



# Школа «Науки о данных»

Трек «Биоинформатика» 25-26

07.09.2025



# UCSC Genome Browser

UCSC Genome Browser Home

http://genome.ucsc.edu/

## UCSC Genome Bioinformatics

[Genomes](#) - [Blat](#) - [Tables](#) - [Gene Sorter](#) - [PCR](#) - [VisiGene](#) - [Proteome](#) - [Session](#) - [FAQ](#) - [Help](#)

[Genome Browser](#)

[ENCODE](#)

[Neandertal](#)

[Blat](#)

[Table Browser](#)

[Gene Sorter](#)

[In Silico PCR](#)

[Genome Graphs](#)

[Galaxy](#)

[VisiGene](#)

[Proteome Browser](#)

[Utilities](#)

[Downloads](#)

[Release Log](#)

### About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to the [ENCODE](#) and [Neandertal](#) projects.

We encourage you to explore these sequences with our tools. The [Genome Browser](#) zooms and scrolls over chromosomes, showing the work of annotators worldwide. The [Gene Sorter](#) shows expression, homology and other information on groups of genes that can be related in many ways. [Blat](#) quickly maps your sequence to the genome. The [Table Browser](#) provides convenient access to the underlying database. [VisiGene](#) lets you browse through a large collection of *in situ* mouse and frog images to examine expression patterns. [Genome Graphs](#) allows you to upload and display genome-wide data sets.

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the Center for Biomolecular Science and Engineering ([CBSE](#)) at the University of California Santa Cruz ([UCSC](#)). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our [public mailing list](#).

### News

[News Archives](#) ▶

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the [genome-announce](#) mailing list.

**10 February 2012 - dbSNP 135 Available for hg19**

We are pleased to announce the release of four tracks derived from dbSNP build 135, available on the human assembly (GRCh37/hg19). dbSNP build 135 is available at NCBI. The new tracks contain additional annotation data not included in previous dbSNP tracks, with corresponding coloring and filtering options in the Genome Browser.

As for dbSNP build 132, there are four tracks in this release. One is a track containing all mappings of reference SNPs to the human assembly, labeled "All SNPs (135)". The other three tracks are subsets of this track and show interesting and easily defined subsets of dbSNP:

Done

# UCSC Genome Browser

Human (Homo sapiens) Genome Browser Gateway

http://genome.ucsc.edu/cgi-bin/hgGateway

Home Genomes Blat Tables Gene Sorter PCR Session FAQ Help

## Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the [Genome Bioinformatics Group of UC Santa Cruz](#).  
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade genome assembly position or search term [gene](#)

Mammal Human Feb. 2009 (GRCh37/hg19) chr21:33,031,597-33,041,570 submit

Mammal Human Feb. 2009 (GRCh37/hg19) submit  
Vertebrate Mar. 2006 (NCBI36/hg18)  
Deuterostome Gorilla May 2004 (NCBI35/hg17)  
Insect Orangutan July 2003 (NCBI34/hg16)  
Nematode Gibbon  
Other Rhesus  
Marmoset  
Mouse  
Rat  
Guinea pig  
Rabbit  
Pig  
Sheep  
Cow  
Horse  
Cat  
Dog  
Panda  
Microbat  
Elephant

interface settings to their defaults.  
configure tracks and display clear position

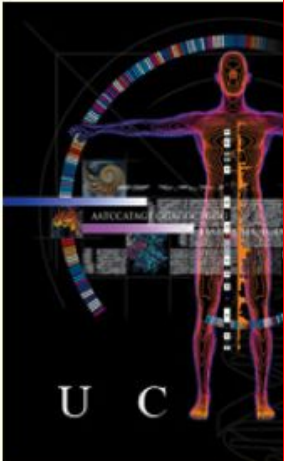
### About the Human Feb. 2009 (GRCh37) assembly (sequences)

The February 2009 human reference sequence (GRCh37) was produced by the [Genome Reference Consortium](#). For more information about the assembly, see [GRCh37](#) in the NCBI Assembly database.

### Sample position queries

A genome position can be specified by a chromosome number, a number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal band, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the [User's Guide](#) for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_g1000212	Displays all of the unplaced contig g1000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000



- ATP
- ATP10A
- ATP10B
- ATP10D
- ATP11A
- ATP11B
- ATP11C
- ATP12A
- ATP13A1
- ATP13A2
- ATP13A3
- ATP13A4
- ATP13A5
- ATP1A1
- ATP1A10S
- ATP1A2
- ATP1A3
- ATP1A4
- ATP1B1
- ATP1B2
- ATP1B3
- ATP1B4
- ATP2A1
- ATP2A2
- ATP2A3
- ATP2B1
- ATP2B2
- ATP2B3
- ATP2B4
- ATP2C1

Done



## Human Gene ATP1A1 (uc010owv.1) Description and Page Index

**Description:** Homo sapiens ATPase, Na+/K+ transporting, alpha 1 polypeptide (ATP1A1), transcript variant 4, mRNA.

**RefSeq Summary (NM\_001160234):** The protein encoded by this gene belongs to the family of P-type cation transport ATPases, and to the subfamily of Na+/K+-ATPases. Na+/K+-ATPase is an integral membrane protein responsible for establishing and maintaining the electrochemical gradients of Na and K ions across the plasma membrane. These gradients are essential for osmoregulation, for sodium-coupled transport of a variety of organic and inorganic molecules, and for electrical excitability of nerve and muscle. This enzyme is composed of two subunits, a large catalytic subunit (alpha) and a smaller glycoprotein subunit (beta). The catalytic subunit of Na+/K+-ATPase is encoded by multiple genes. This gene encodes an alpha 1 subunit. Multiple transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, May 2009].

**Strand:** + **Genomic Size:** 31594 **Exon Count:** 22 **Coding Exon Count:** 21

# UCSC Genome Browser

Human chr1:116,915,795-116,947,396 +

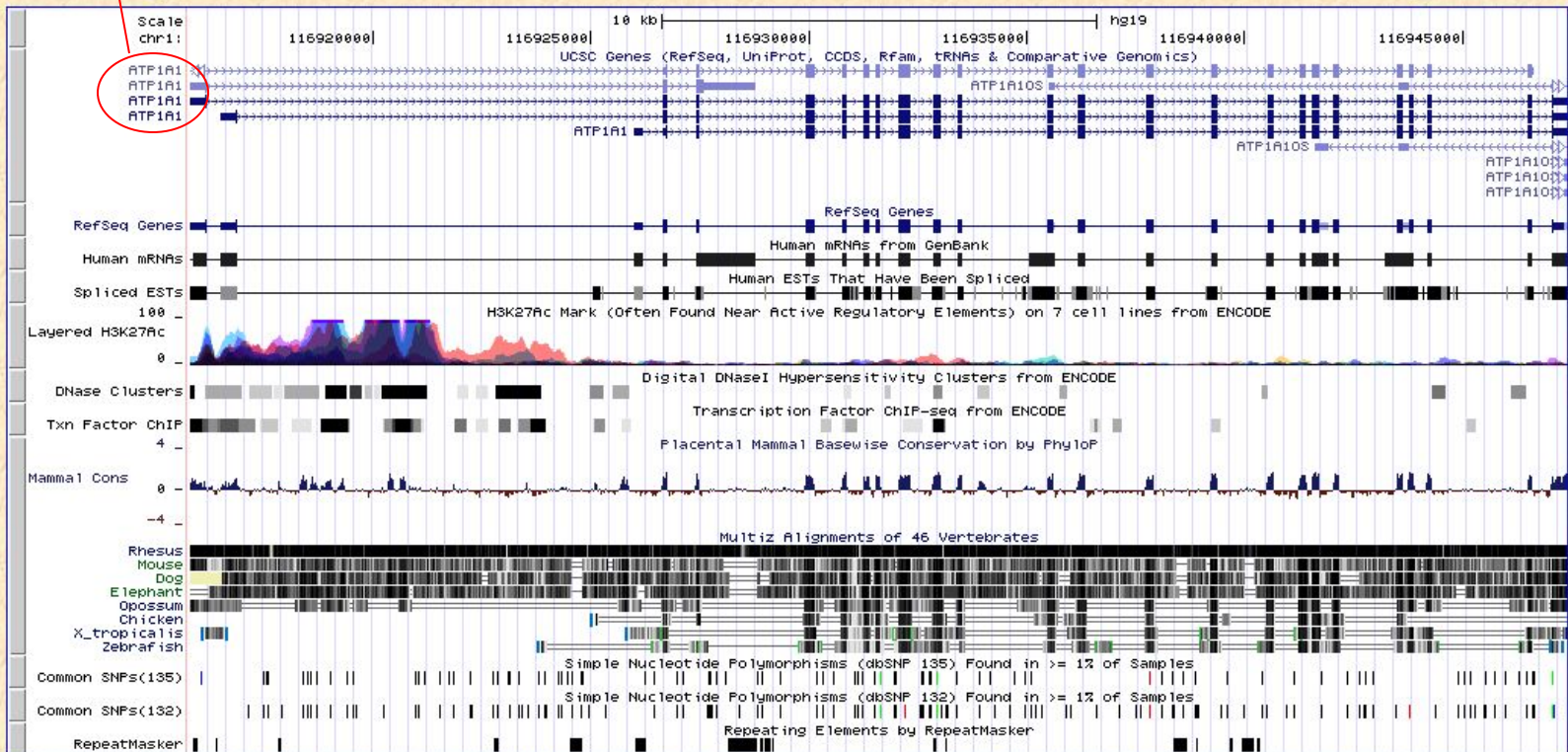
Home Genomes Blat Tables Gene Sorter PCR DNA Convert PDF/PS Session Ensembl NCBI Help

## UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr1:116,915,795-116,947,396    size 31,602 bp.

chr1 (p13.1) 33 1p31.1 1q12 52.1 1q41 q43 q44



move start

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options.

move end

< 2.0 >

Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

< 2.0 >



# UCSC Genome Browser

**Regulation** refresh

<input checked="" type="checkbox"/> <a href="#">ENCODE Regulation...</a> show ▾	<input checked="" type="checkbox"/> <a href="#">CD34 DnaseI</a> hide ▾	<a href="#">CpG Islands</a> hide ▾	<input checked="" type="checkbox"/> <a href="#">ENC DNA Methyl...</a> hide ▾	<input checked="" type="checkbox"/> <a href="#">ENC DNase/FAIRE...</a> hide ▾	<input checked="" type="checkbox"/> <a href="#">ENC Histone...</a> hide ▾
<input checked="" type="checkbox"/> <a href="#">ENC RNA Binding...</a> hide ▾	<input checked="" type="checkbox"/> <a href="#">ENC TF Binding...</a> hide ▾	<input checked="" type="checkbox"/> <a href="#">ORegAnno</a> hide ▾	<input checked="" type="checkbox"/> <a href="#">Stanf Nucleosome</a> hide ▾	<input checked="" type="checkbox"/> <a href="#">SUNY SwitchGear</a> hide ▾	<input checked="" type="checkbox"/> <a href="#">SwitchGear TSS</a> hide ▾
<a href="#">TFBS Conserved</a> hide ▾	<a href="#">TS miRNA sites</a> hide ▾	<a href="#">UMMS Brain Hist</a> hide ▾	<a href="#">Vista Enhancers</a> hide ▾	<input checked="" type="checkbox"/> <a href="#">NKI Nuc Lamina...</a> hide ▾	<input checked="" type="checkbox"/> <a href="#">UCSF Brain Methyl</a> hide ▾

**Comparative Genomics** refresh

<a href="#">Conservation</a> full ▾	<input checked="" type="checkbox"/> <a href="#">Cons Indels MmCf</a> hide ▾	<a href="#">GERP</a> hide ▾	<input checked="" type="checkbox"/> <a href="#">Evo Cpg</a> hide ▾	<a href="#">Primate Chain/Net</a> hide ▾	<a href="#">Placental Chain/Net</a> hide ▾
<a href="#">hg19Patch2 Chain/Net</a> hide ▾	<a href="#">Vertebrate Chain/Net</a> hide ▾				

**Neandertal Assembly and Analysis** refresh

**Variation and Repeats** refresh

<a href="#">Common SNPs(135)</a> dense ▾	<a href="#">Flagged SNPs(135)</a> hide ▾	<a href="#">Mult. SNPs(135)</a> hide ▾	<a href="#">All SNPs(135)</a> hide ▾	<a href="#">Common SNPs(132)</a> dense ▾	<a href="#">Flagged SNPs(132)</a> hide ▾
<a href="#">Mult. SNPs(132)</a> hide ▾	<a href="#">All SNPs(132)</a> hide ▾	<a href="#">SNPs (131)</a> hide ▾	<a href="#">Arrays</a> hide ▾	<input checked="" type="checkbox"/> <a href="#">GIS DNA PET</a> hide ▾	<input checked="" type="checkbox"/> <a href="#">HAIB Genotype</a> hide ▾
<input checked="" type="checkbox"/> <a href="#">SNP Arrays</a> hide ▾	<a href="#">HGDP Allele Freq</a> hide ▾	<input checked="" type="checkbox"/> <a href="#">HapMap SNPs</a> hide ▾	<a href="#">DGV Struct Var</a> hide ▾	<a href="#">Segmental Dups</a> hide ▾	<a href="#">RepeatMasker</a> dense ▾
<a href="#">Interrupted Rpts</a> hide ▾	<a href="#">Simple Repeats</a> hide ▾	<a href="#">Microsatellite</a> hide ▾	<a href="#">Self Chain</a> hide ▾	<input checked="" type="checkbox"/> <a href="#">Genome Variants</a> hide ▾	<a href="#">NumtS Sequence</a> hide ▾

refresh

# UCSC Genome Browser

UCSC Genome Browser Home

http://genome.ucsc.edu/

## UCSC Genome Bioinformatics

Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Proteome - Session - FAQ - Help

**Genome Browser**

**ENCODE**

**Neandertal**

**Blat**

**Table Browser**

**Gene Sorter**

**In Silico PCR**

**Genome Graphs**

**Galaxy**

**VisiGene**

**Proteome Browser**

**Utilities**

**Downloads**

**Release Log**

### VERTEBRATES - Complete annotation sets

<a href="#">Human</a>	<a href="#">Horse</a>	<a href="#">Platypus</a>
<a href="#">Cat</a>	<a href="#">Lamprey</a>	<a href="#">Rabbit</a>
<a href="#">Chicken</a>	<a href="#">Lizard</a>	<a href="#">Rat</a>
<a href="#">Chimpanzee</a>	<a href="#">Marmoset</a>	<a href="#">Rhesus</a>
<a href="#">Cow</a>	<a href="#">Medaka</a>	<a href="#">Sheep</a>
<a href="#">Dog</a>	<a href="#">Microbat</a>	<a href="#">Stickleback</a>
<a href="#">Elephant</a>	<a href="#">Mouse</a>	<a href="#">Tetraodon</a>
<a href="#">Fugu</a>	<a href="#">Opossum</a>	<a href="#">Turkey</a>
<a href="#">Gibbon</a>	<a href="#">Orangutan</a>	<a href="#">Wallaby</a>
<a href="#">Gorilla</a>	<a href="#">Panda</a>	<a href="#">X. tropicalis</a>
<a href="#">Guinea pig</a>	<a href="#">Pig</a>	<a href="#">Zebra finch</a>
		<a href="#">Zebrafish</a>

### VERTEBRATES - Sequence downloads only

<a href="#">Armadillo</a>	<a href="#">Tenrec</a>
<a href="#">Bushbaby</a>	<a href="#">Tree shrew</a>
<a href="#">European hedgehog</a>	<a href="#">J. Craig Venter</a>
<a href="#">Shrew</a>	

### DEUTEROSTOMES

- [C. intestinalis](#)
- [Lancelet](#)
- [S. purpuratus](#)

### INSECTS

<a href="#">A. gambiae</a>	<a href="#">D. pseudoobscura</a>
<a href="#">A. mellifera</a>	<a href="#">D. sechellia</a>
<a href="#">D. ananassae</a>	<a href="#">D. simulans</a>
<a href="#">D. erecta</a>	<a href="#">D. virilis</a>
<a href="#">D. grimshawi</a>	<a href="#">D. willistoni</a>
<a href="#">D. melanogaster</a>	<a href="#">D. yakuba</a>
<a href="#">D. mojavensis</a>	<a href="#">T. castaneum</a>
<a href="#">D. persimilis</a>	

### YEAST AND OTHERS

- [S. cerevisiae](#)
- [Sea hare](#)
- [Denisova](#)

### OTHER DOWNLOADS

- [Shared Data \(Protein DBs, hgFixed, visiGene\)](#)
- [LiftOver Files](#)
- [ENCODE Project Files \(Genome-wide Phase\)](#)
- [ENCODE Project Files \(Pilot Phase\)](#)

Working draft assemblies for a large

...lls over chromosomes, showing the

...n on groups of genes that can be

...s convenient access to the

...images to examine expression

...cross-departmental team within the

...z (UCSC). If you have feedback or

...list.

News Archives ▶

...aining seminars by email, subscribe

...the human assembly

...on data not included in previous

...of reference SNPs to the human

...sting and easily defined subsets of

Done



# Complete human genome

## Human Genome

Feb. 2009 (hg19, GRCh37)

- [Full data set](#)
- [Data set by chromosome](#)
- [Annotation database](#)
- [GC percent data](#)
- [Protein database for hg19](#)
- [SNP135-masked FASTA files](#)
- [SNP132-masked FASTA files](#)
- [SNP131-masked FASTA files](#)
- [LiftOver files](#)
- Pairwise Alignments
  - [Human self alignments](#)
  - [Human/Chimp \(panTro3\)](#)
  - [Human/Chimp \(panTro2\)](#)
  - [Human/Gorilla \(gorGor3\)](#)
  - [Human/Gorilla \(gorGor1\)](#)
  - [Human/Orangutan \(ponAbe2\)](#)
  - [Human/Gibbon \(nomLeu1\)](#)
  - [Human/Rhesus \(rheMac2\)](#)
  - [Human/Baboon \(papHam1\)](#)
  - [Human/Marmoset \(calJac3\)](#)
  - [Human/Marmoset \(calJac1\)](#)
  - [Human/Bushbaby \(otoGar1\)](#)
  - [Human/Mouse lemur \(micMur1\)](#)
  - [Human/Tree shrew \(tupBel1\)](#)
  - [Human/Shrew \(sorAra1\)](#)
  - [Human/Mouse \(mm9\)](#)
  - [Human/Rat \(rn4\)](#)
  - [Human/Squirrel \(speTri1\)](#)
  - [Human/Guinea pig \(cavPor3\)](#)
  - [Human/Rabbit \(oryCun2\)](#)
  - [Human/Pika \(ochPri2\)](#)
  - [Human/Hedgehog \(eriEur1\)](#)
  - [Human/Cat \(felCat4\)](#)
  - [Human/Cat \(felCat3\)](#)
  - [Human/Panda \(ailMel1\)](#)
  - [Human/Dog \(canFam2\)](#)
  - [Human/Horse \(equCab2\)](#)
  - [Human/Microbat \(myoLuc1\)](#)
  - [Human/Pig \(susScr2\)](#)
  - [Human/Megabat \(pteVam1\)](#)
  - [Human/Cow \(bosTau6\)](#)
  - [Human/Cow \(bosTau4\)](#)
  - [Human/Alpaca \(vicPac1\)](#)
  - [Human/Dolphin \(turTru1\)](#)
  - [Human/Tarsier \(tarSyr1\)](#)
  - [Human/Rock hyrax \(proCap1\)](#)
  - [Human/Kangaroo rat \(dipOrd1\)](#)
  - [Human/Sloth \(choHof1\)](#)
  - [Human/Wallaby \(macEug1\)](#)
  - [Human/Armadillo \(dasNov2\)](#)
  - [Human/Tenrec \(eTenrec1\)](#)
  - [Human/Elephant \(eleGor1\)](#)
  - [Human/Opossum \(mopMus1\)](#)
  - [Human/Chicken \(galGal4\)](#)
  - [Human/Turkey \(turChro1\)](#)
  - [Human/Zebra finch \(taeGut1\)](#)
  - [Human/Lizard \(anoCar2\)](#)
  - [Human/Lizard \(anoCar1\)](#)
  - [Human/X. tropicalis \(xenTro3\)](#)
  - [Human/X. tropicalis \(xenTro2\)](#)
  - [Human/Zebrafish \(danRer7\)](#)
- Multiple Alignments
  - [Multiple alignments of 45 vertebrate genomes with Human](#)
  - [Conservation scores for alignments of 45 vertebrate genomes with Human](#)
  - [Basewise conservation scores \(phyloP\) of 45 vertebrate genomes with Human](#)
  - [FASTA alignments of 45 vertebrate genomes with Human for CDS regions](#)

<a href="#">Name</a>	<a href="#">Last modified</a>	<a href="#">Size</a>	<a href="#">Description</a>
<a href="#">Parent Directory</a>		-	
<a href="#">chromApp.tar.gz</a>	20-Mar-2009 09:02	538K	
<a href="#">chromFa.tar.gz</a>	20-Mar-2009 09:21	905M	
<a href="#">chromFaMasked.tar.gz</a>	20-Mar-2009 09:30	477M	
<a href="#">chromOut.tar.gz</a>	20-Mar-2009 09:03	163M	
<a href="#">chromTrf.tar.gz</a>	20-Mar-2009 09:30	7.6M	
<a href="#">est.fa.gz</a>	11-Feb-2012 23:01	1.4G	
<a href="#">est.fa.gz.md5</a>	11-Feb-2012 23:01	44	
<a href="#">hg19.2bit</a>	08-Mar-2009 15:29	778M	
<a href="#">md5sum.txt</a>	29-Jul-2009 10:04	457	
<a href="#">mrna.fa.gz</a>	11-Feb-2012 22:33	166M	
<a href="#">mrna.fa.gz.md5</a>	11-Feb-2012 22:33	45	
<a href="#">refMrna.fa.gz</a>	11-Feb-2012 23:01	40M	
<a href="#">refMrna.fa.gz.md5</a>	11-Feb-2012 23:01	48	
<a href="#">upstream1000.fa.gz</a>	12-Feb-2012 20:32	7.5M	
<a href="#">upstream1000.fa.gz.md5</a>	12-Feb-2012 20:32	53	
<a href="#">upstream2000.fa.gz</a>	12-Feb-2012 20:34	14M	
<a href="#">upstream2000.fa.gz.md5</a>	12-Feb-2012 20:34	53	
<a href="#">upstream5000.fa.gz</a>	12-Feb-2012 20:37	35M	
<a href="#">upstream5000.fa.gz.md5</a>	12-Feb-2012 20:37	53	
<a href="#">xenoMrna.fa.gz</a>	11-Feb-2012 22:41	1.9G	
<a href="#">xenoMrna.fa.gz.md5</a>	11-Feb-2012 22:41	49	

chromFa.tar.gz - The assembly sequence in one file per chromosome. Repeats from RepeatMasker and Tandem Repeats Finder (with period of 12 or less) are shown in lower case; non-repeating sequence is shown in upper case.

chromFaMasked.tar.gz - The assembly sequence in one file per chromosome. Repeats are masked by capital Ns; non-repeating sequence is shown in upper case.

chromOut.tar.gz - RepeatMasker .out files (one file per chromosome). RepeatMasker was run with the -s (sensitive) setting. Using: Jan 29 2009 (open-3-2-7) version of RepeatMasker and RELEASE 20090120 of library RepeatMaskerLib.embl

# Transposable Elements

- 45% of the human genome is occupied by transposons and transposon-like repetitive elements.
- Barbara McClintock (1902-1992) in 50s.
- Nobel prize in 1983

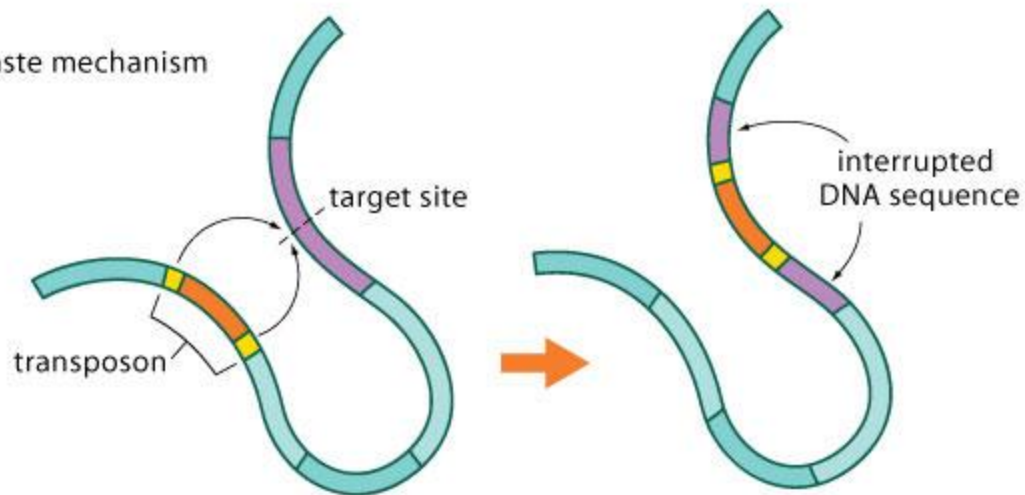




## Two methods of transposition:

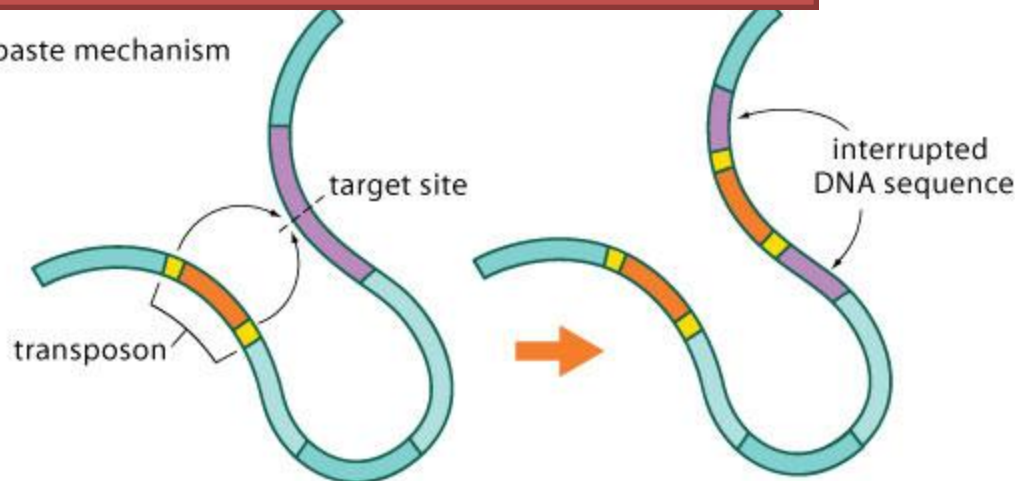
1. Cut-and-paste mechanism

Class II



Class I – retrotransposons (via RNA intermediate)

2. Copy-and-paste mechanism



Схожесть с ретровирусами  
Retrovirus reverse transcripton

<http://www.youtube.com/watch?v=eS1GODinO8w>

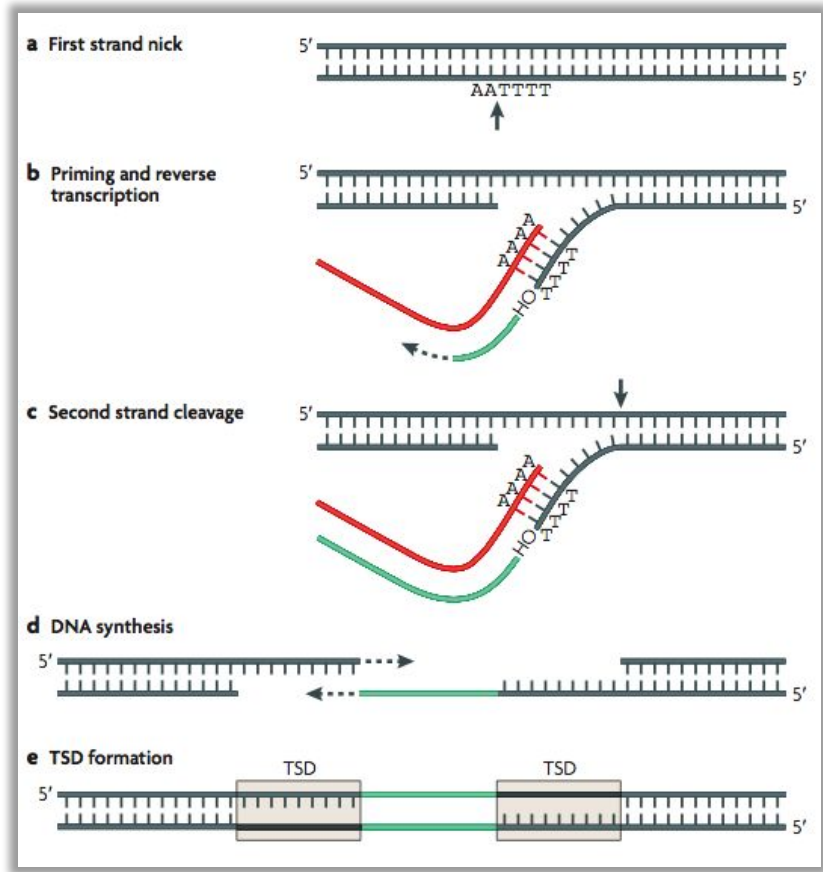
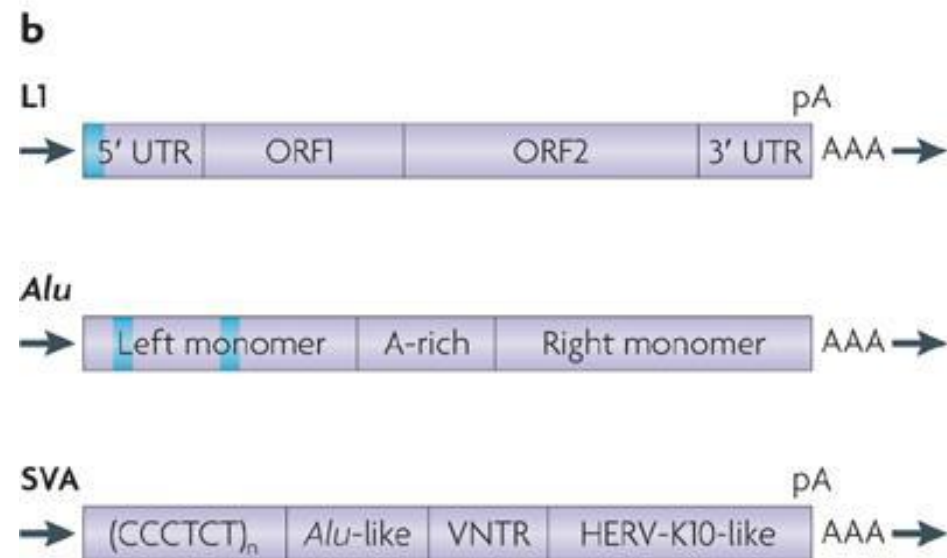
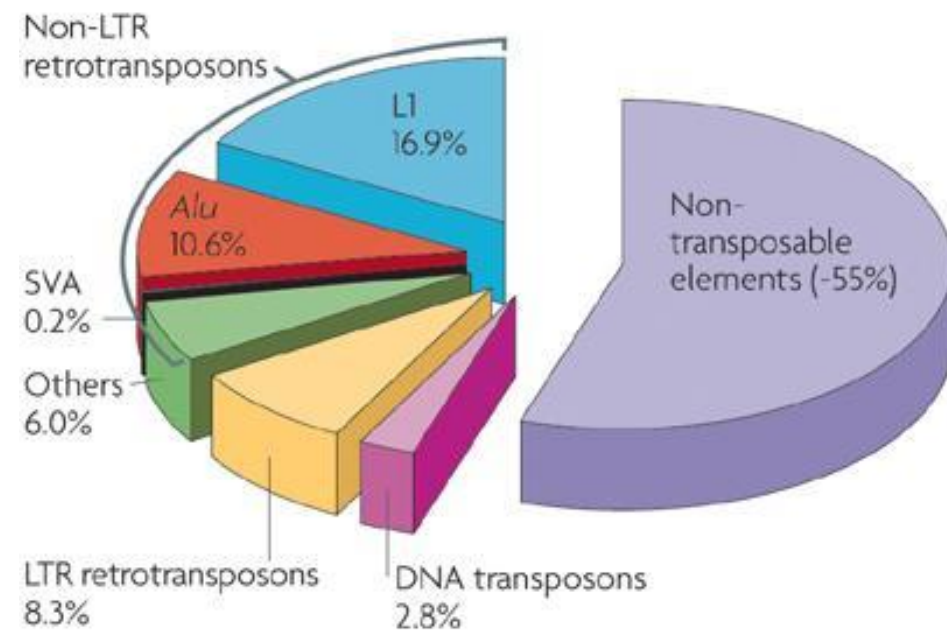
**TABLE 17.1****Categorization of Transposable Elements by Transposition Mechanism**

<b>Category</b>	<b>Examples</b>	<b>Host Organism</b>
<b>I. Cut-and-paste transposons</b>	<b>IS elements (e.g., IS50)</b>	<b>Bacteria</b>
	<b>Composite transposons (e.g., Tn5)</b>	<b>Bacteria</b>
	<b><i>Ac/Ds</i> elements</b>	<b>Maize</b>
	<b><i>P</i> elements</b>	<b><i>Drosophila</i></b>
	<b><i>hobo</i> elements</b>	<b><i>Drosophila</i></b>
	<b><i>piggyBac</i></b>	<b>moth</b>
	<b><i>Sleeping Beauty</i></b>	<b>salmon</b>
<b>II. Replicative transposons</b>	<b>Tn3 elements</b>	<b>Bacteria</b>
<b>III. Retrotransposons</b>		
<b>A. Retroviruslike elements (also called long terminal repeat, or LTR, retrotransposons)</b>	<b>Ty1</b>	<b>Yeast</b>
	<b><i>copia</i></b>	<b><i>Drosophila</i></b>
	<b><i>gypsy</i></b>	<b><i>Drosophila</i></b>
<b>B. Retroposons</b>	<b><i>F</i>, <i>G</i>, and <i>I</i> elements</b>	<b><i>Drosophila</i></b>
	<b>Telomeric retroposons</b>	<b><i>Drosophila</i></b>
	<b>LINEs (e.g., <i>L1</i>)</b>	<b>Humans</b>
	<b>SINEs (e.g., <i>Alu</i>)</b>	<b>Humans</b>



# The impact of retrotransposons on human genome evolution

Richard Cordaux & Mark A. Batzer



# Which transposable elements are active in the human genome?

Ryan E. Mills<sup>1,2</sup>, E. Andrew Bennett<sup>1,3</sup>, Rebecca C. Iskow<sup>1,3</sup> and Scott E. Devine<sup>1,2,3</sup>

<sup>1</sup> Department of Biochemistry, Emory University School of Medicine, Atlanta, GA 30322, USA

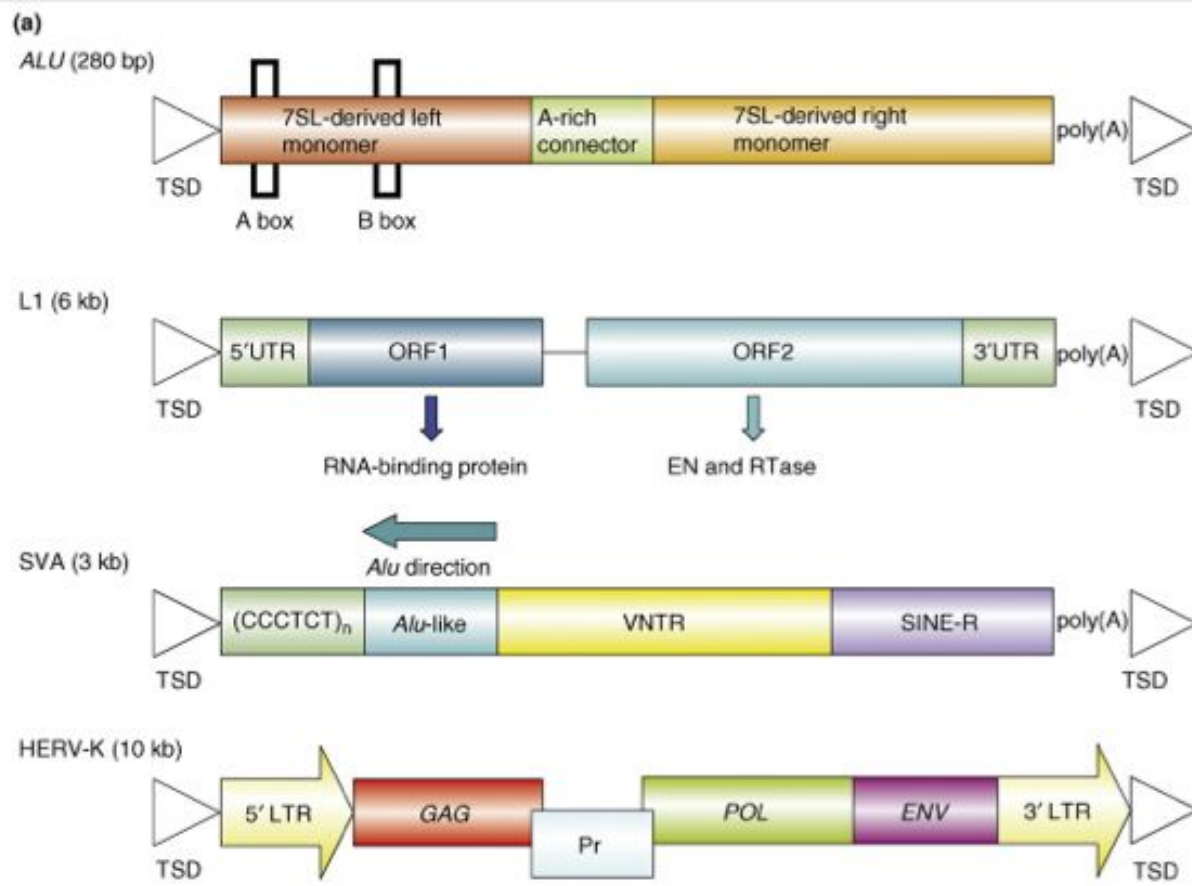
<sup>2</sup> Center for Bioinformatics, Emory University School of Medicine, Atlanta, GA 30322, USA

<sup>3</sup> Genetics and Molecular Biology Graduate Program, Emory University School of Medicine, Atlanta, GA 30322, USA

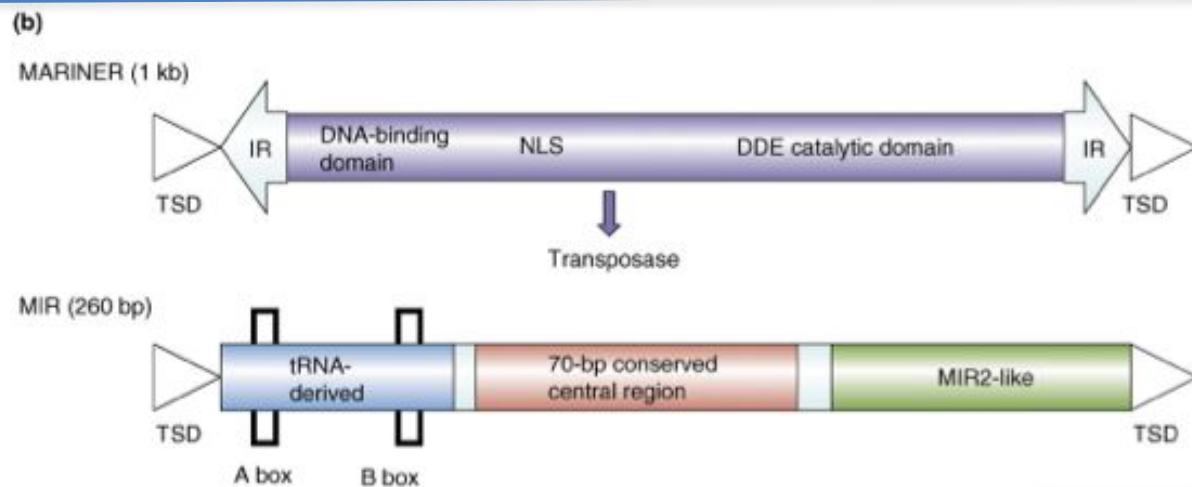
**Although a large proportion (44%) of the human genome is occupied by transposons and transposon-like repetitive elements, only a small proportion (<0.05%) of these elements remain active today. Recent evidence indicates that ~35–40 subfamilies of *Alu*, L1 and SVA elements (and possibly HERV-K elements) remain actively mobile in the human genome. These active transposons are of great interest because they continue to produce genetic diversity in human populations and also cause human diseases by integrating into genes. In this review, we examine these active human transposons and explore mechanistic factors that influence their mobilization.**



Active



Non-Active



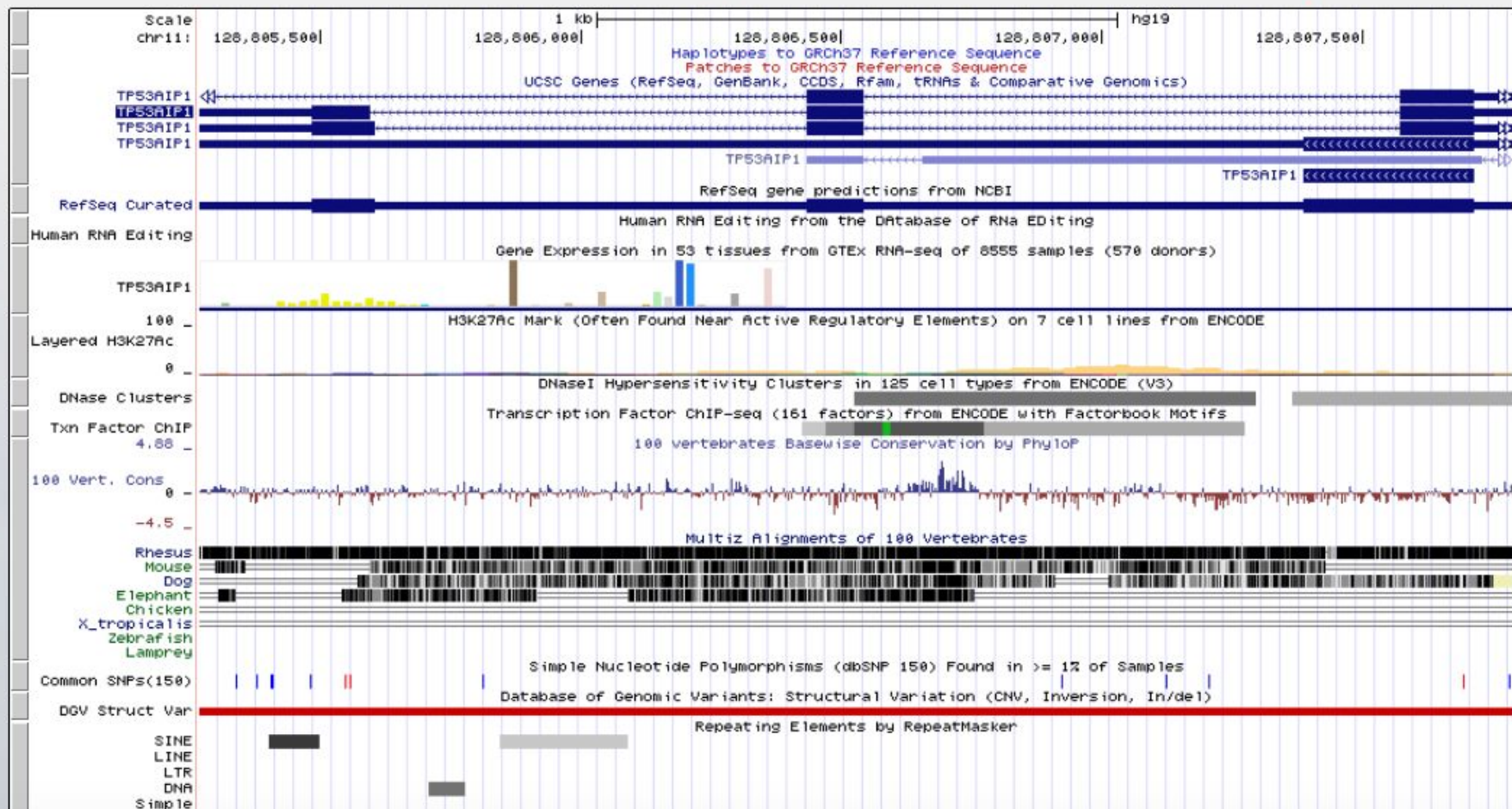
# First Layer of Genome Annotation

## UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr11:128,805,270-128,807,789 2,520 bp. enter position, gene symbol, HGVS or search terms

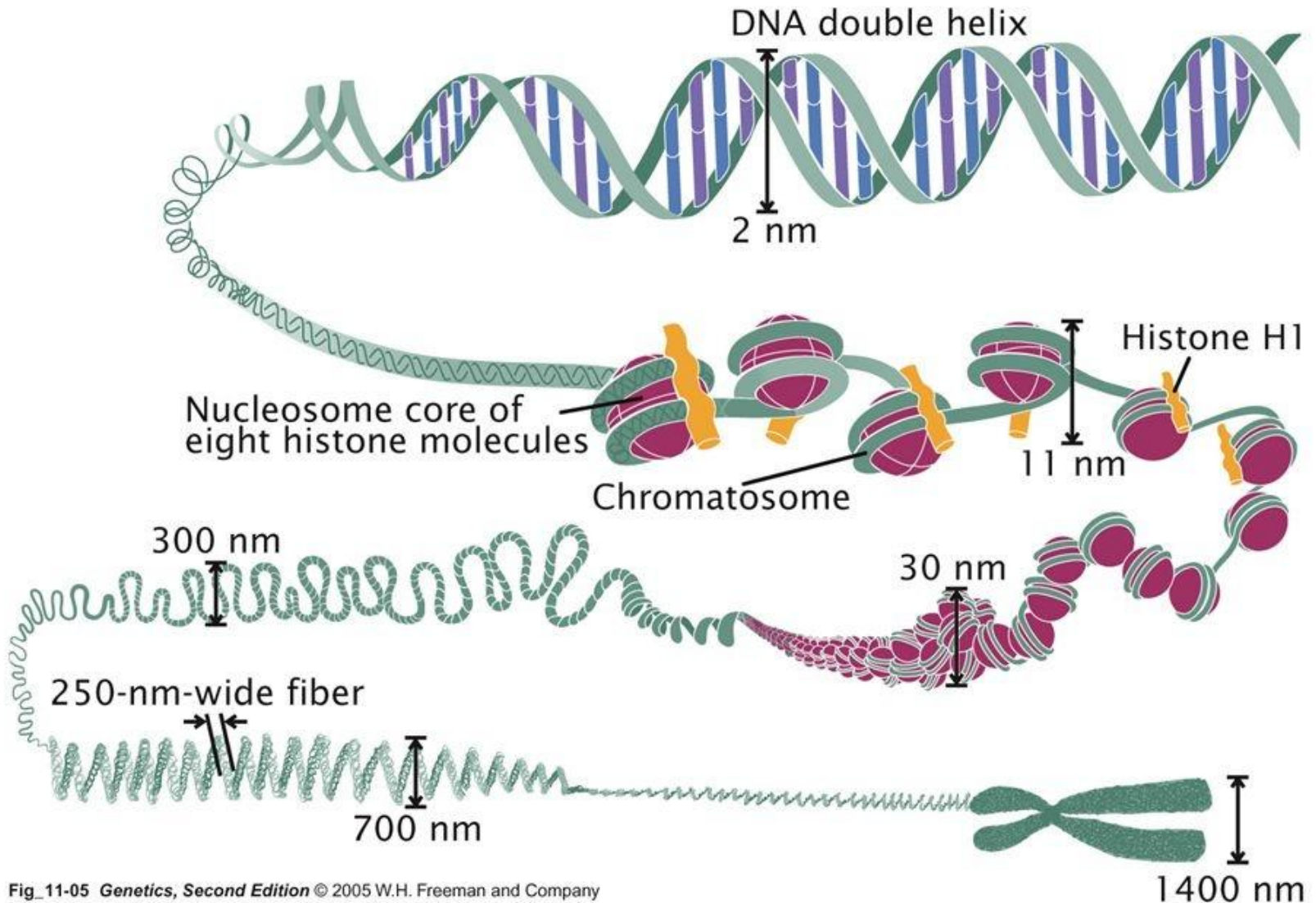
go [See us FREE @ ASHG Wed 7](#)





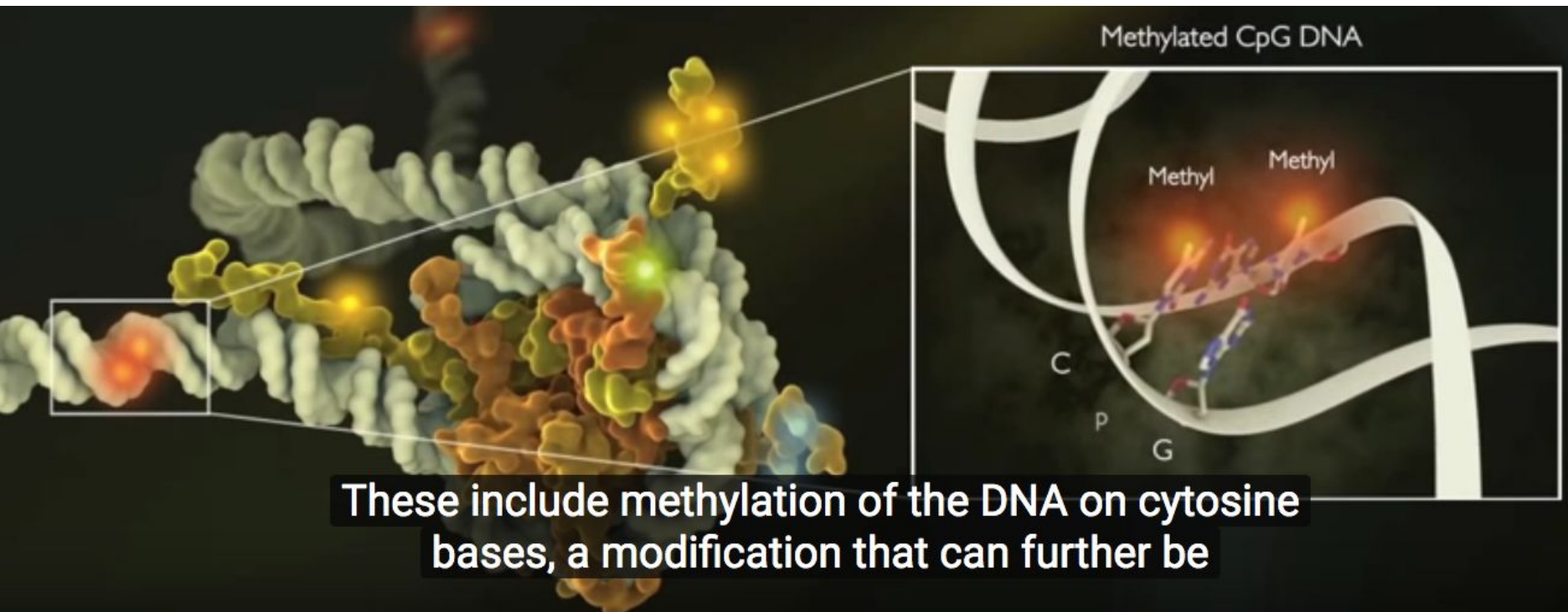
# Epigenetics

## Levels of DNA Packaging in Eukaryotes



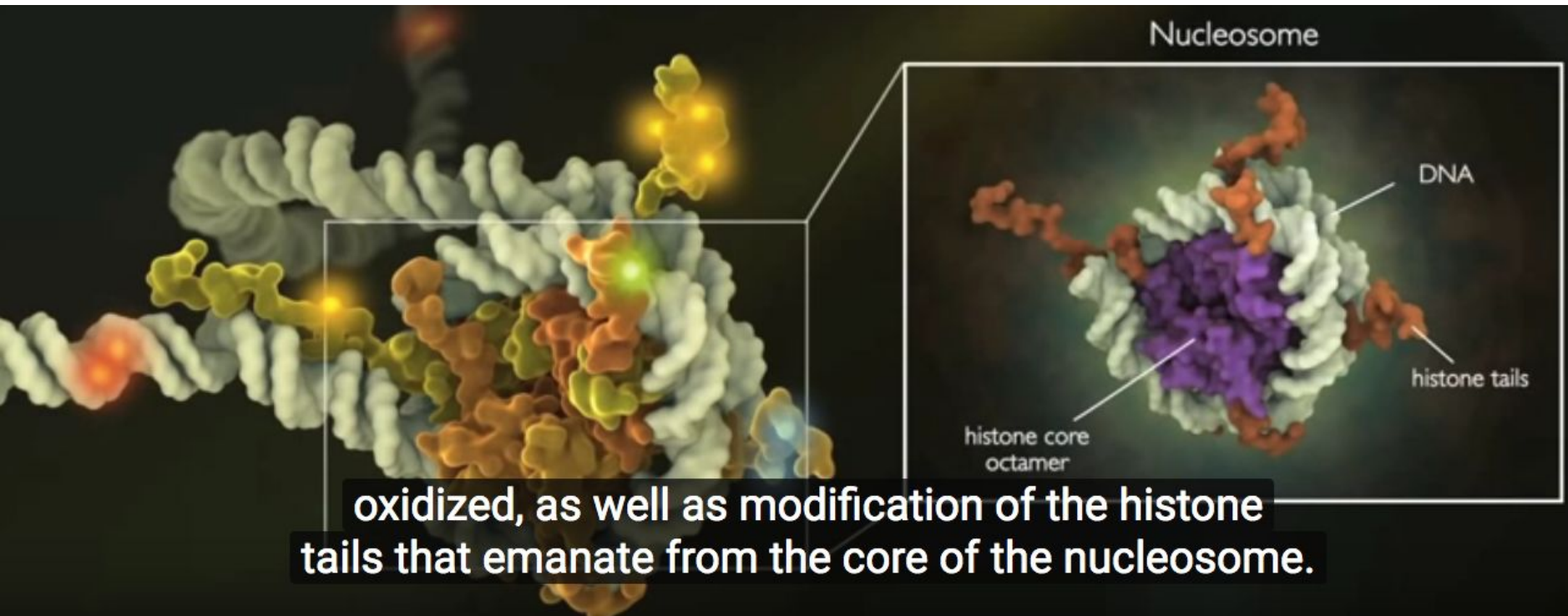
Fig\_11-05 Genetics, Second Edition © 2005 W.H. Freeman and Company

# Second Layer of Genome Annotation

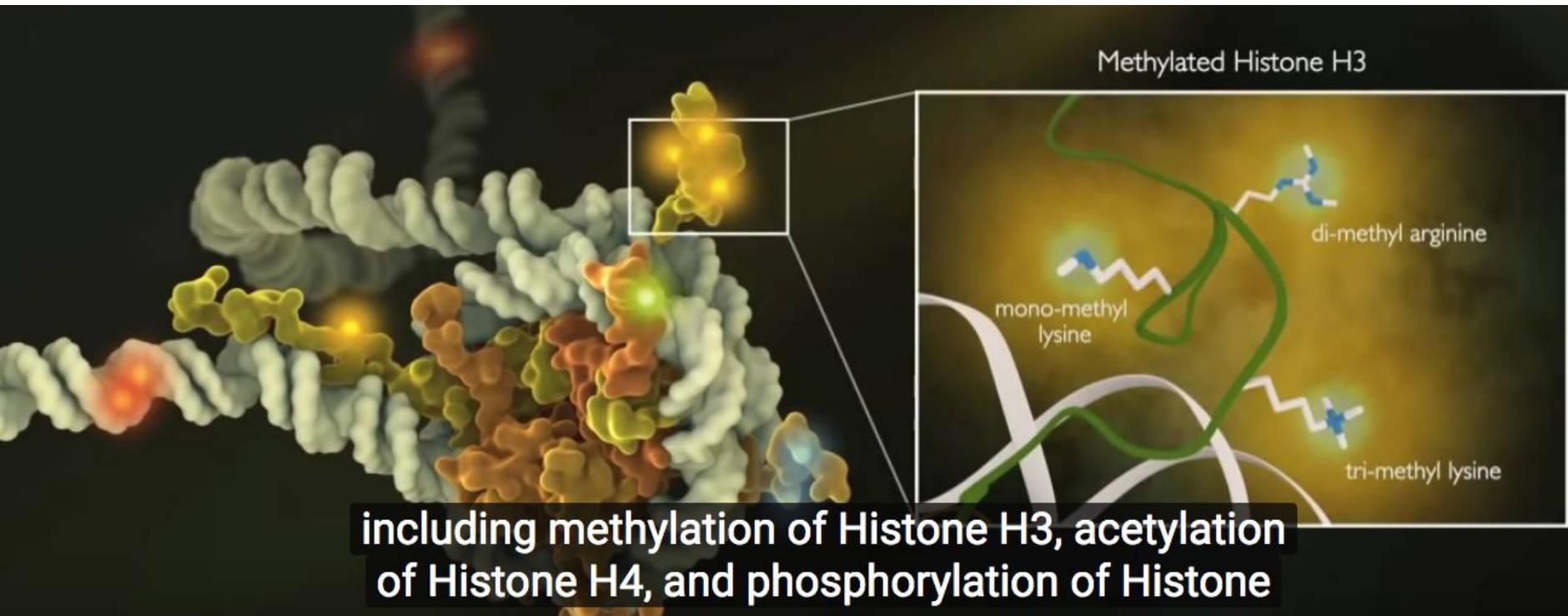




# Second Layer of Genome Annotation

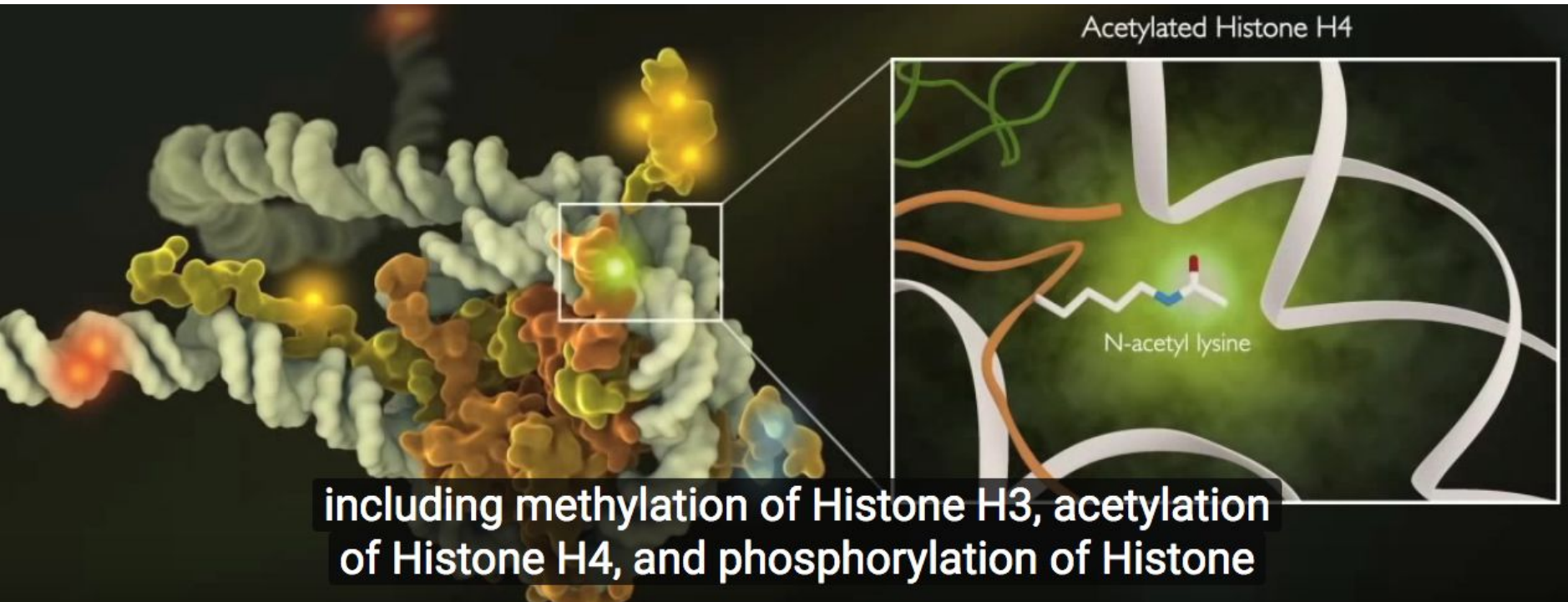


# Second Layer of Genome Annotation

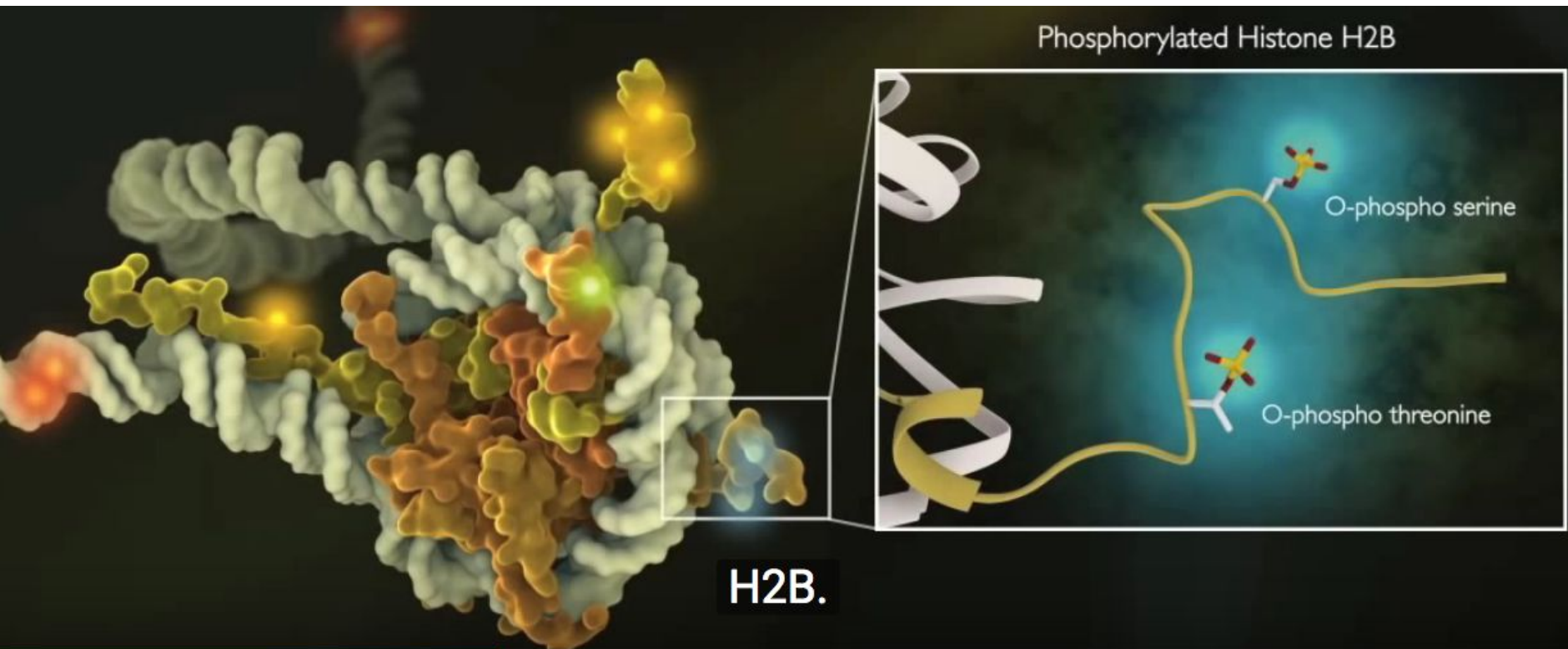




# Second Layer of Genome Annotation

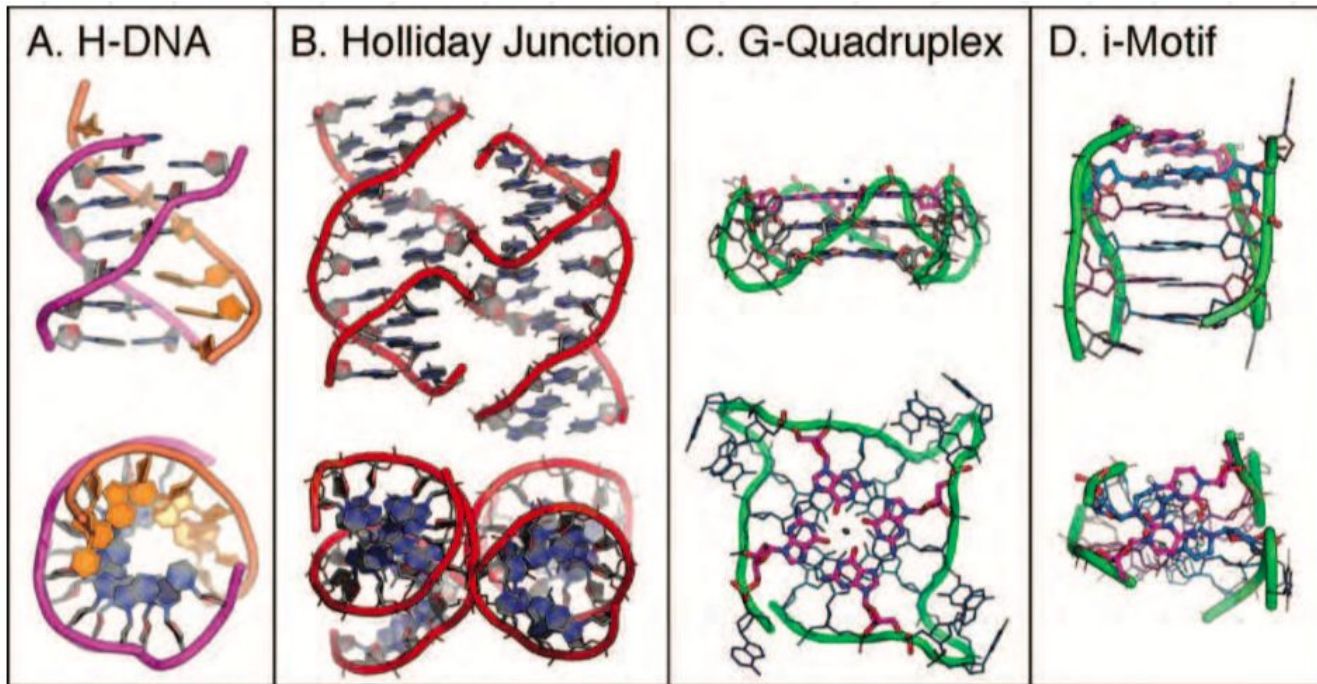
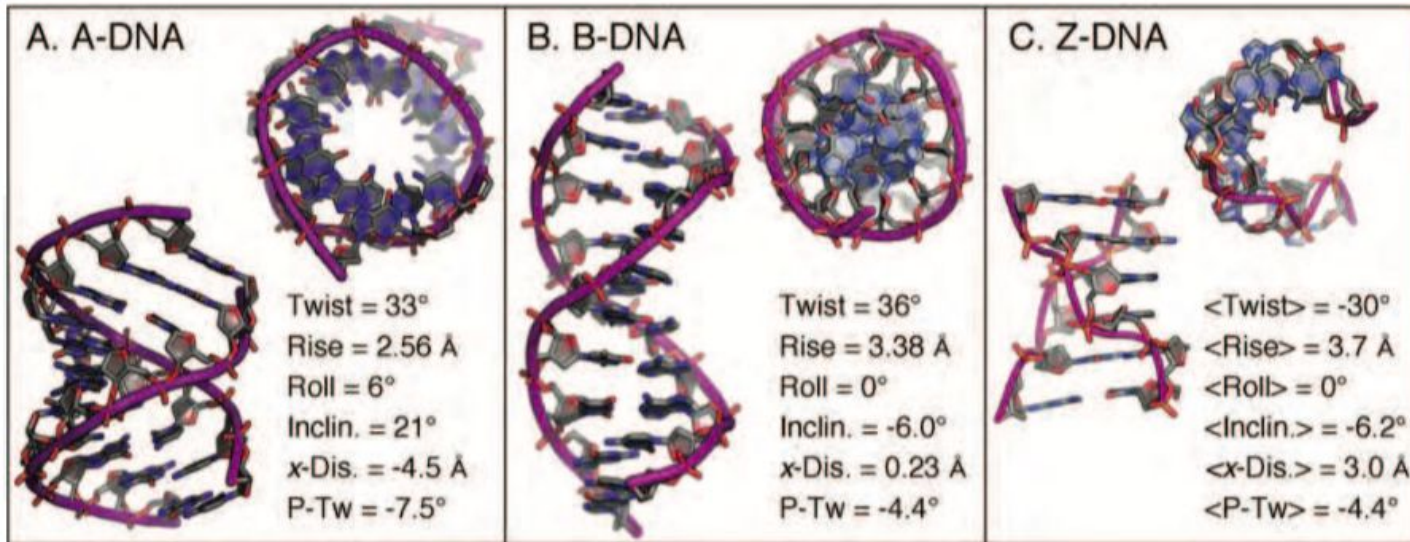


# Second Layer of Genome Annotation





# Third Layer of Genome Annotation



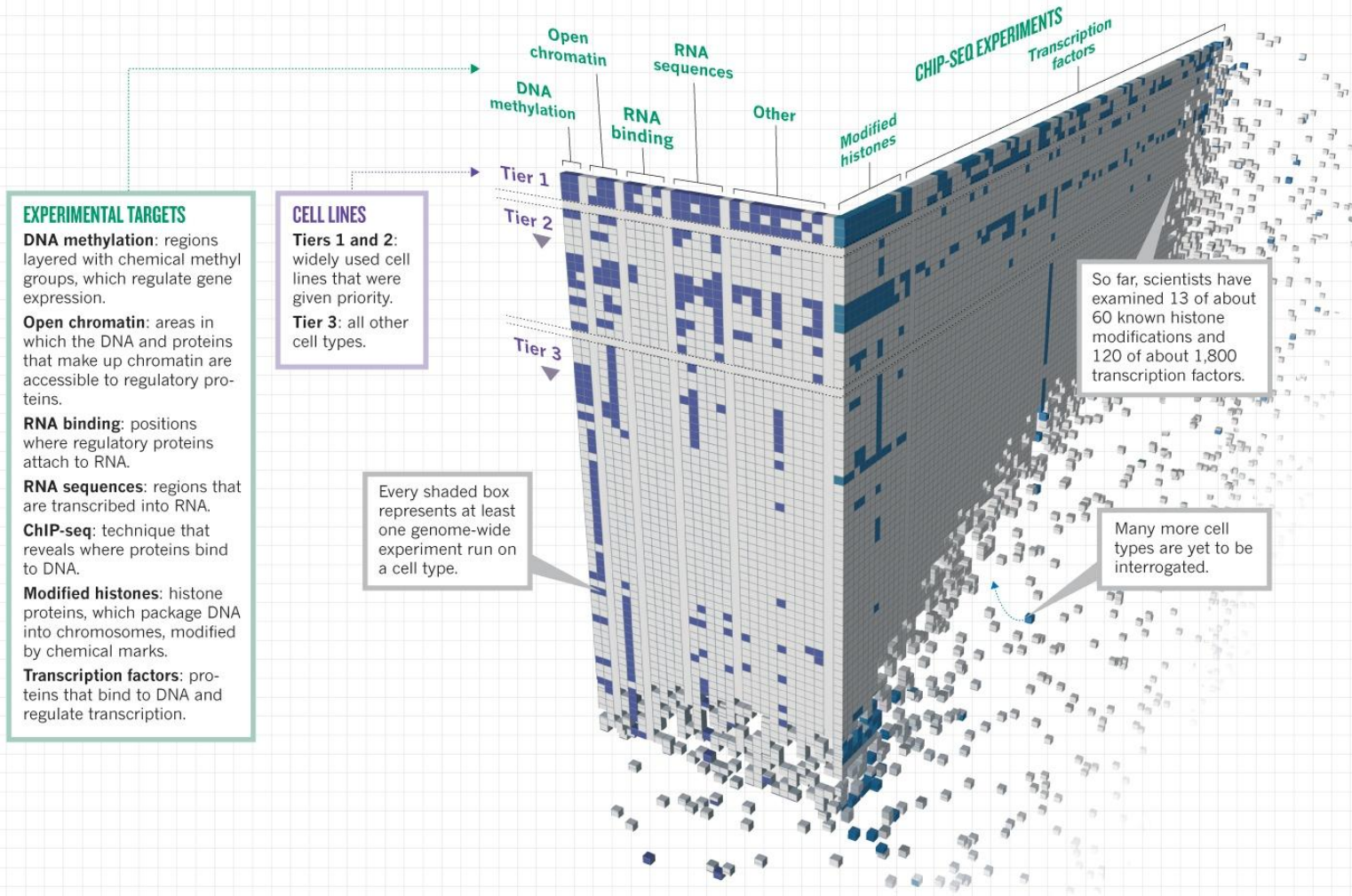
# Data Accumulation



# ENCODE: Encyclopedia of DNA Elements

## MAKING A GENOME MANUAL

Scientists in the Encyclopedia of DNA Elements Consortium have applied 24 experiment types (across) to more than 150 cell lines (down) to assign functions to as many DNA regions as possible — but the project is still far from complete.



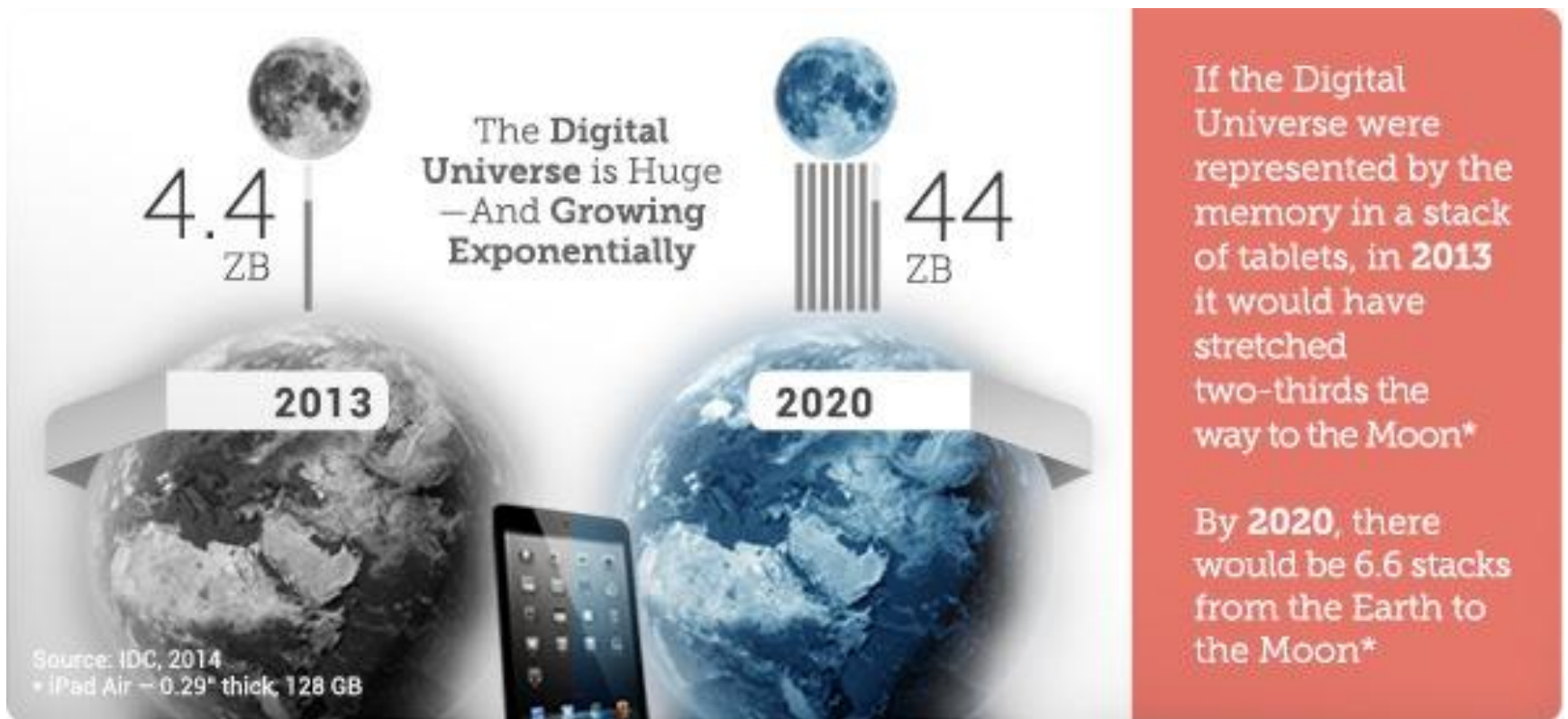
# Digital Universe

- Like the Physical Universe the Digital Universe is also expanding but much faster doubling every two years – and by 2020 will be 44 zettabytes ( $10^{21}$ )
- Every second a new 205 000 bytes come to being
- At the end of this lecture the digital universe will grow by 2 214 000 000 bytes or 2.2 GB.



# Digital Universe

Data Universe Will Expand To 44 Trillion GBs By 2020



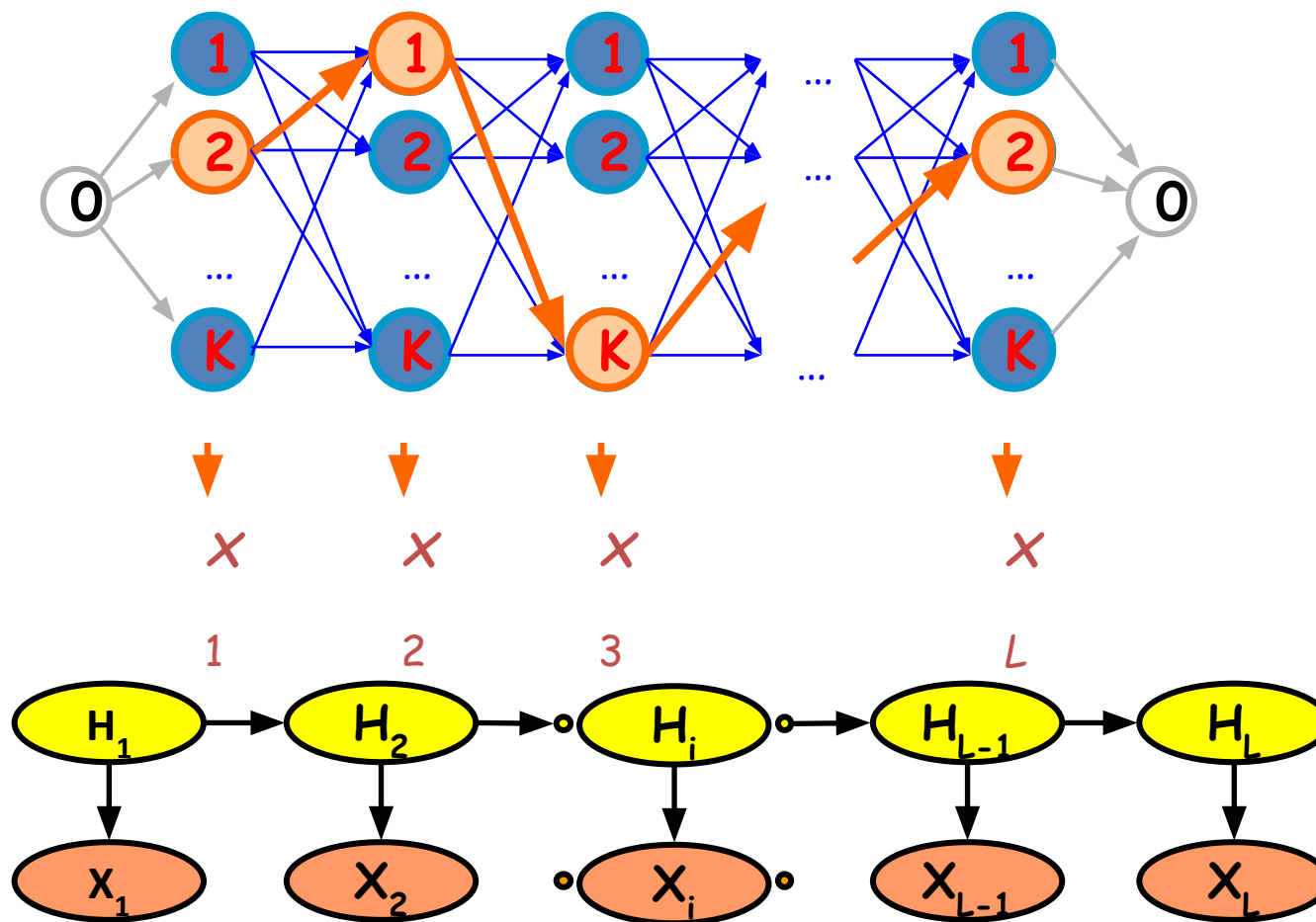
# Что делать?





Что получилось?  
(Success Stories)

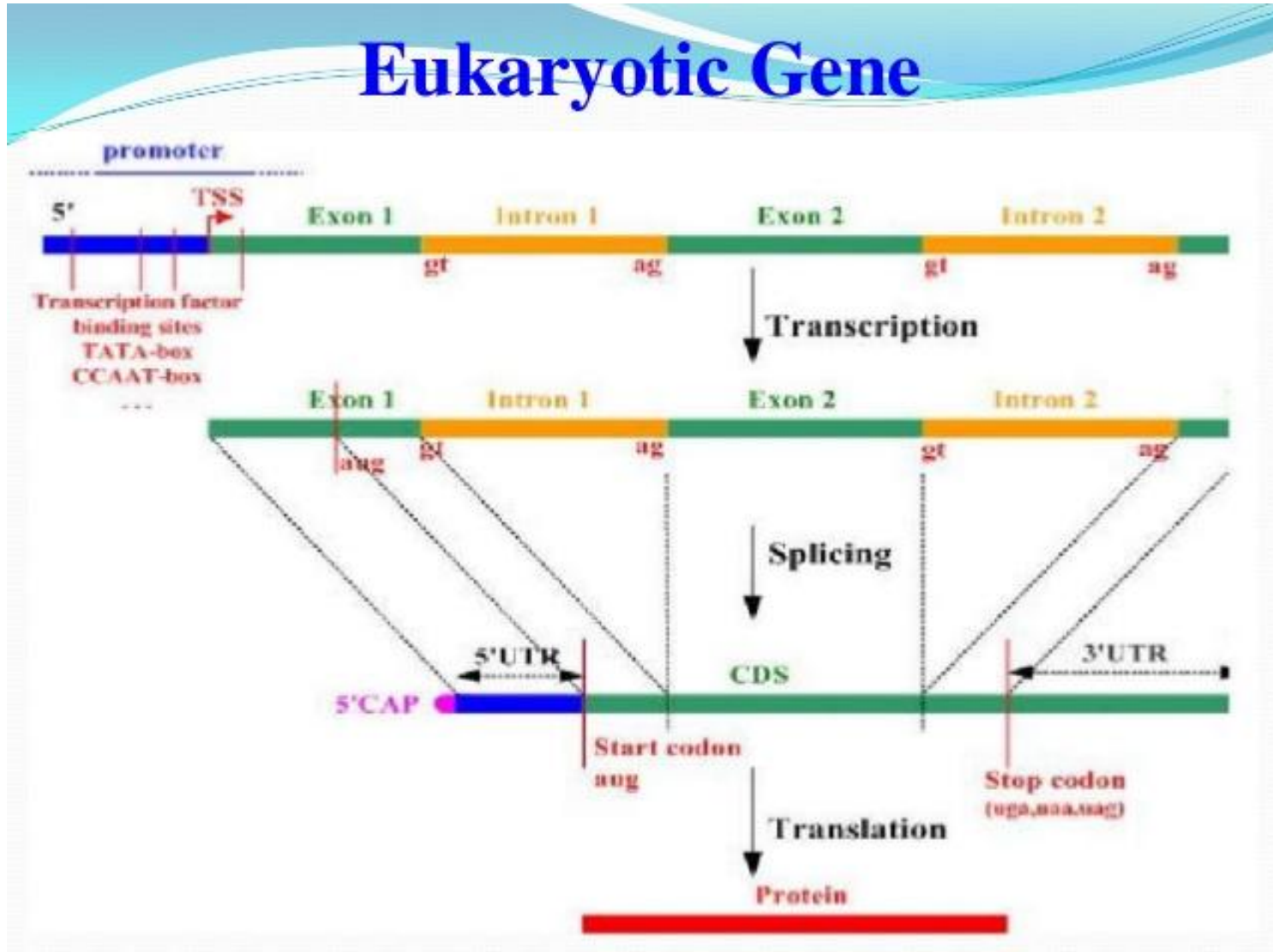
# СКРЫТЫЕ ЦЕПИ МАРКОВА





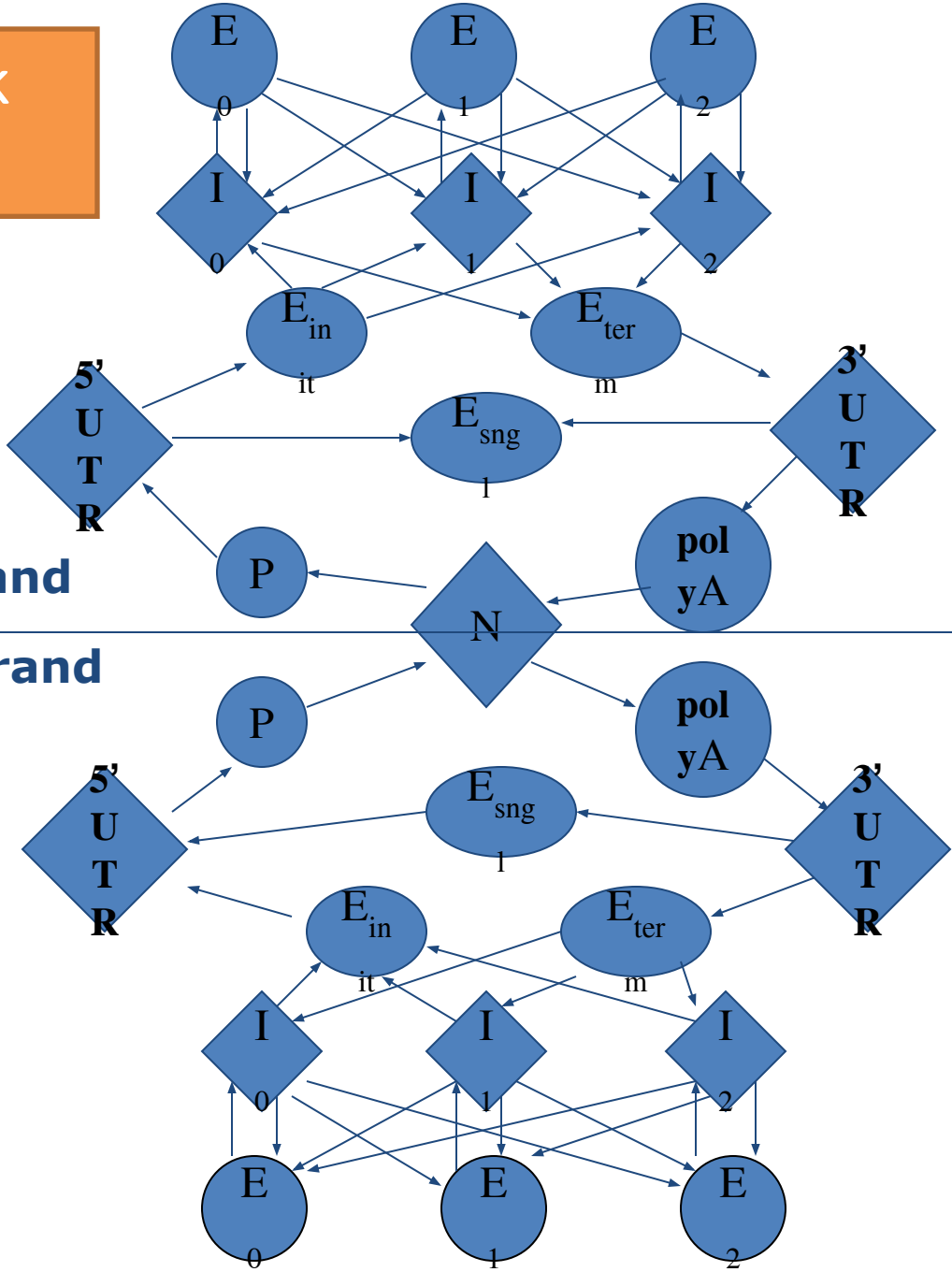
# Gene Prediction

## Eukaryotic Gene



# GeneMark HMM

- E- exons
- I- introns
- single exon
- 5' UTRs
- 3' UTRs
- P- promoter region polyA site
- N- intergenic region





# Promoter prediction

## McPromoter

- Hidden Markov model with six interpolated Markov chain submodels
  - upstream 1 and 2,
  - TATA box, spacer,
  - Initiator
  - downstream.
  - Gaussian densities of DNA physicochemical properties.
- Neural network classifier

# Deciphering the splicing code

Nature 2010

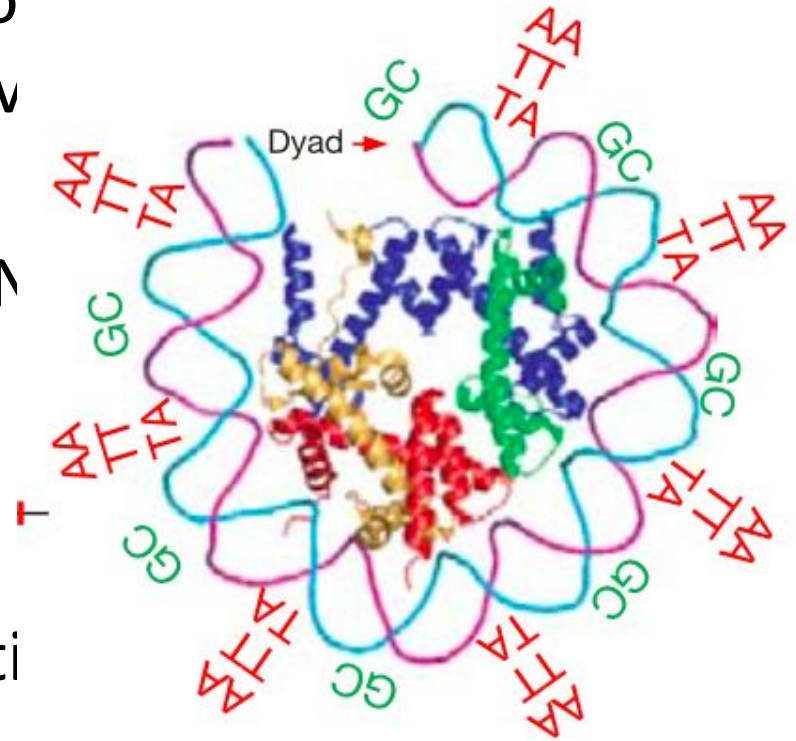
Yoseph Barash<sup>1,2\*</sup>, John A. Calarco<sup>2\*</sup>, Weijun Gao<sup>1</sup>, Qun Pan<sup>2</sup>, Xinchun Wang<sup>1,2</sup>, Ofer Shai<sup>1</sup>, Benjamin J. Blencowe<sup>2</sup>  
& Brendan J. Frey<sup>1,2,3</sup>

- predict tissue-dependent changes in alternative splicing for thousands of exons.
- 1,014 features: known motifs, new motifs, short motifs and features describing transcript structure
- trained on RNA-seq data
- single-layer logistic Bayesian network or neural network, or a weighted combination of single-layer decision trees.

# A genomic code for nucleosome positioning

Eran Segal<sup>1</sup>, Yvonne Fondufe-Mittendorf<sup>2</sup>, Lingyi Chen<sup>2</sup>, AnnChristine Thåström<sup>2</sup>, Yair Field<sup>1</sup>, Irene K. Moore<sup>2</sup>, Ji-Ping Z. Wang<sup>3</sup> & Jonathan Widom<sup>2</sup>

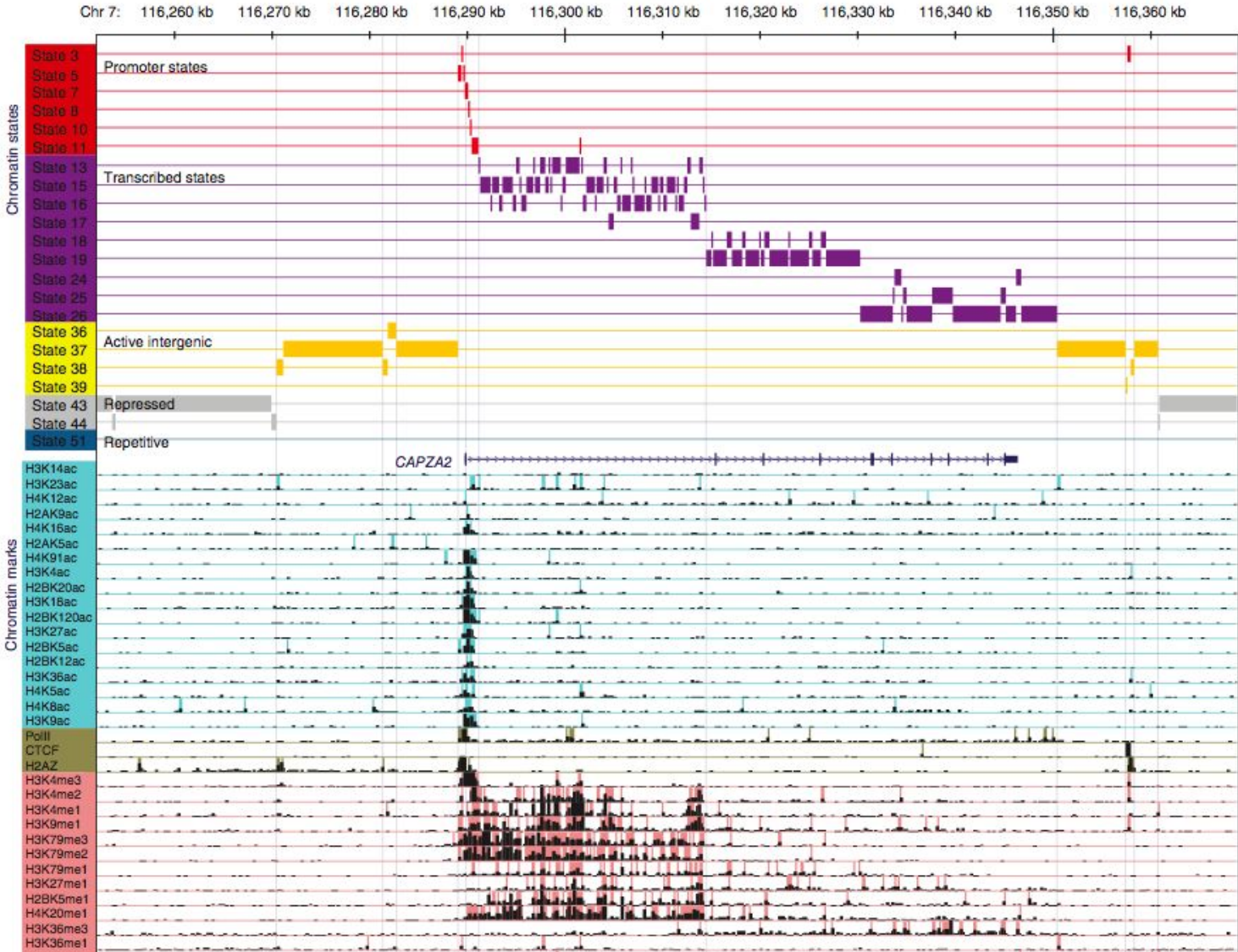
- Genome intrinsic organization<sup>f</sup> can explain ,50% of the in vivo nucleosome positions
- Probabilistic nucleosome–DN interaction model - built on dinucleotide distribution
- Thermodynamic model for predicting nucleosome position genome-wide.



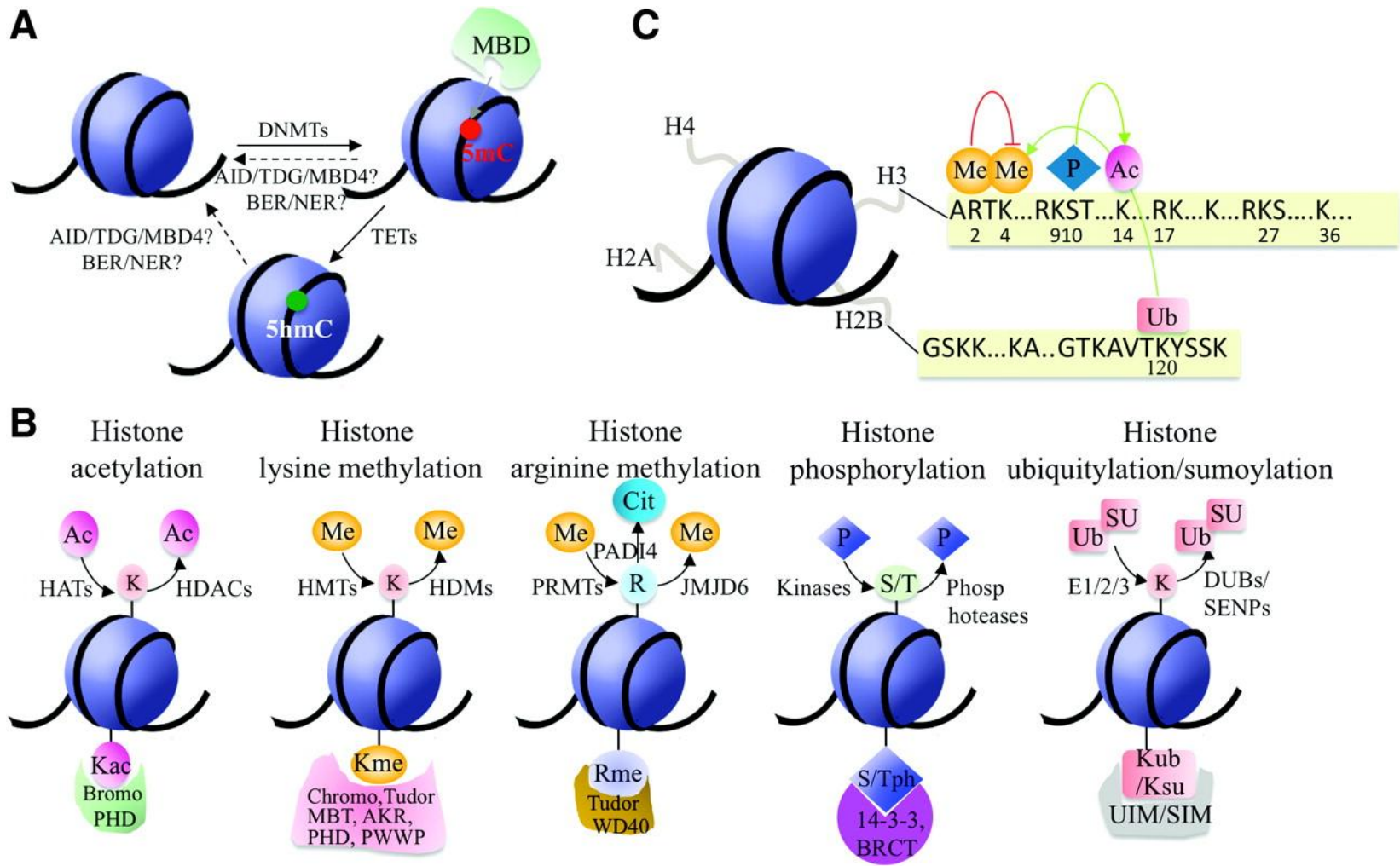


# Discovery and characterization of chromatin states for systematic annotation of the human genome

Jason Ernst<sup>1,2</sup> & Manolis Kellis<sup>1,2</sup>



# Schematic overview of epigenetic regulatory mechanisms.



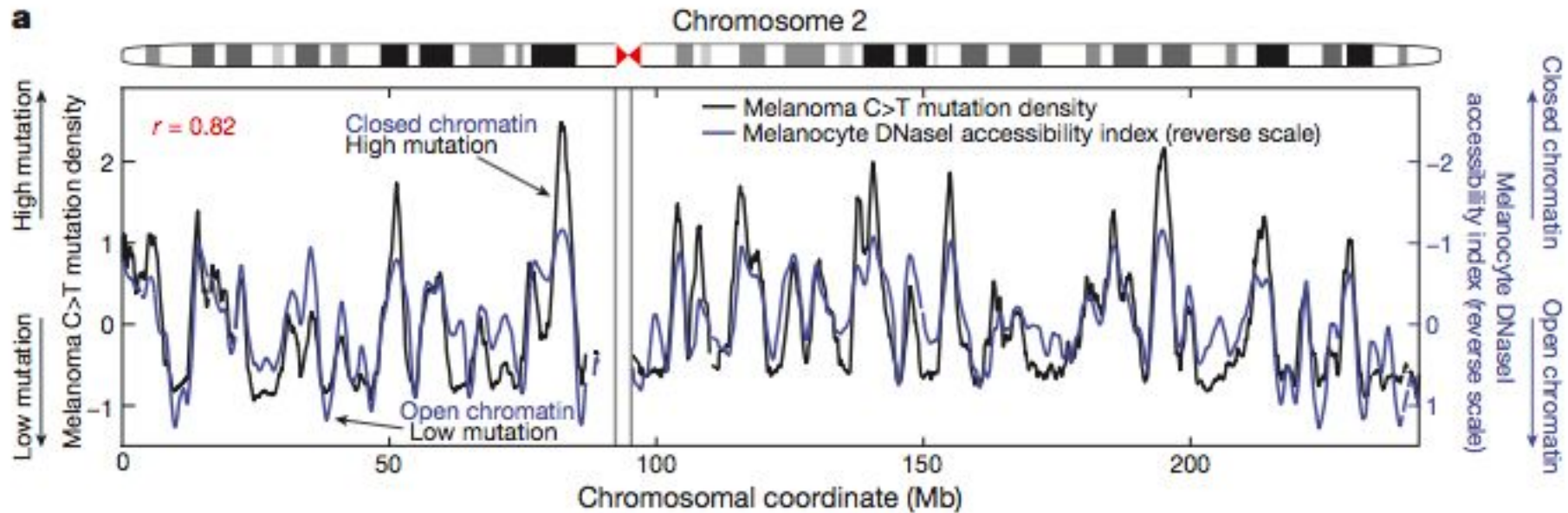
Yonggang Zhou et al. *Circ Res.* 2011;109:1067-1081



# Cell-of-origin chromatin organization shapes the mutational landscape of cancer

Paz Polak<sup>1,2\*</sup>, Rosa Karlič<sup>3\*</sup>, Amnon Koren<sup>2,4</sup>, Robert Thurman<sup>5</sup>, Richard Sandstrom<sup>5</sup>, Michael S. Lawrence<sup>2</sup>, Alex Reynolds<sup>5</sup>, Eric Rynes<sup>5</sup>, Kristian Vlahoviček<sup>3,6</sup>, John A. Stamatoyannopoulos<sup>5,7</sup> & Shamil R. Sunyaev<sup>1,2</sup>

Random Forest model predicts cancer mutation densities from epigenomic mark ups





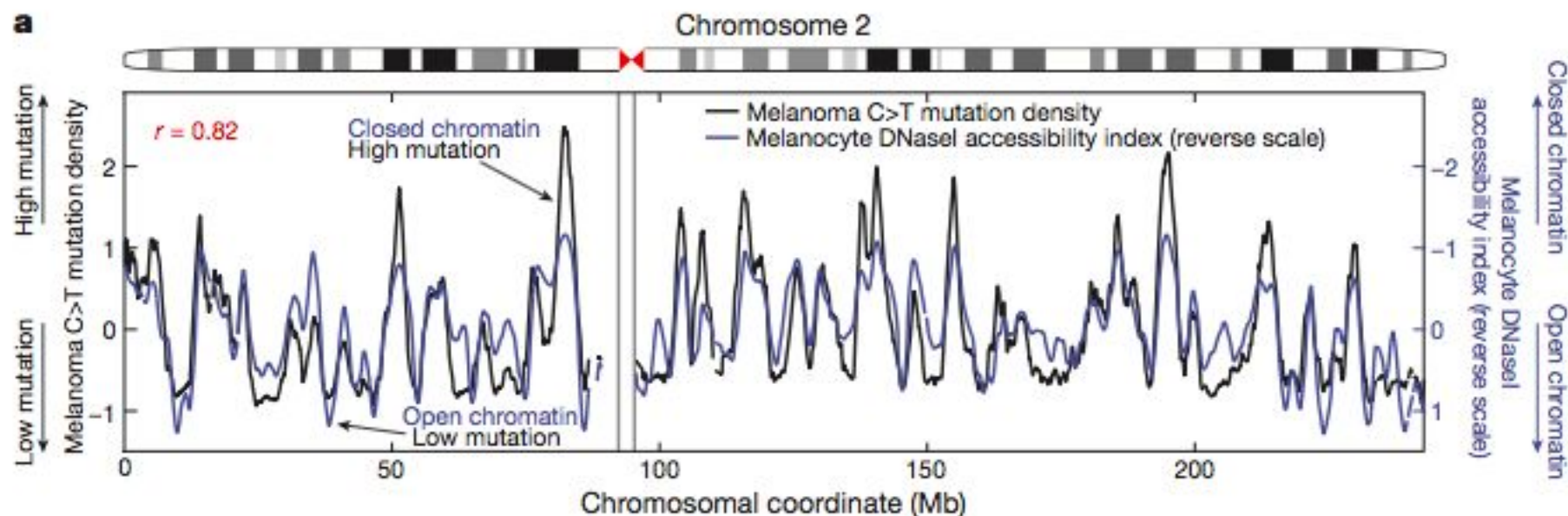
# Cell-of-origin chromatin organization shapes the mutational landscape of cancer

Paz Polak<sup>1,2\*</sup>, Rosa Karlič<sup>3\*</sup>, Amnon Koren<sup>2,4</sup>, Robert Thurman<sup>5</sup>, Richard Sandstrom<sup>5</sup>, Michael S. Lawrence<sup>2</sup>, Alex Reynolds<sup>5</sup>, Eric Rynes<sup>5</sup>, Kristian Vlahoviček<sup>3,6</sup>, John A. Stamatoyannopoulos<sup>5,7</sup> & Shamil R. Sunyaev<sup>1,2</sup>

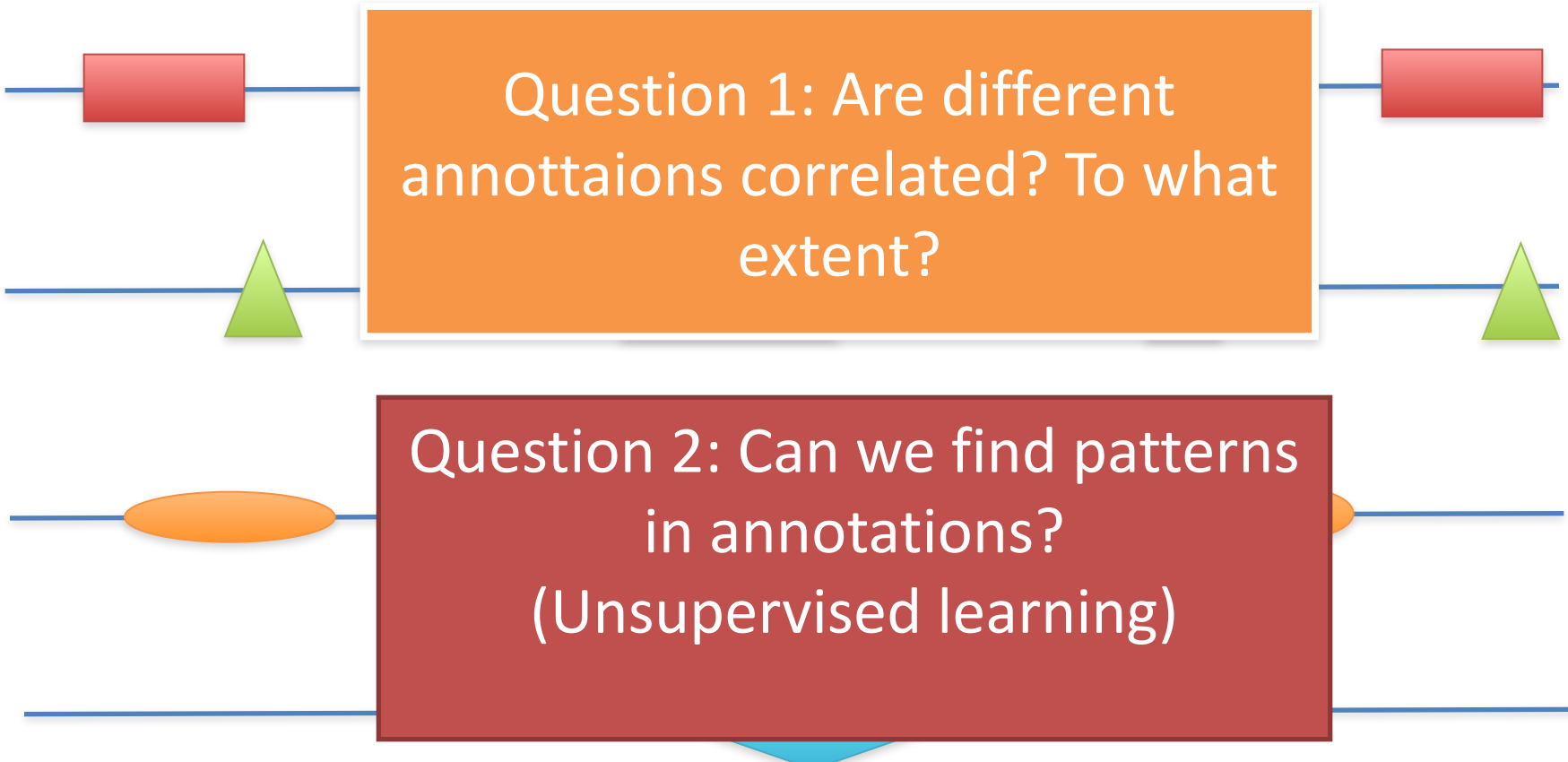
<sup>1</sup>Division of Genetics, Department of Medicine, Brigham & Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>2</sup>The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. <sup>3</sup>Bioinformatics Group, Department of Molecular Biology, Division of Biology, Faculty of Science, University of Zagreb, Horvatovac 102a, 10000 Zagreb, Croatia. <sup>4</sup>Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>5</sup>Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA. <sup>6</sup>Department of Informatics, University of Oslo, P.O. Box 1080, Blindern, NO-0316 Oslo, Norway. <sup>7</sup>Department of Medicine, Division of Oncology, University of Washington, Seattle, Washington 98195, USA.

\*These authors contributed equally to this work.

360 | NATURE | VOL 518 | 19 FEBRUARY 2015



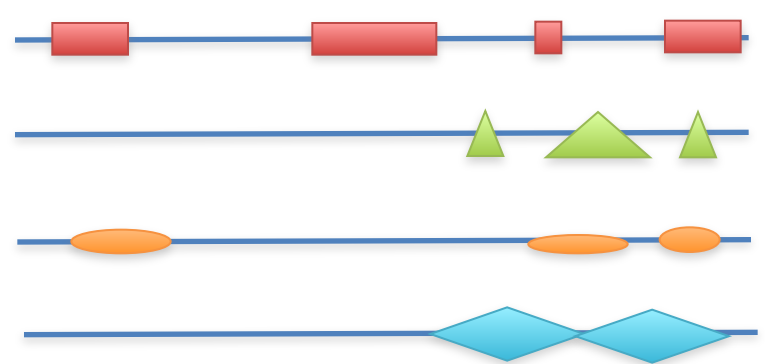
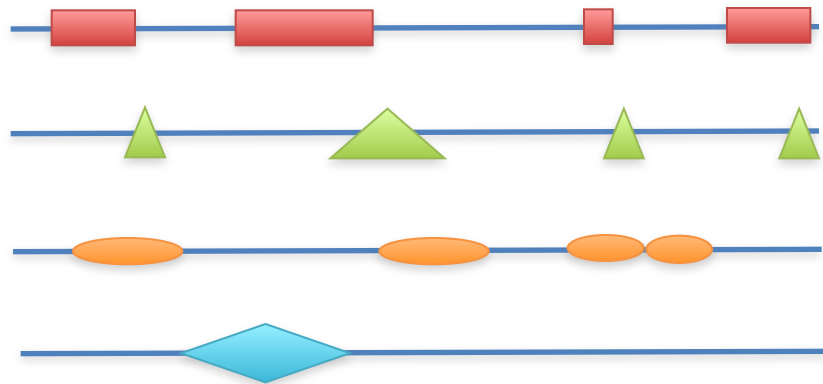
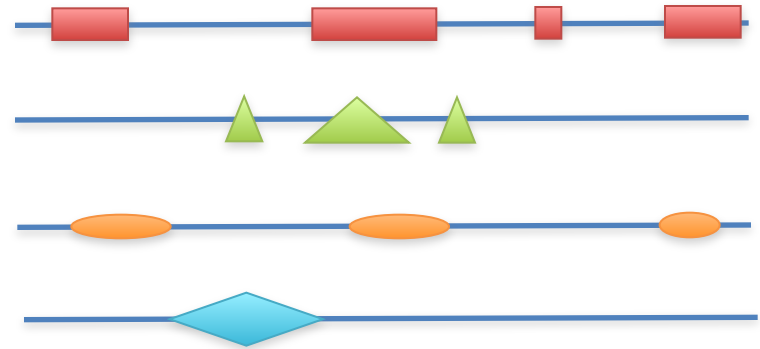
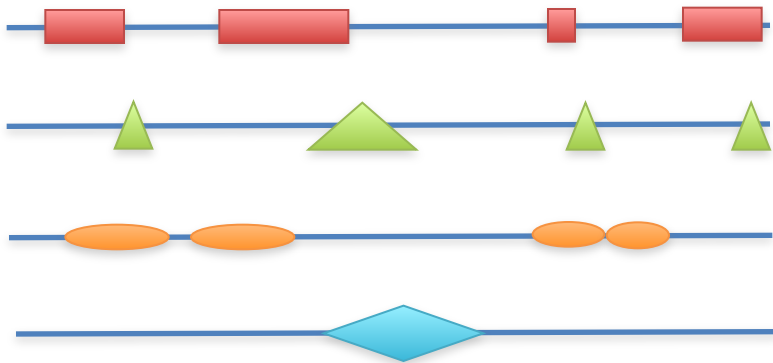
# We have many experimental genome-wide annotations



Question 1: Are different annotations correlated? To what extent?

Question 2: Can we find patterns in annotations?  
(Unsupervised learning)

# Annotations under different conditions





# Как много данных?

- Roadmap Epigenomics  
~ 3 000 полногеномных данных
- ENCODE Encyclopedia of Genomic Elements  
~ 9000 полногеномных данных
- International Cancer Genome Consortium  
~ 20 000 patients (~50 типов рака)
- The Cancer Genome Atlas  
~ patients 11 000 (~33 типа рака)

# Открытые вопросы

- Какие участки кода работают одновременно?
- Как переключать режимы работы клетки?
- Как перепрограммируется код для разных типов тканей?
- Сколько механизмов регуляции существует в клетках (надежда на универсальность)?



Спасибо за  
внимание

