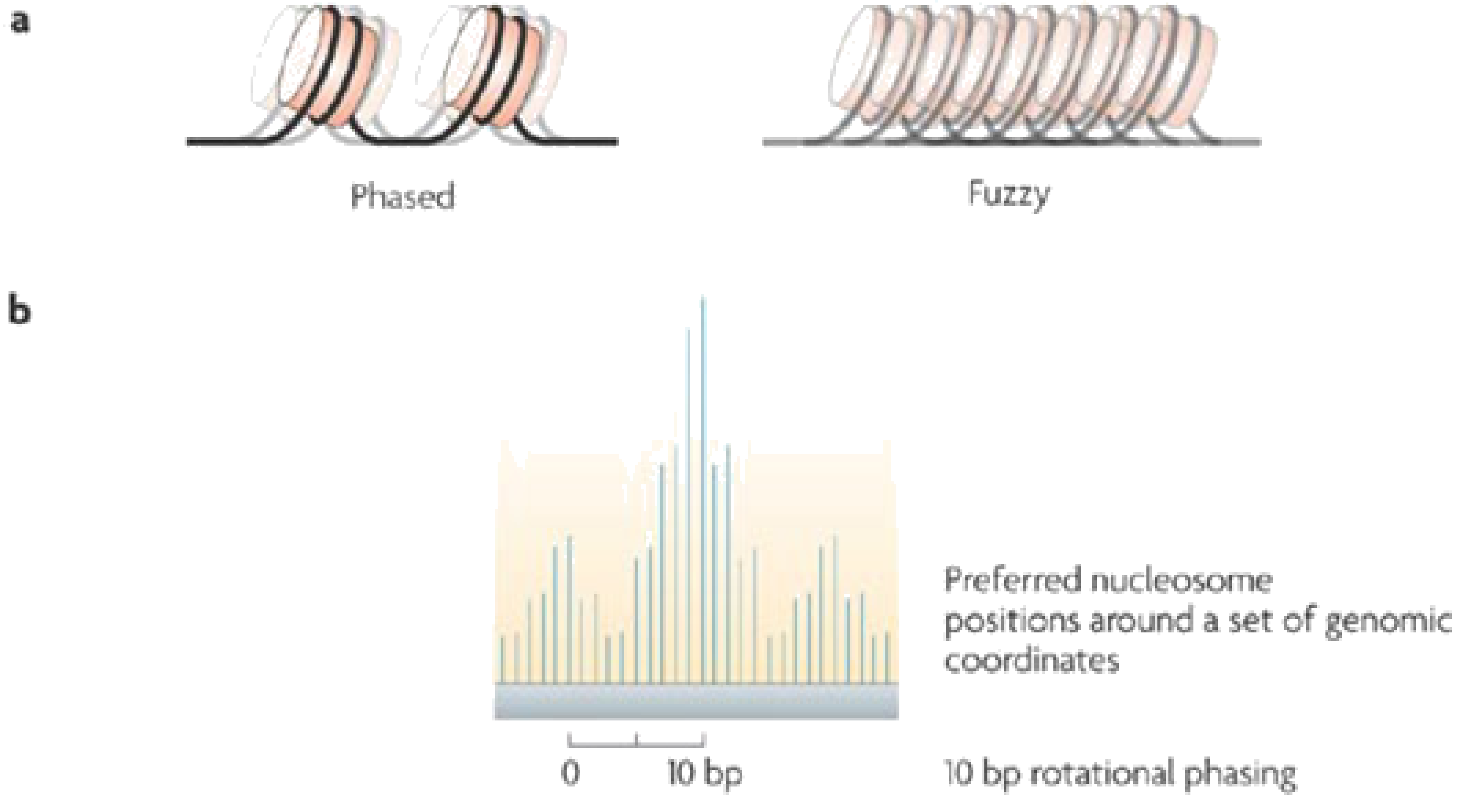
- Genuine positional heterogeneity
- how much is an artifact that is caused by overtrimming or undertrimming of the DNA at nucleosome borders by experiment



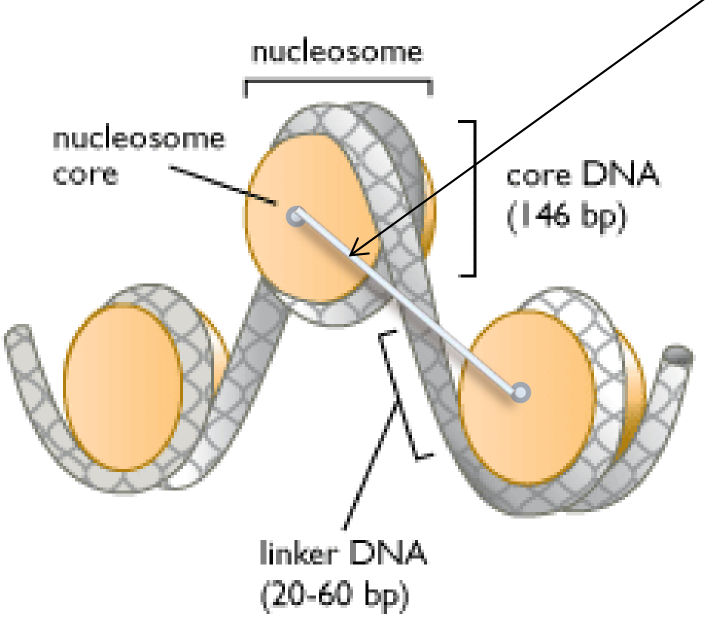
***Phasing**

The distribution of nucleosomes around a particular coordinate in a population of cells.

PHASING INFORMATION AND ROTATIONAL SETTING


a
Phased
Fuzzy
b
Preferred nucleosome positions around a set of genomic coordinates
0
10 bp
10 bp rotational phasing

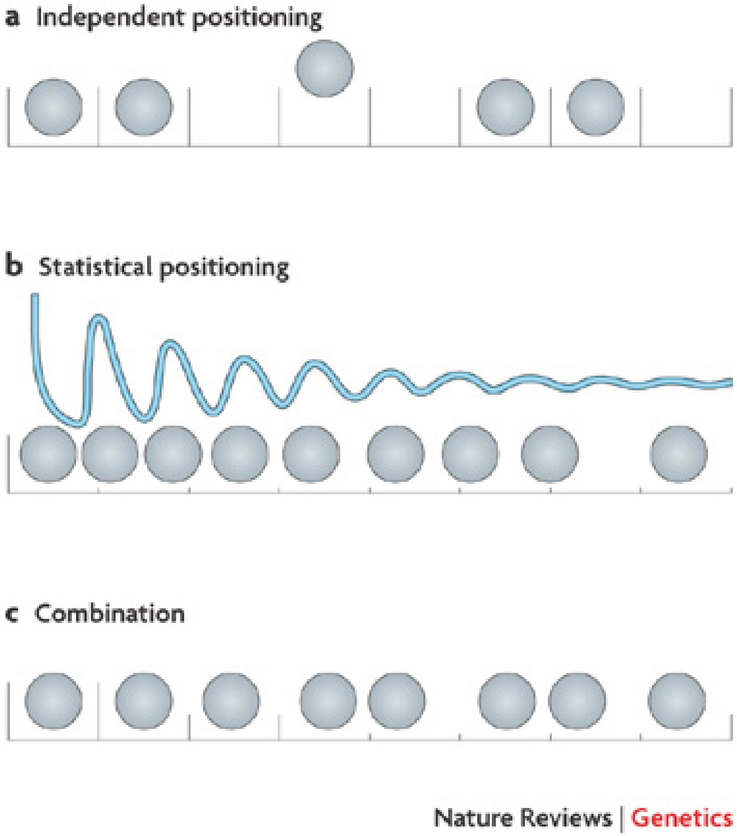
a | In a population, individual nucleosomes are either positioned within a small range of a genomic locus (phased) or with a continuous distribution throughout an array (fuzzy).
b | The bar graph is an idealized distribution of nucleosomal sequence tags, which form a large cluster and several subclusters, in which the subclusters are spaced about 10 bp apart and represent multiple translational settings with a single predominant rotational setting.



Typical distance between adjacent nucleosome midpoints is approximately

- 165 bp (~18 bp linker) in *S. cerevisiae (*yeast*)*
- 175 bp (~28 bp linker) in Drosophila *melanogaster* & *C. elegans*
- 185 bp (~38 bp linker) in humans

SEQUENCE-BASED PACKING VERSUS STATISTICAL PACKING.


a Independent positioning
b Statistical positioning
c Combination
Nature Reviews | Genetics

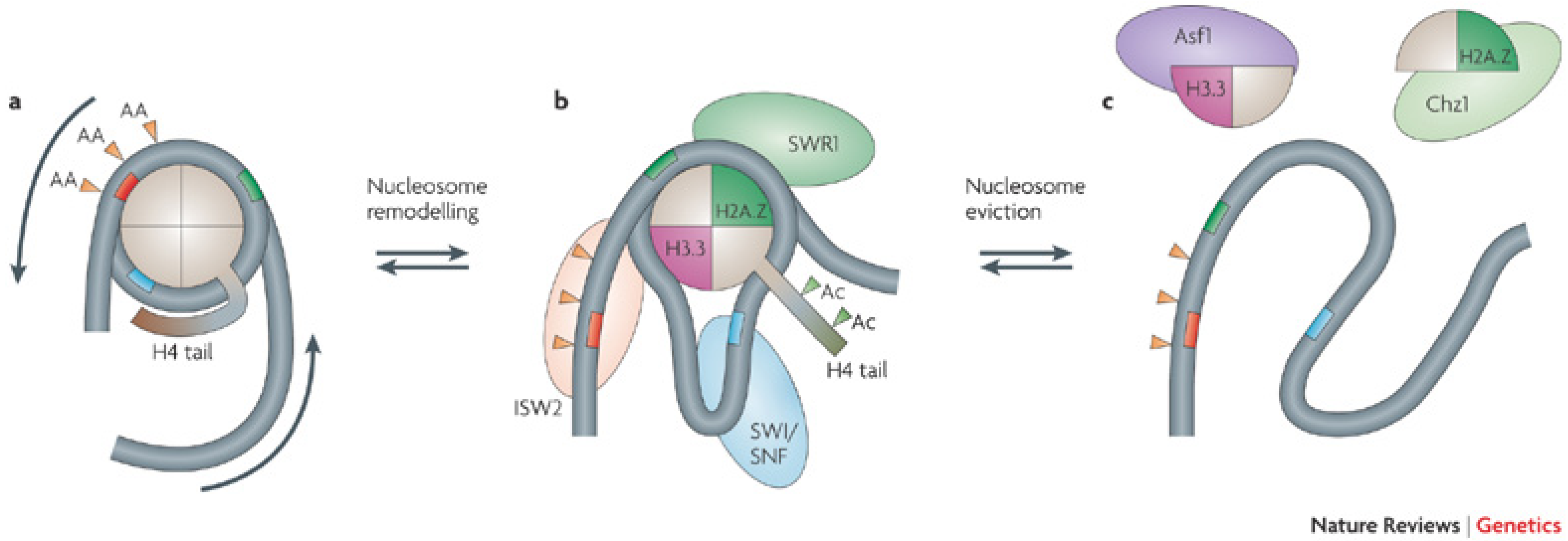
Models about origin of nucleosome positions:

Independent positioning: that the positions of adjacent nucleosomes are independently controlled depending on the binding affinity with the underlying DNA sequence.

Statistical positioning: positioning of one nucleosome in the array forces the positioning of all other nucleosomes, because the tight packing restricts their lateral movement.

a | Individual slots represent nucleosome positioning sequences that define where a nucleosome (grey circle) will reside on a length of DNA. **b** | In its purest form, statistical positioning relies on a single positional barrier (left side), against which nucleosomes are ordered. A probabilistic density trace of where nucleosomes would reside in a population is shown. **c** | The true cellular state is likely to be a combination of both independent and statistical positioning.

MECHANISMS THAT ALLOW DNA ACCESSIBILITY.



a | A stable nucleosome. **b** | A remodelled nucleosome. **c** | An evicted nucleosome. Three transcription factor binding sites are shown in red, green and blue, respectively. The red and blue sites become accessible only during remodelling, either by nucleosome sliding, as indicated by the arrows in **a**, or by chromatin remodelling complexes (for example, ISW2, SWR1 and SWI/SNF) that 'extract' DNA from the nucleosome surface, as shown in **b**. Owing to rotational phasing, the green site is always accessible in the various states. Nucleosome eviction (**c**) might be necessary to assemble a pre-initiation complex and to transcribe the underlying DNA. Anti-silencing function 1 (Asf1) and H2A.Z-specific chaperone (Chz1) are examples of histone chaperones. Ac, acetylation.

MIXTURE MODELS: INTRODUCTION

THE DENSITY ESTIMATION PROBLEM

Density Estimation Problem: (loose definition)

Given a set of N points in D dimensions, $\mathrm{x}_1, \ldots, \mathrm{x}_N \in R^D$, and a family F of probability density function on R^D, find the probability density functions (pdf) on R^D, find pdf $f(\mathrm{x}) \in F$ that is most likely to have generated the given points.

Defining F : give each of it's members the same mathematical form, and to distinguish different members by different values of a set of parameters θ.

EX> Mixture of PDFs

$$f(\mathbf{x};\theta) = \sum_{k=1}^{K} \pi_k g(\mathbf{x};\theta_k)$$

$$\int g(\mathbf{x};\theta_k)d\mathbf{x} = 1$$

PDF

$$\int f(\mathbf{x};\ \theta)d\mathbf{x} = 1$$

Mixture of PDFs

$$\sum_{k=1}^{K} \pi_k \ = 1; \quad \pi_k > 0$$

Mixing probability

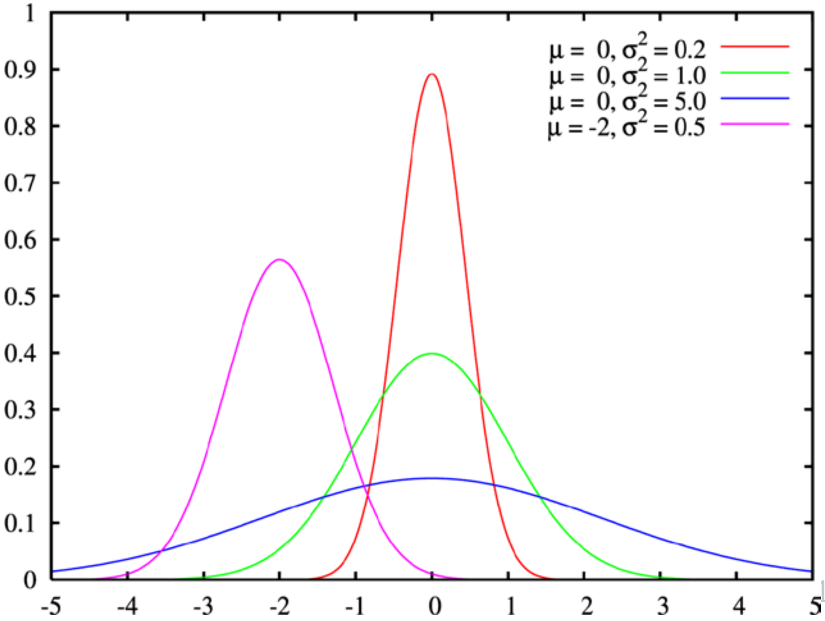
MIXTURE MODEL AND CLUSTERING

Example: Gaussian Mixture Models.

$$\sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$$

$$N(\mathbf{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) = \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma_k}|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu)^T \boldsymbol{\Sigma^{-1}}(x-\mu) \quad)$$

Each cluster is assigned a Gaussian, with **mean** being the center of cluster and **standard deviation** being the spread of data for the cluster.



LIKELIHOOD FUNCTION IN MIXTURE MODEL

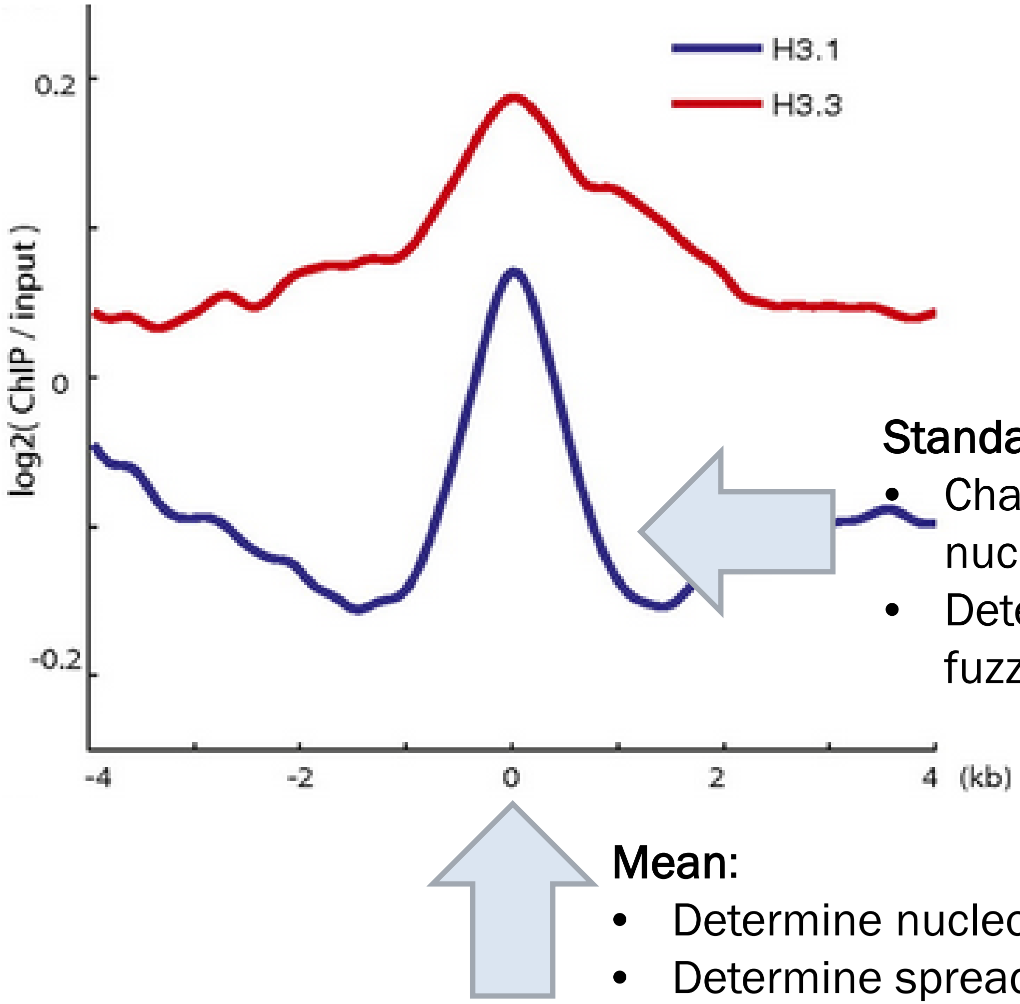
Density estimations in other words: **finding the parameters $\boldsymbol{\theta}$** that specifies the model from which the points are most likely to be drawn.

Estimation of parameters can be done by solving for **maximum likelihood parameter** that explains the **data X** the best.

We will talk about likelihood function formation next class.

We will talk about maximum likely estimate with **E**xpectation **M**aximization.

GAUSSIAN MIXTURE MODEL AND NUCLEOSOME POSITION



Standard deviation:

- Characterize nucleosome stability
- Determine phased or fuzzy.

Mean:

- Determine nucleosome center position
- Determine spread of nucleosome