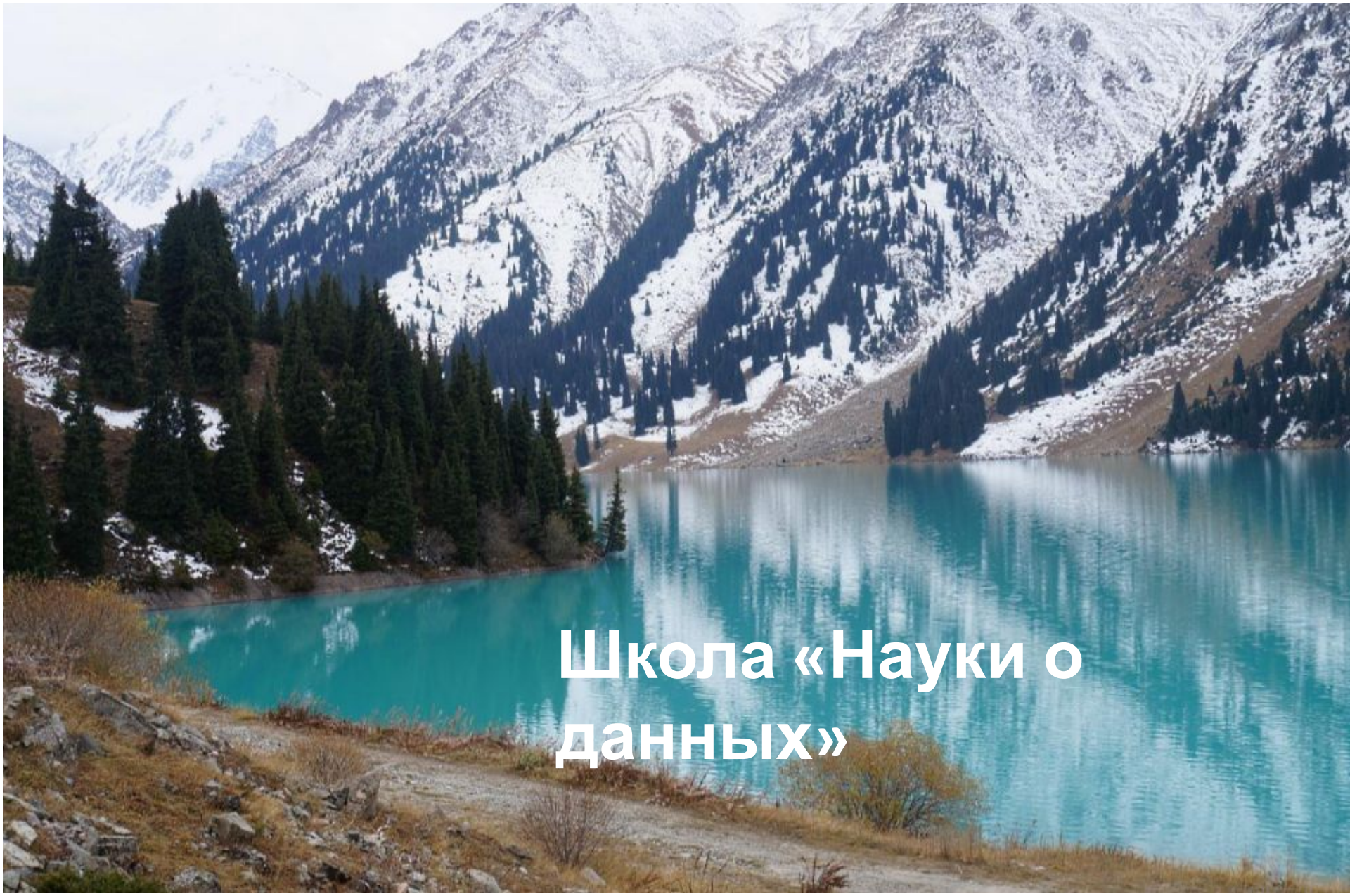
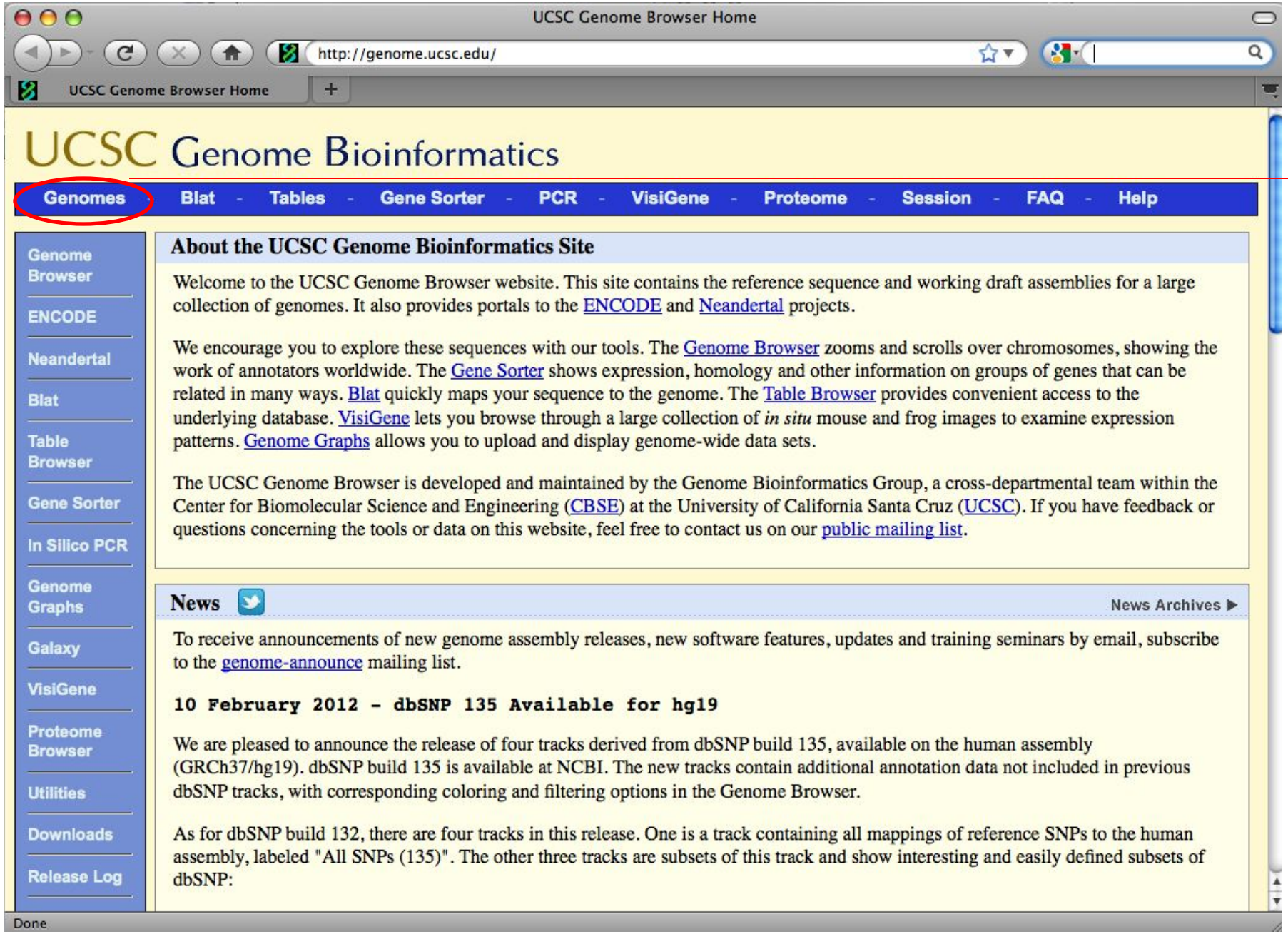
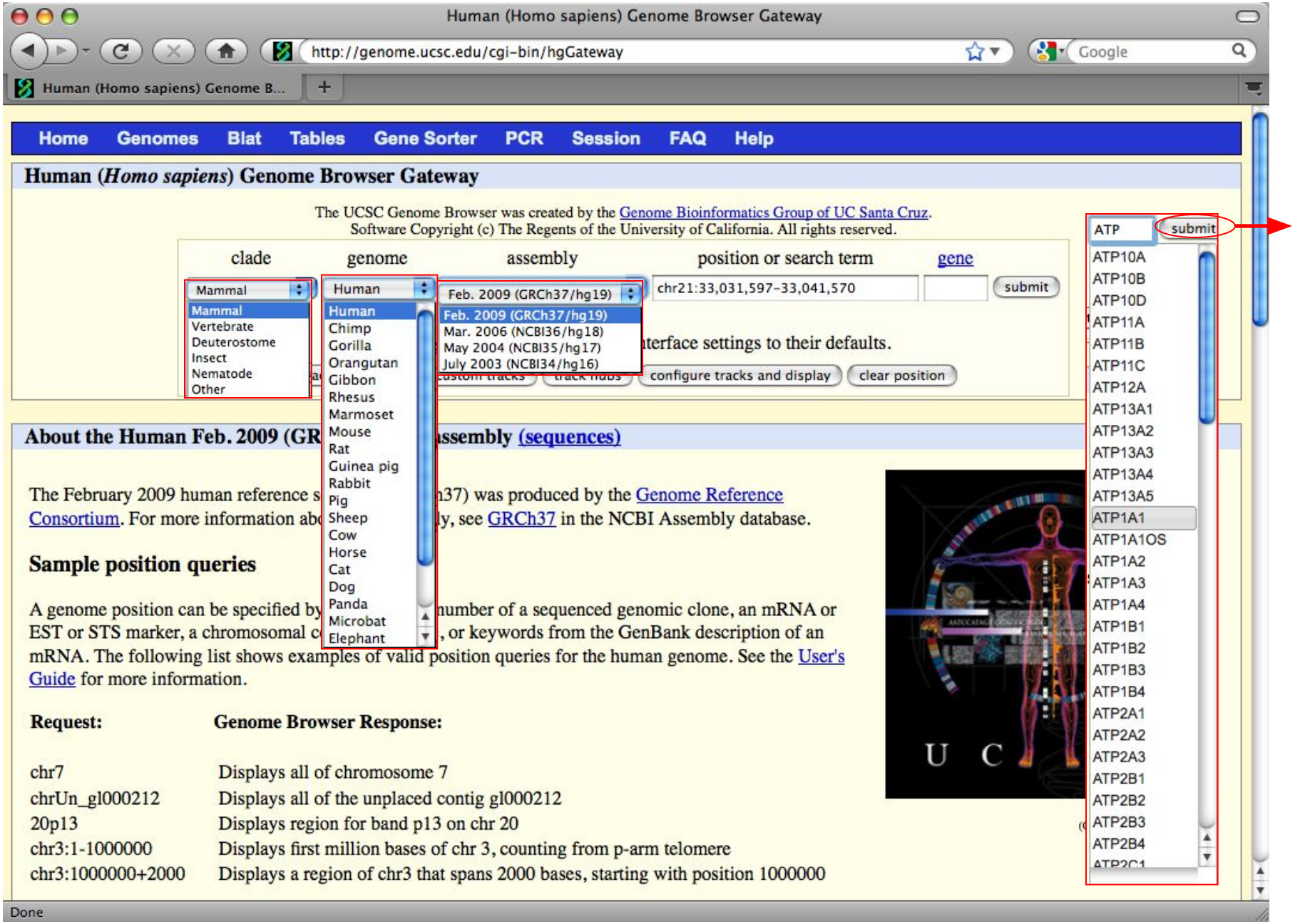


Трек «Биоинформатика» 25-26 апреля

UCSC Genome Browser



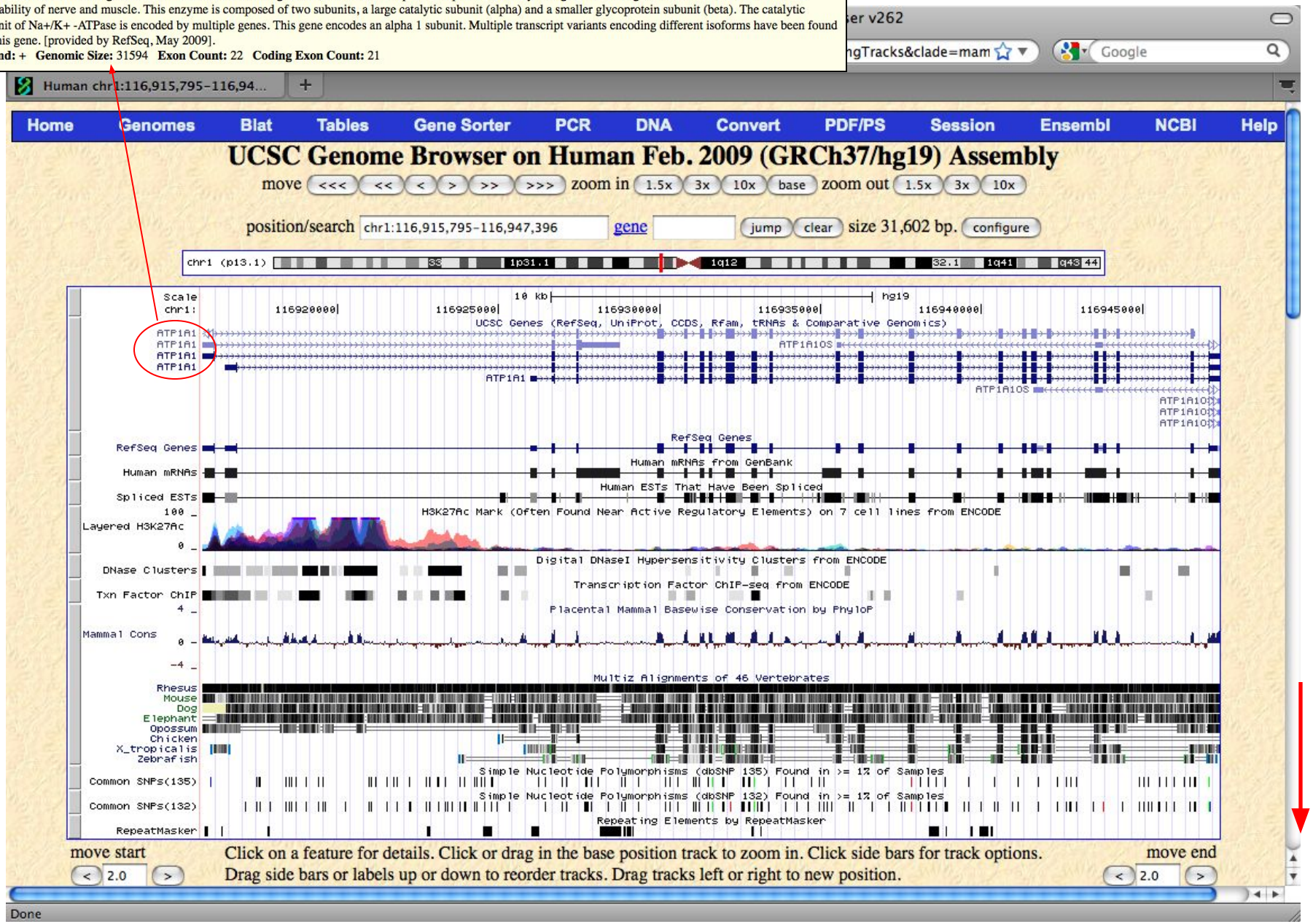
UCSC Genome Browser



UCSC Genome Browser

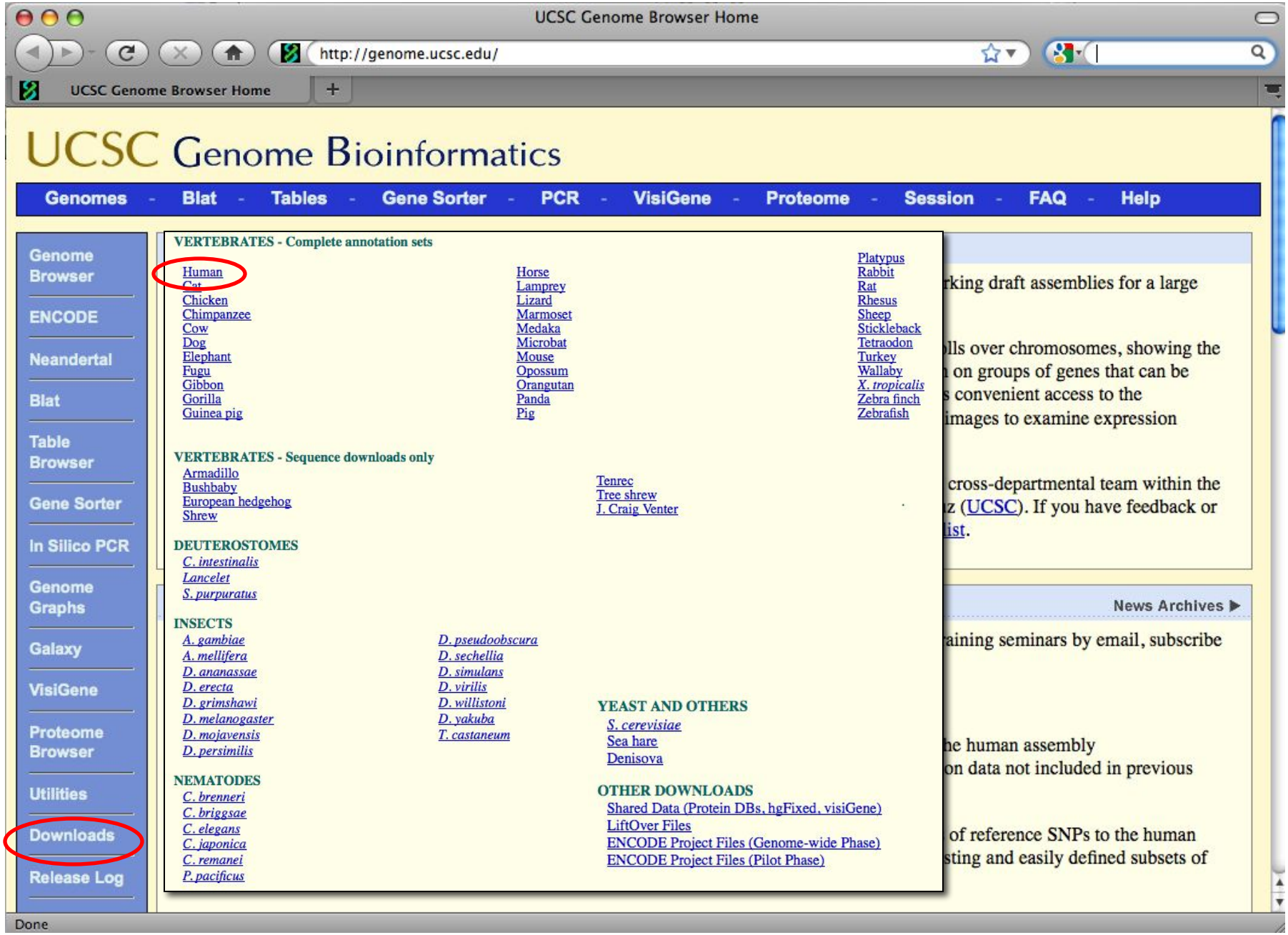
Human Gene ATP1A1 (uc010owv.1) Description and Page Index

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

UCSC Genome Browser



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 (e
  - Human/Elephant
  - Human/Opossum
  - Human/Chicken
  - Human/Turkey (
  - Human/Zebra finch
  - Human/Lizard (anoCar2)
  - Human/Lizard (anoCar1)
  - Human/*X. tropicalis* (xenTro3)
  - Human/*X. tropicalis* (xenTro2)
  - Human/Zebrafish (danRer7)
  - Hu
  - Hu
  - Hu
  - Hu
  - Hu
  - Hu
  - Hu
- 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

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

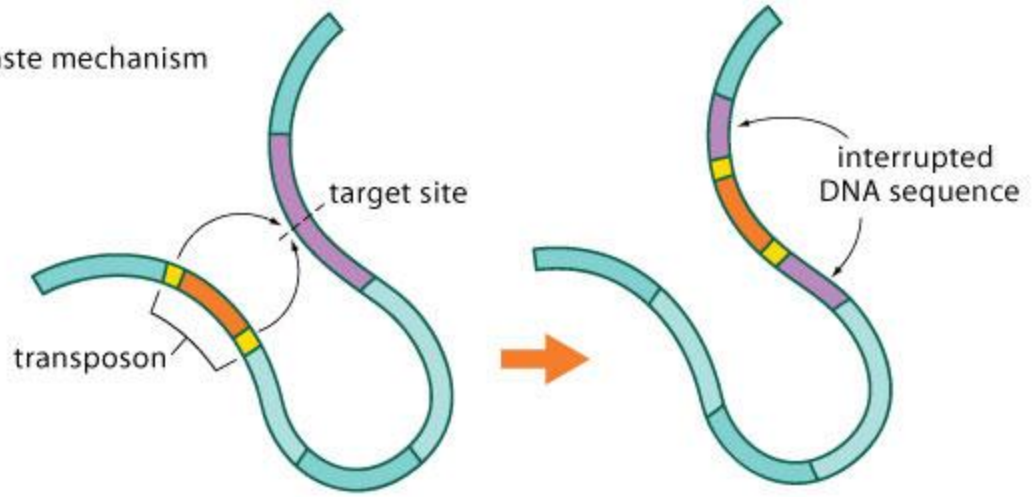
- Barbara McClintock (1902-1992) in 50s.
- Nobel prize in 1983

Gregory G. Dimijian, M.D./Photo Researchers

Two methods of transposition:

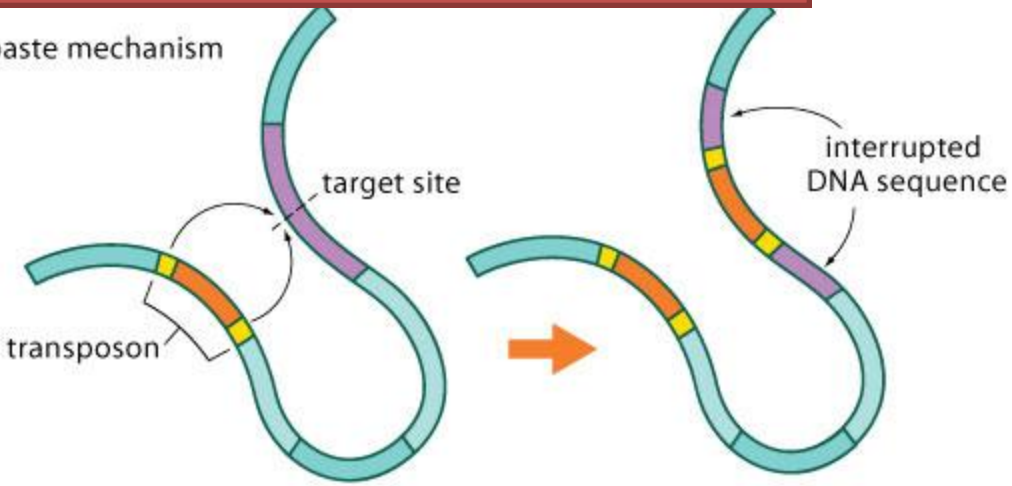
1. Cut-and-paste mechanism

Class II



Class I – retrotransposons (via RNA intermediate)

2. Copy-and-paste mechanism



Схожесть с ретровирусами

Retrovirus reverse transcripiton

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

TABLE 17.1

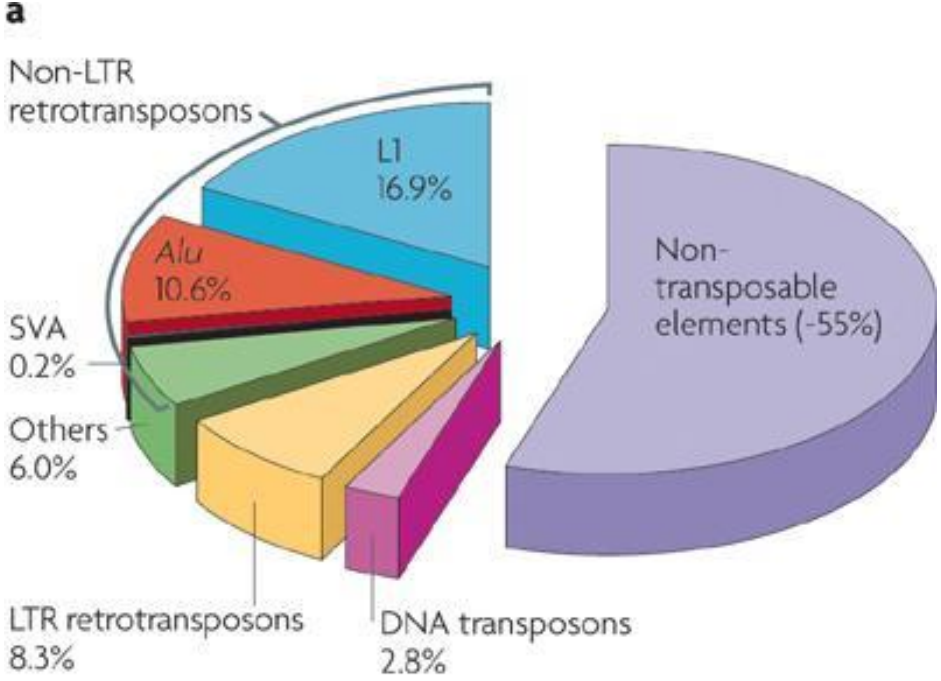
Categorization of Transposable Elements by Transposition Mechanism

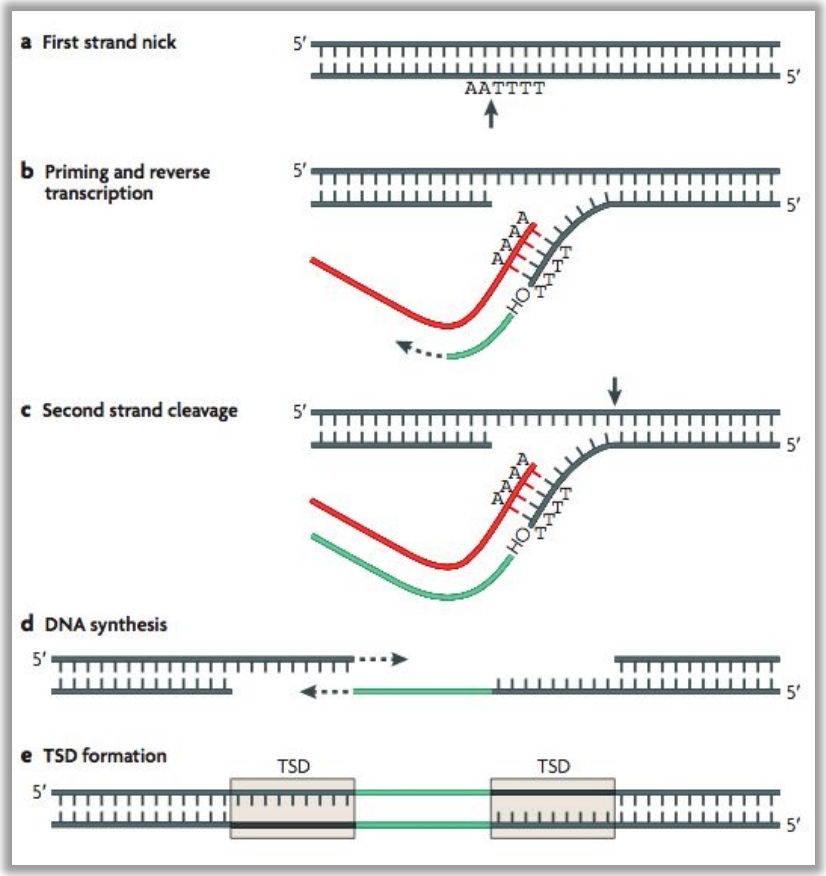
Category	Examples	Host Organism
I. Cut-and-paste transposons	IS elements (e.g., IS*50*)	Bacteria
	Composite transposons (e.g., Tn*5*)	Bacteria
	Ac/*Ds* elements	Maize
	P elements	*Drosophila*
	hobo elements	*Drosophila*
	piggyBac	moth
	Sleeping Beauty	salmon
II. Replicative transposons	Tn*3* elements	Bacteria
III. Retrotransposons		
A. Retroviruslike elements (also called long terminal repeat, or LTR, retrotransposons)	Ty*1*	Yeast
	copia	*Drosophila*
	gypsy	*Drosophila*
B. Retroposons	*F*, *G*, and *I* elements	*Drosophila*
	Telomeric retroposons	*Drosophila*
	LINEs (e.g., *L1*)	Humans
	SINEs (e.g., *Alu*)	Humans

Nature Reviews Genetics **10**, 691-703 (October 2009) | doi:10.1038/

The impact of retrotransposons on human genome evolution

Richard Cordaux & Mark A. Batzer





Review TRENDS in Genetics Vol.23 No.4

Which transposable elements are active in the human genome?

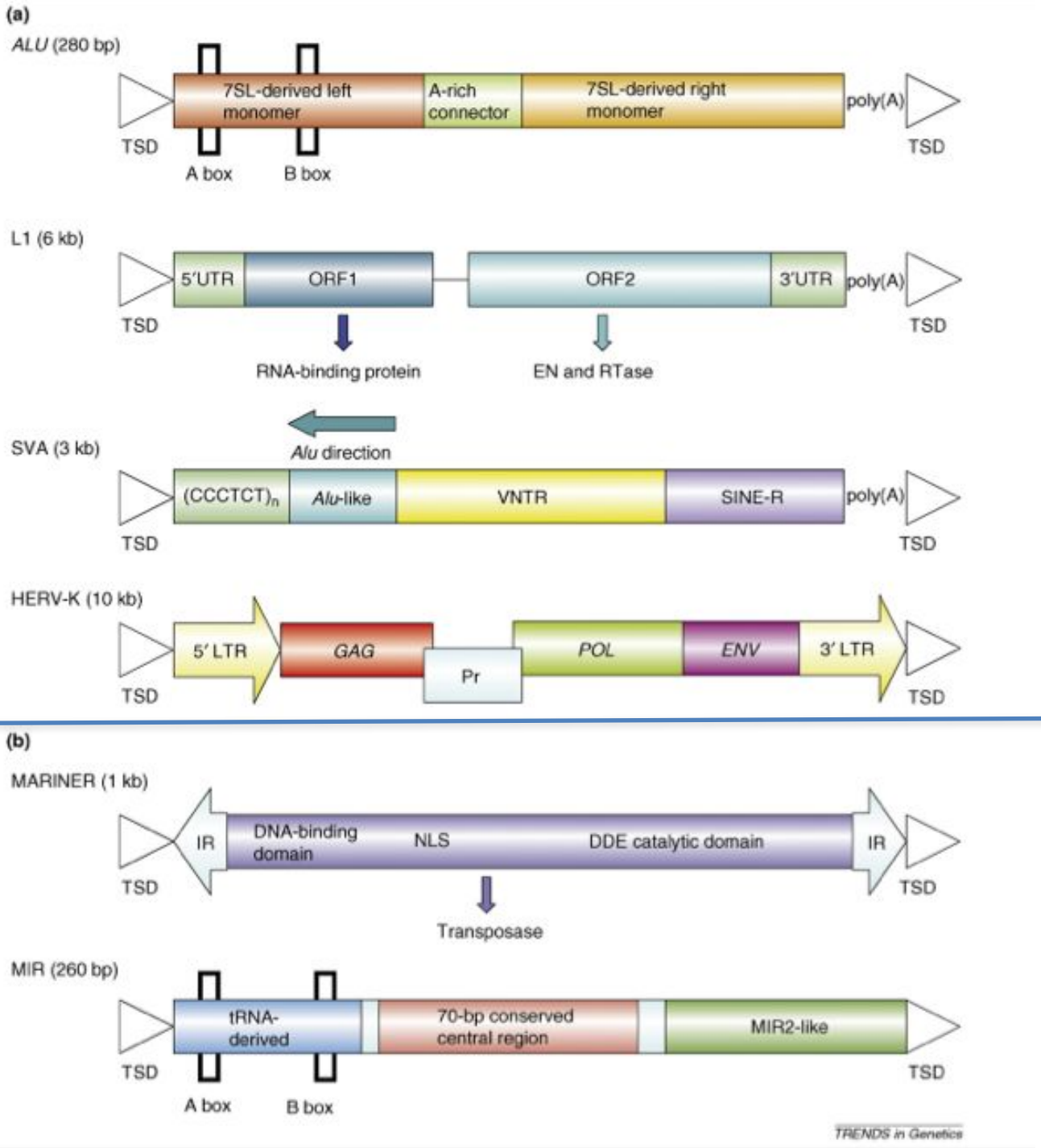
Ryan E. Mills[1,2], E. Andrew Bennett[1,3], Rebecca C. Iskow[1,3] and Scott E. Devine[1,2,3]

[1] Department of Biochemistry, Emory University School of Medicine, Atlanta, GA 30322, USA
[2] Center for Bioinformatics, Emory University School of Medicine, Atlanta, GA 30322, USA
[3] Genetics and Molecular Biology Graduate Program, Emory University School of Medicine, Atlanta, GA 30322, USA

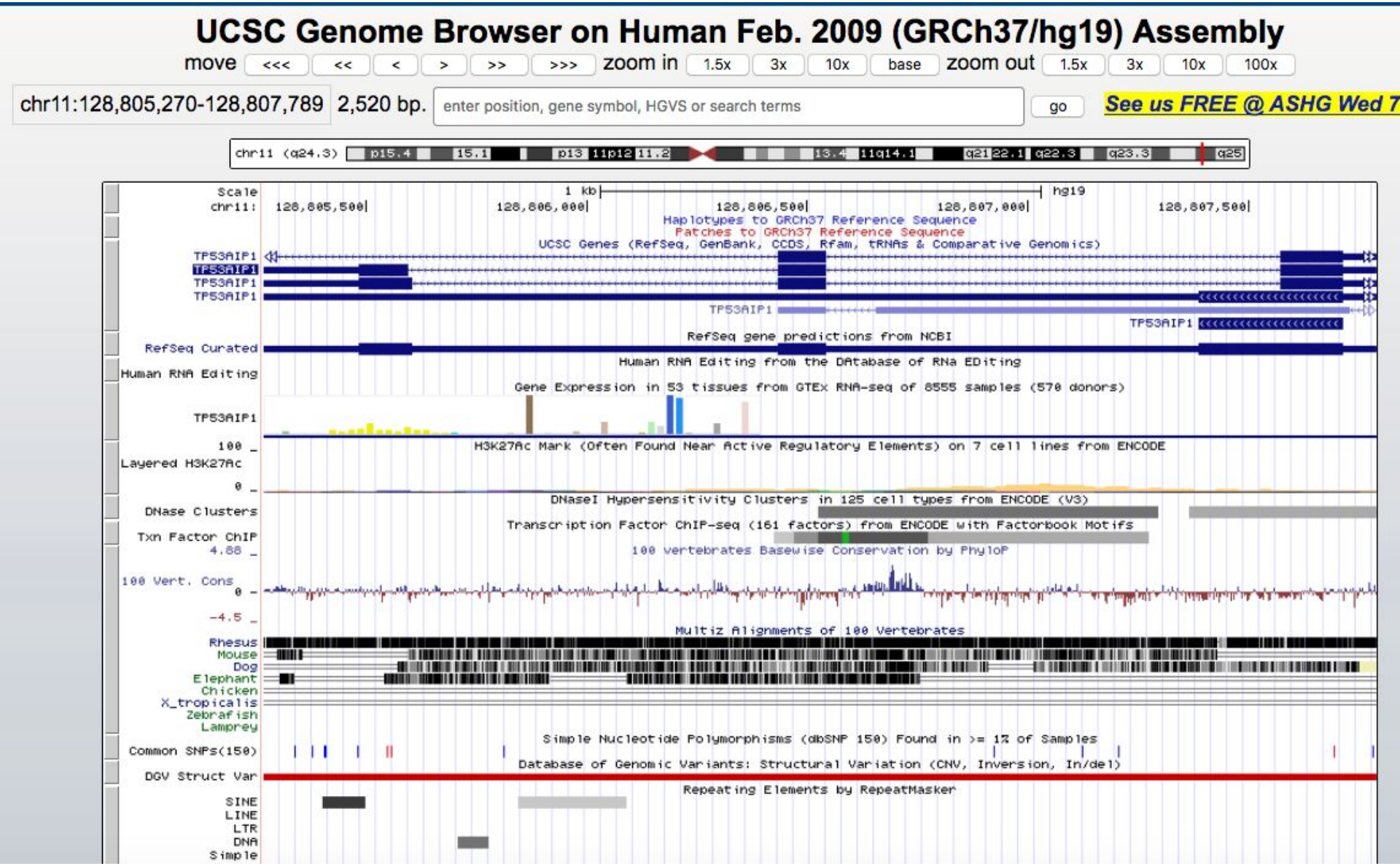
Although a large proportion (44%) of the human genome is occupied by transposons and transposon-like repetitive elements, only a small proportion (<0.05%) of these elements remain active today. Recent evidence indicates that ~35–40 subfamilies of *Alu*, L1 and SVA elements (and possibly HERV-K elements) 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

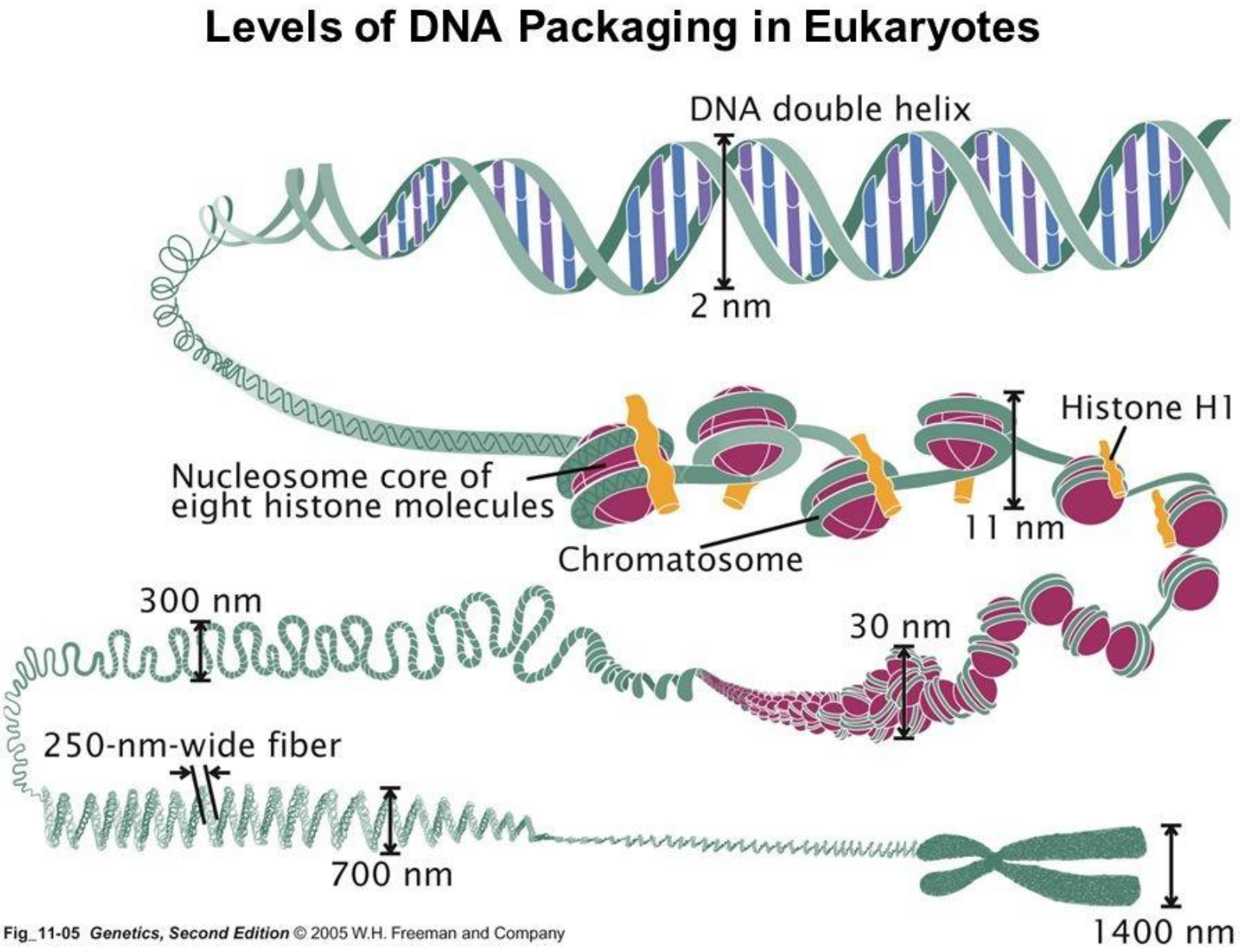
Non-Active



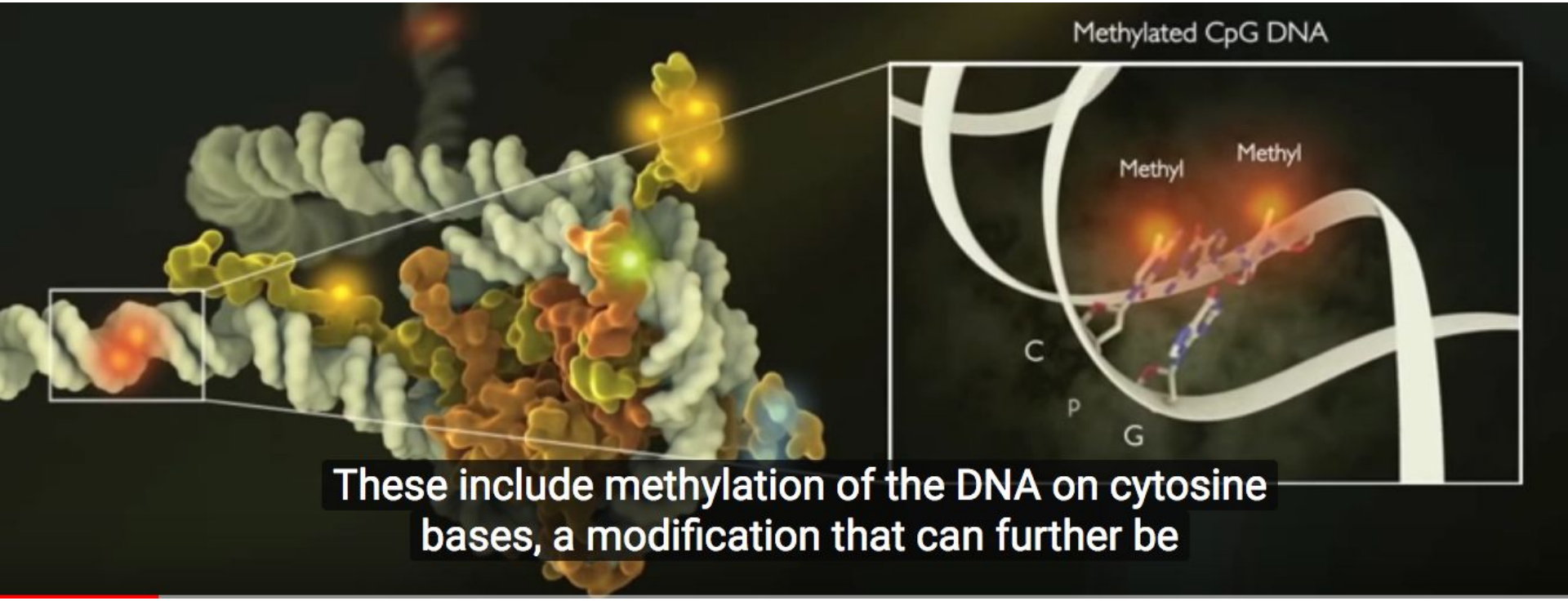
First Layer of Genome Annotation



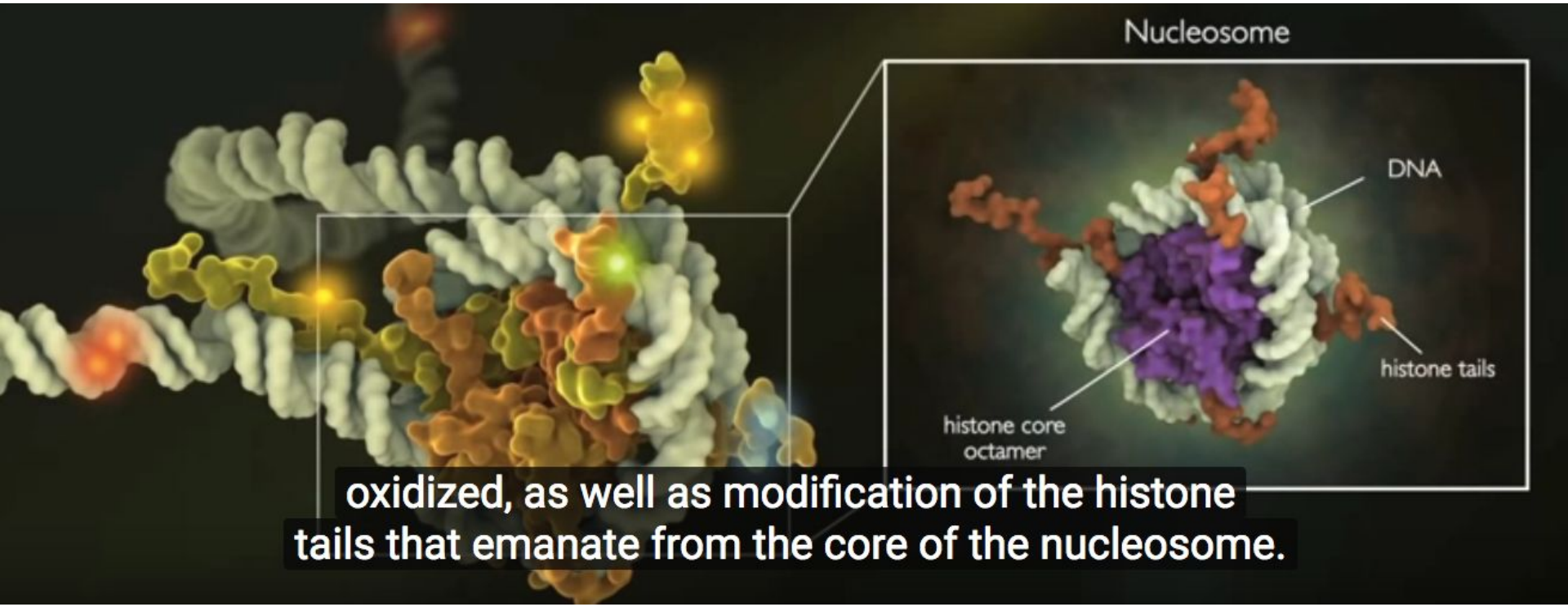
Epigenetics



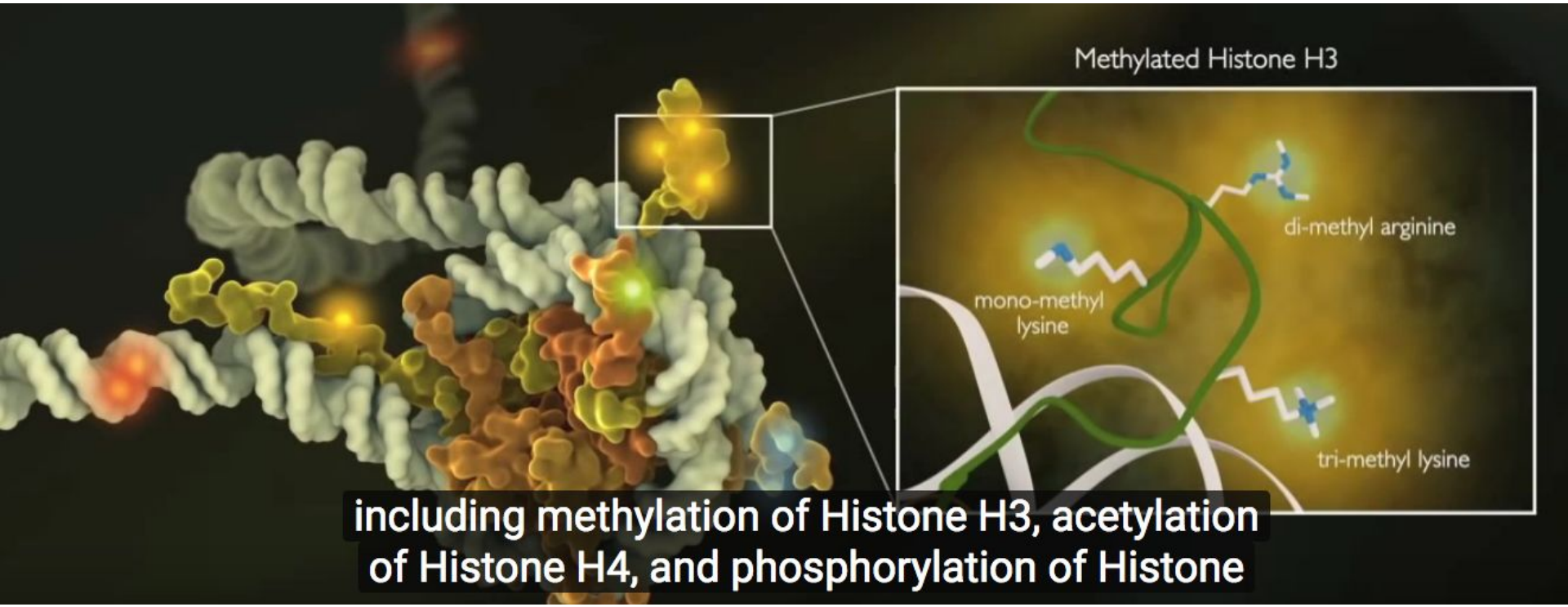
Second Layer of Genome Annotation



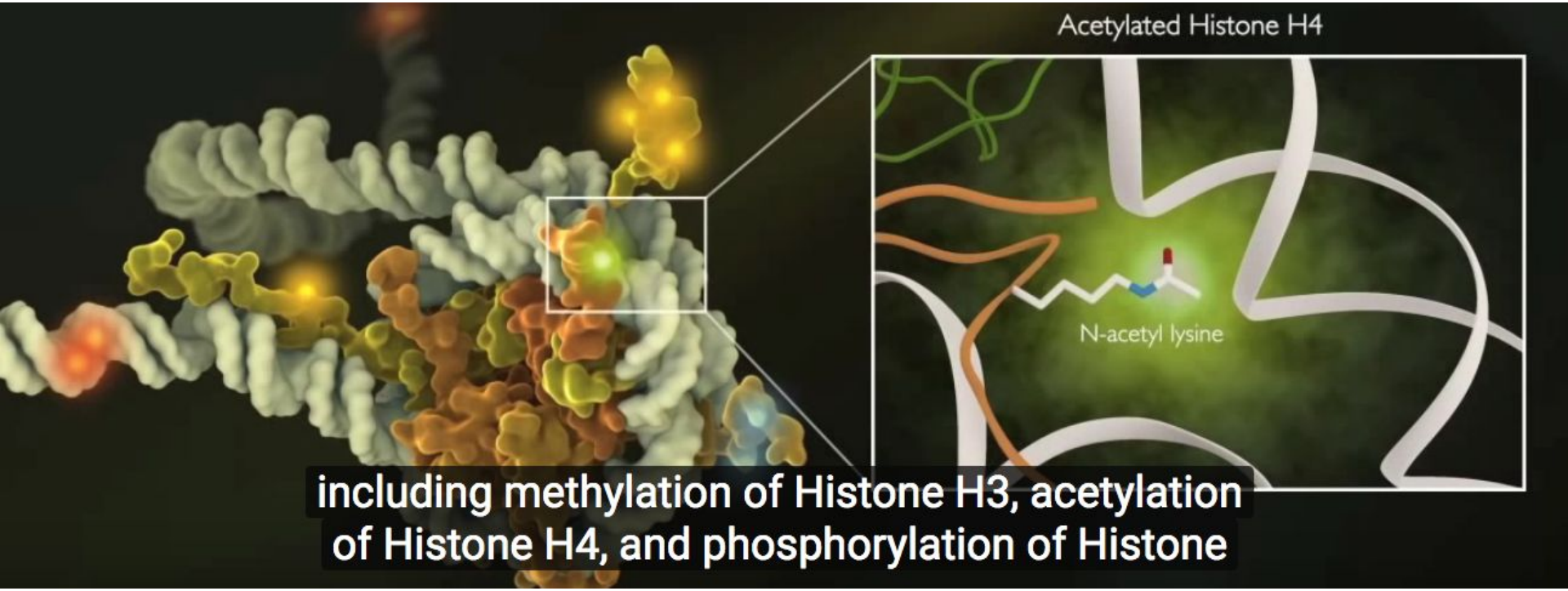
Second Layer of Genome Annotation



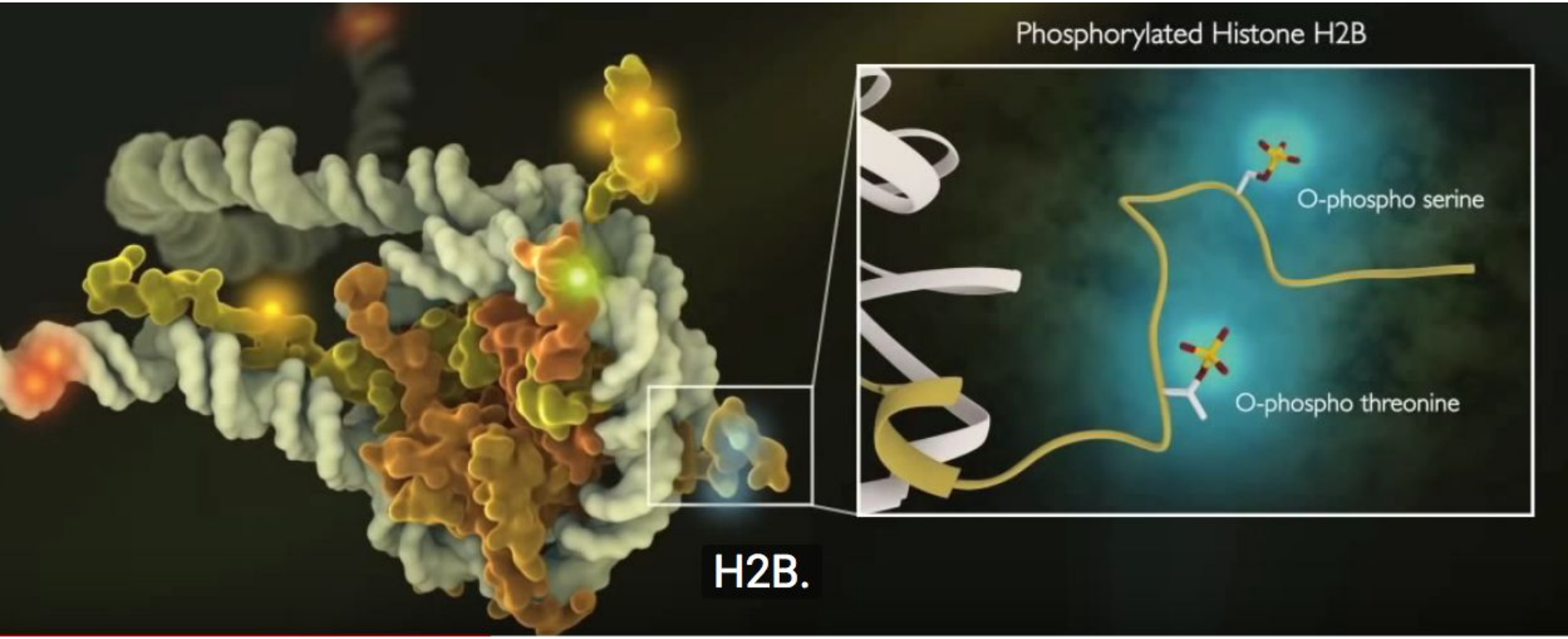
Second Layer of Genome Annotation



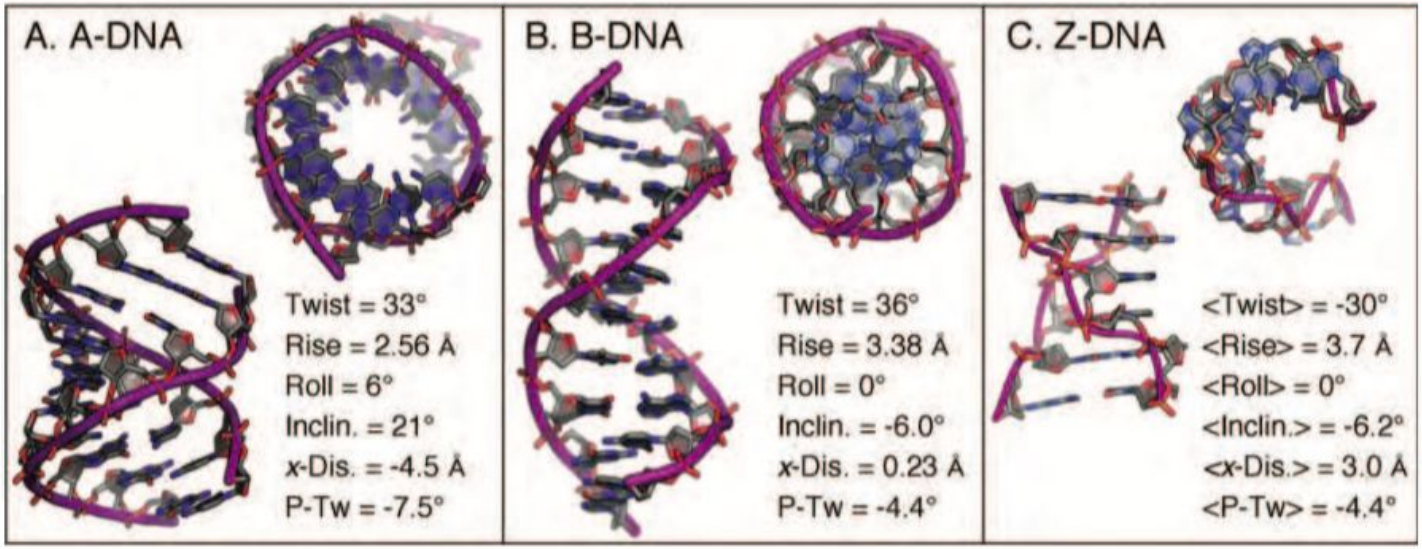
Second Layer of Genome Annotation

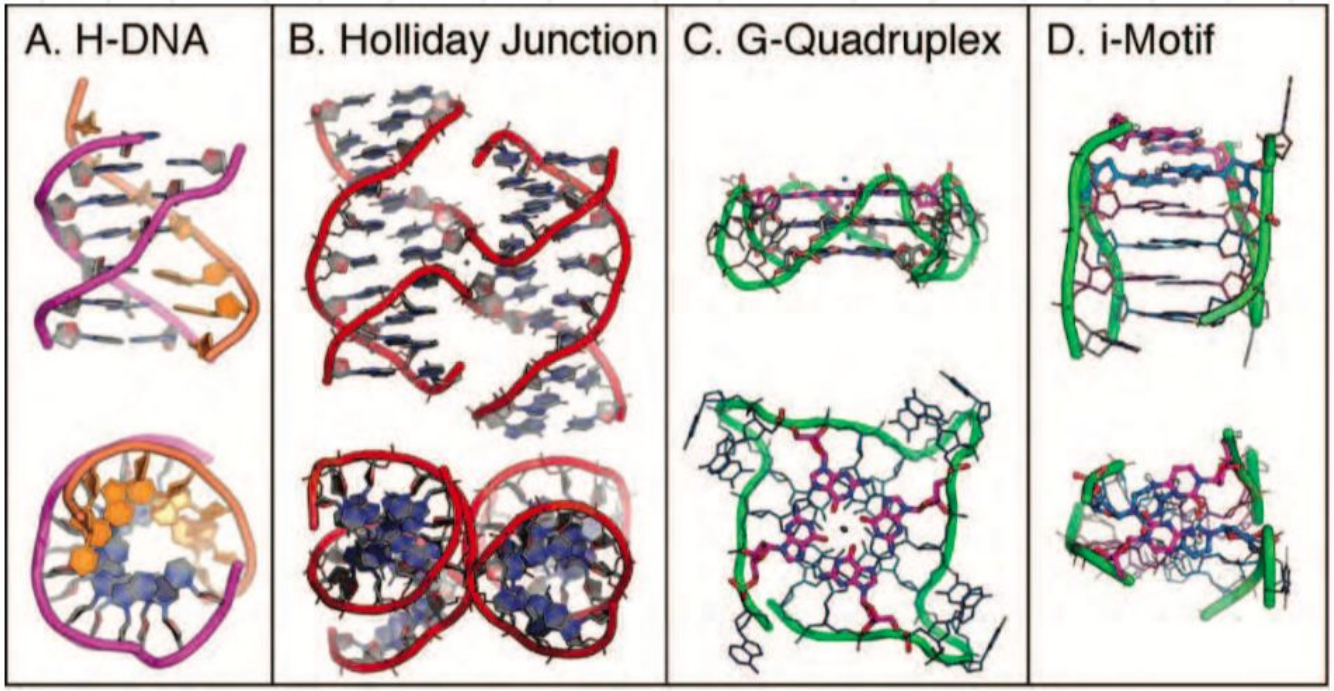


Second Layer of Genome Annotation



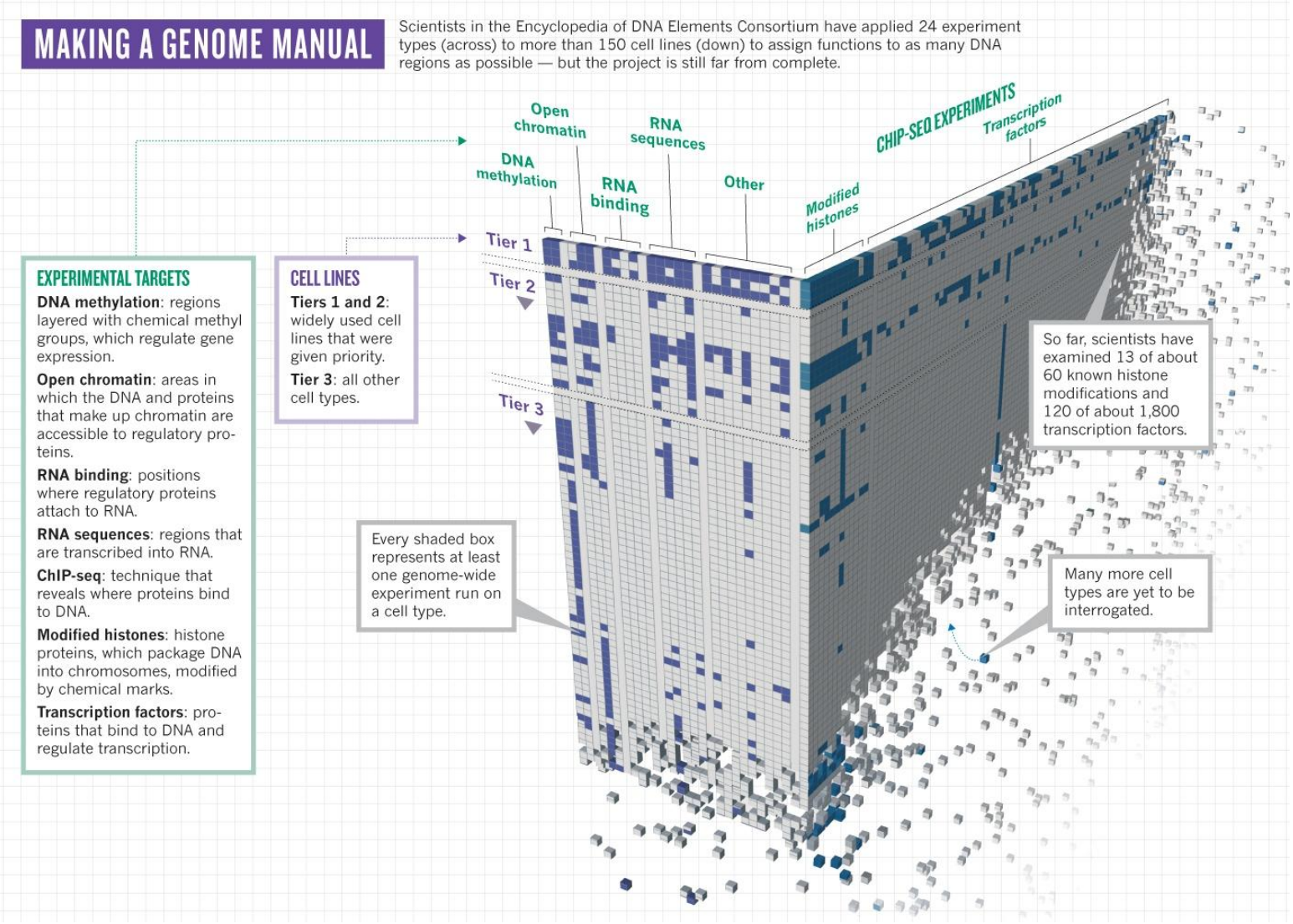
Third Layer of Genome Annotation





Data Accumulation

ENCODE: Encyclopedia of DNA Elements

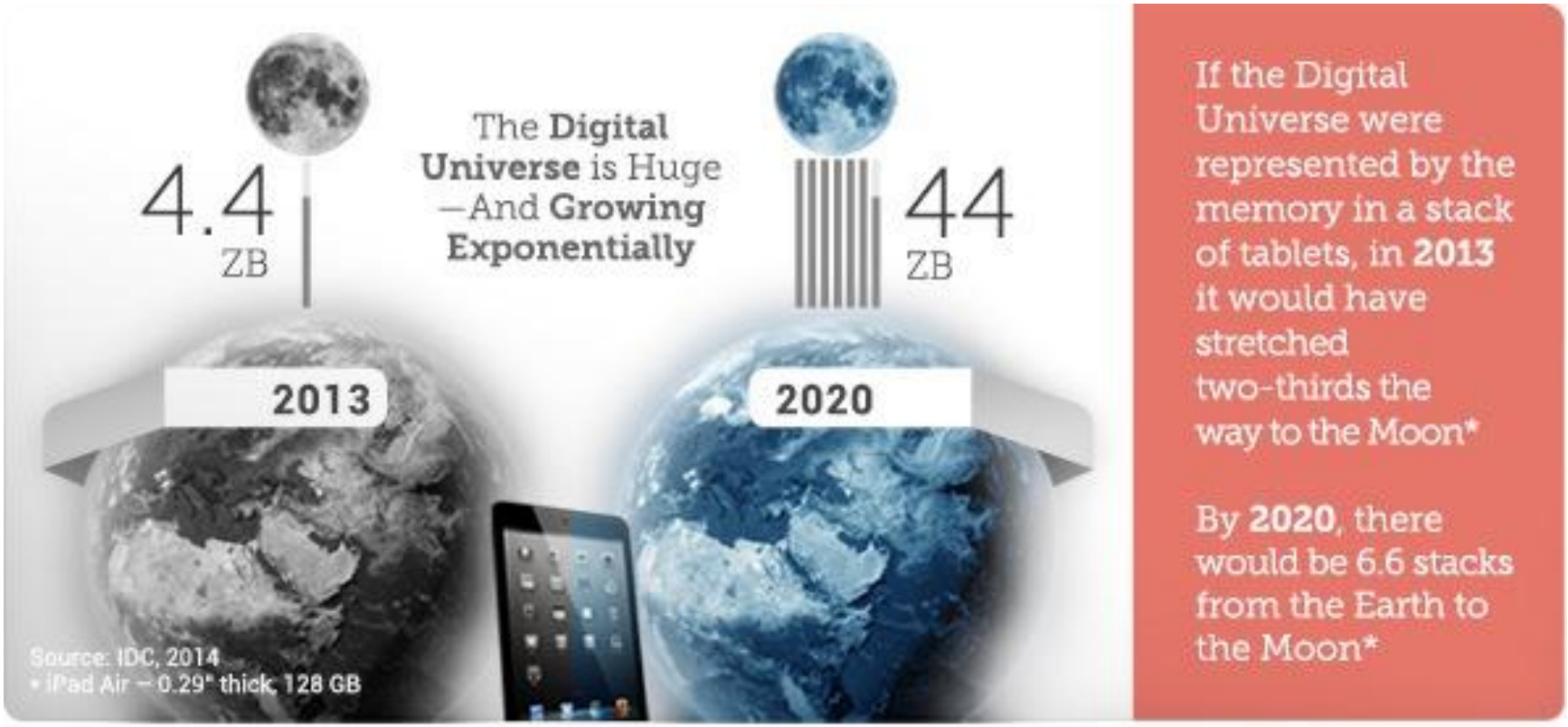


Digital Universe

- 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

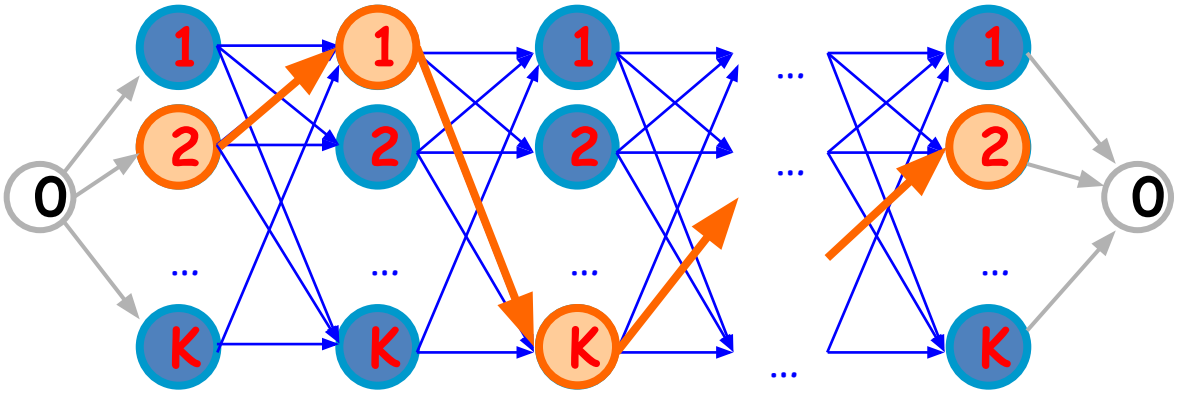
Data Universe Will Expand To 44 Trillion GBs By 2020

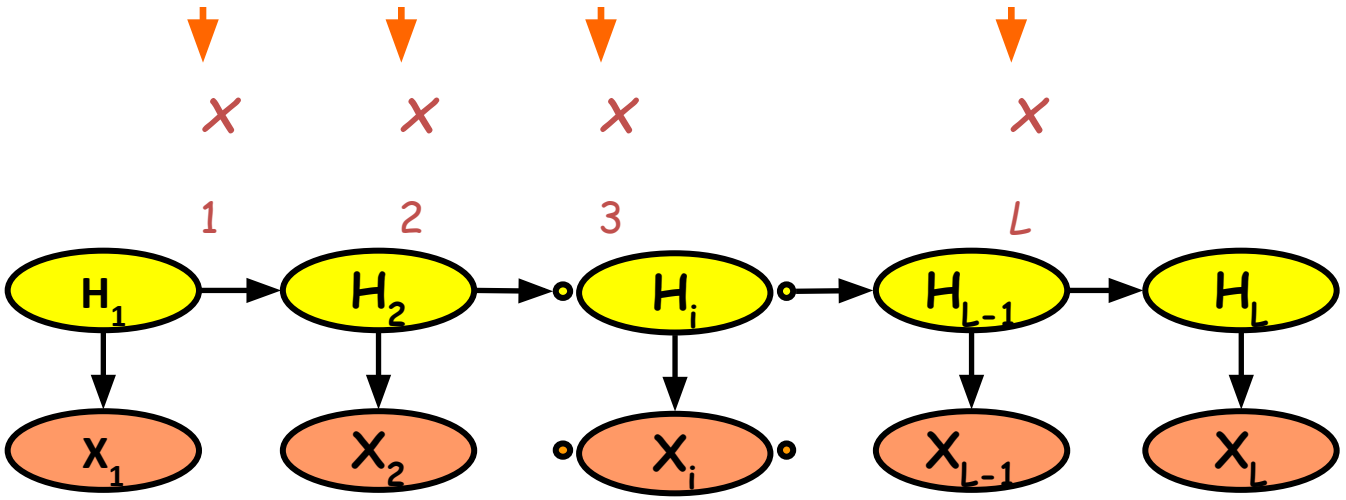


Что делать?

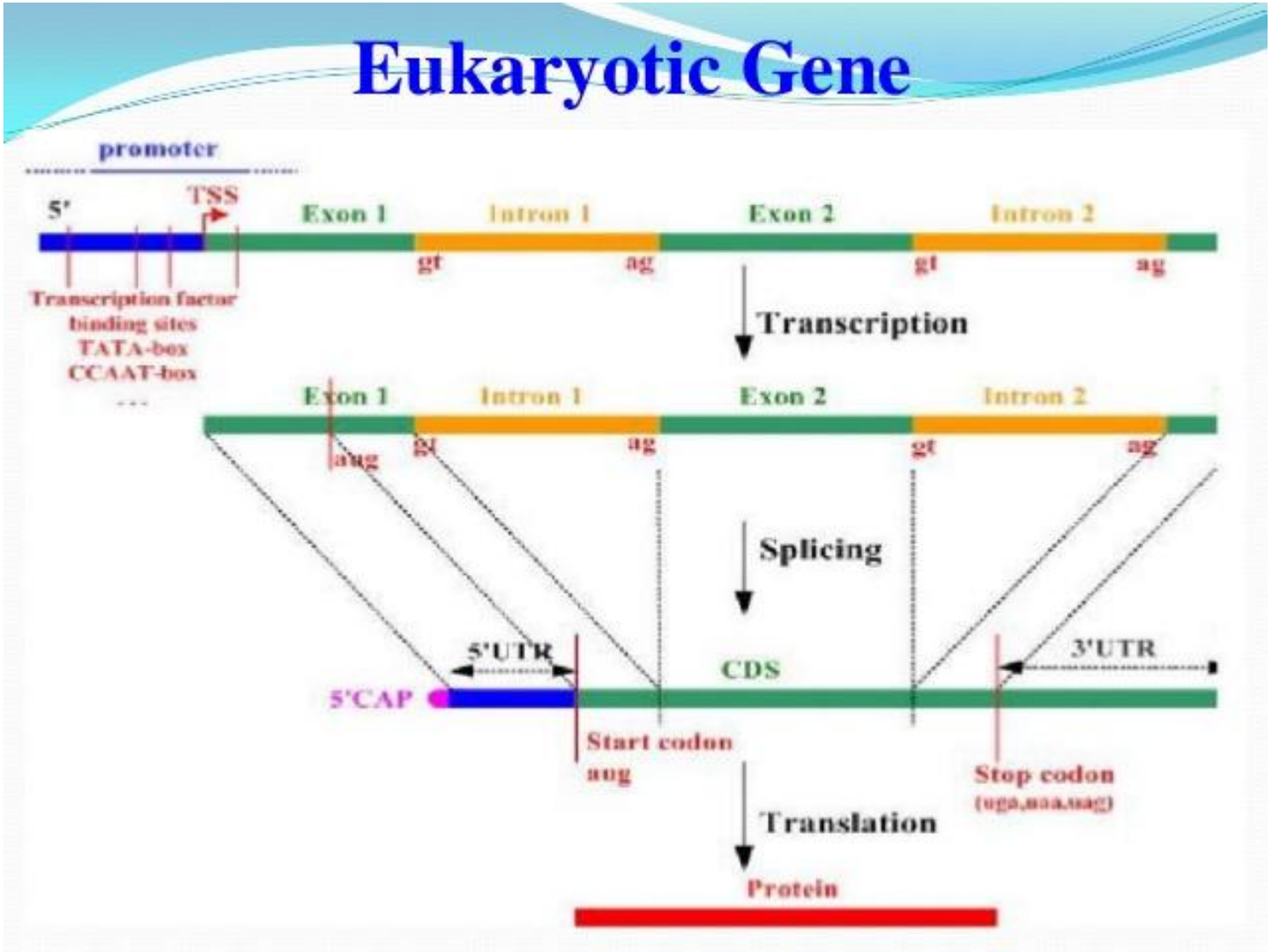
Что получилось?
(Success Stories)

СКРЫТЫЕ ЦЕПИ МАРКОВА

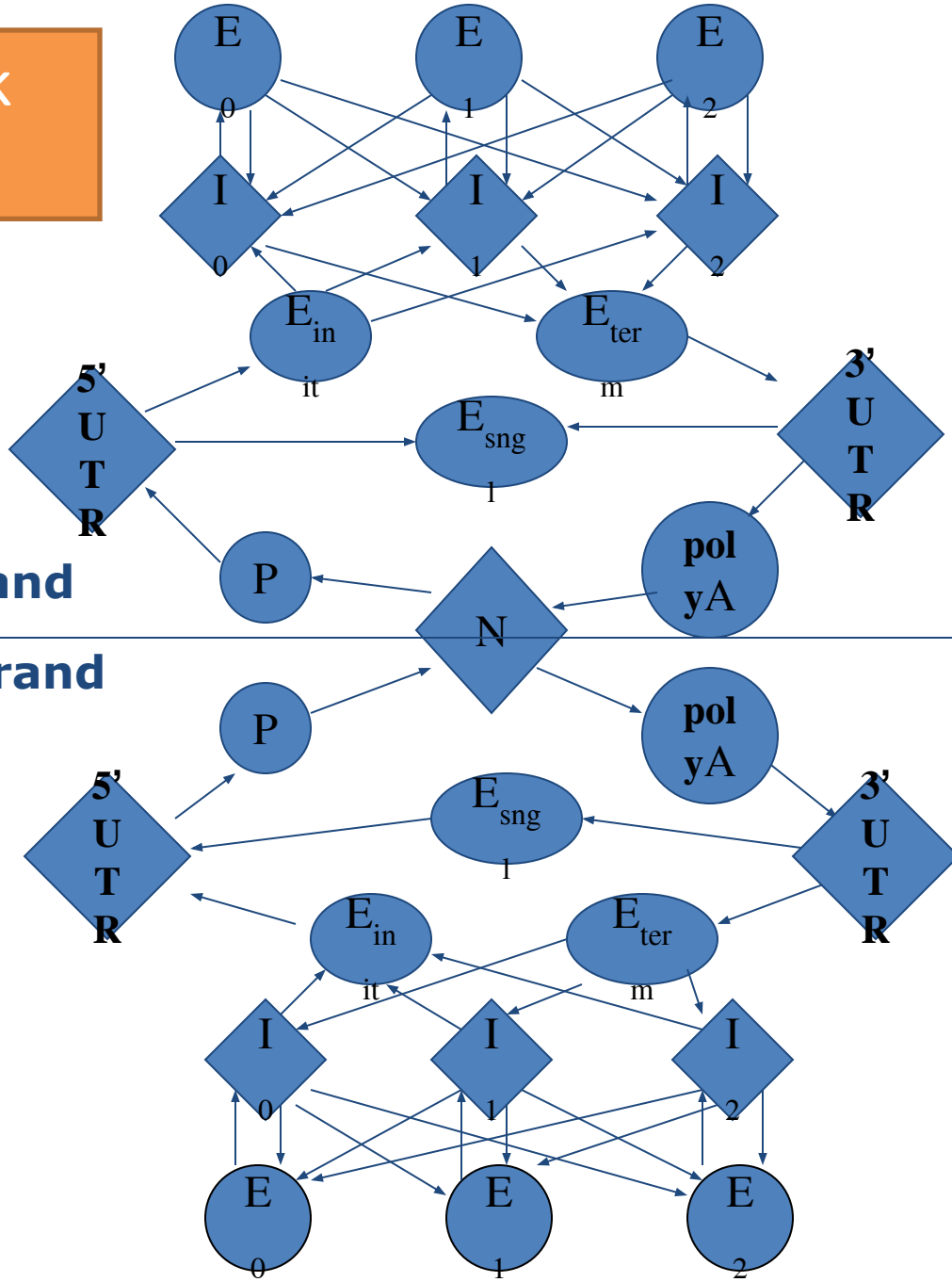




Gene Prediction



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

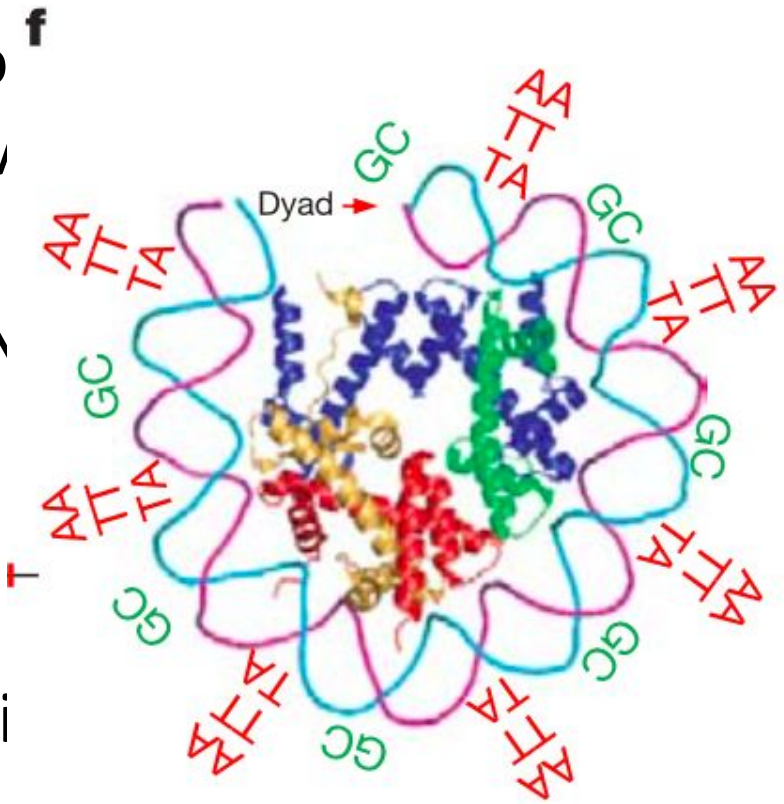
Yoseph Barash[1,2]*, John A. Calarco[2]*, Weijun Gao[1], Qun Pan[2], Xinchen Wang[1,2], Ofer Shai[1], Benjamin J. Blencowe[2] & Brendan J. Frey[1,2,3]

- 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

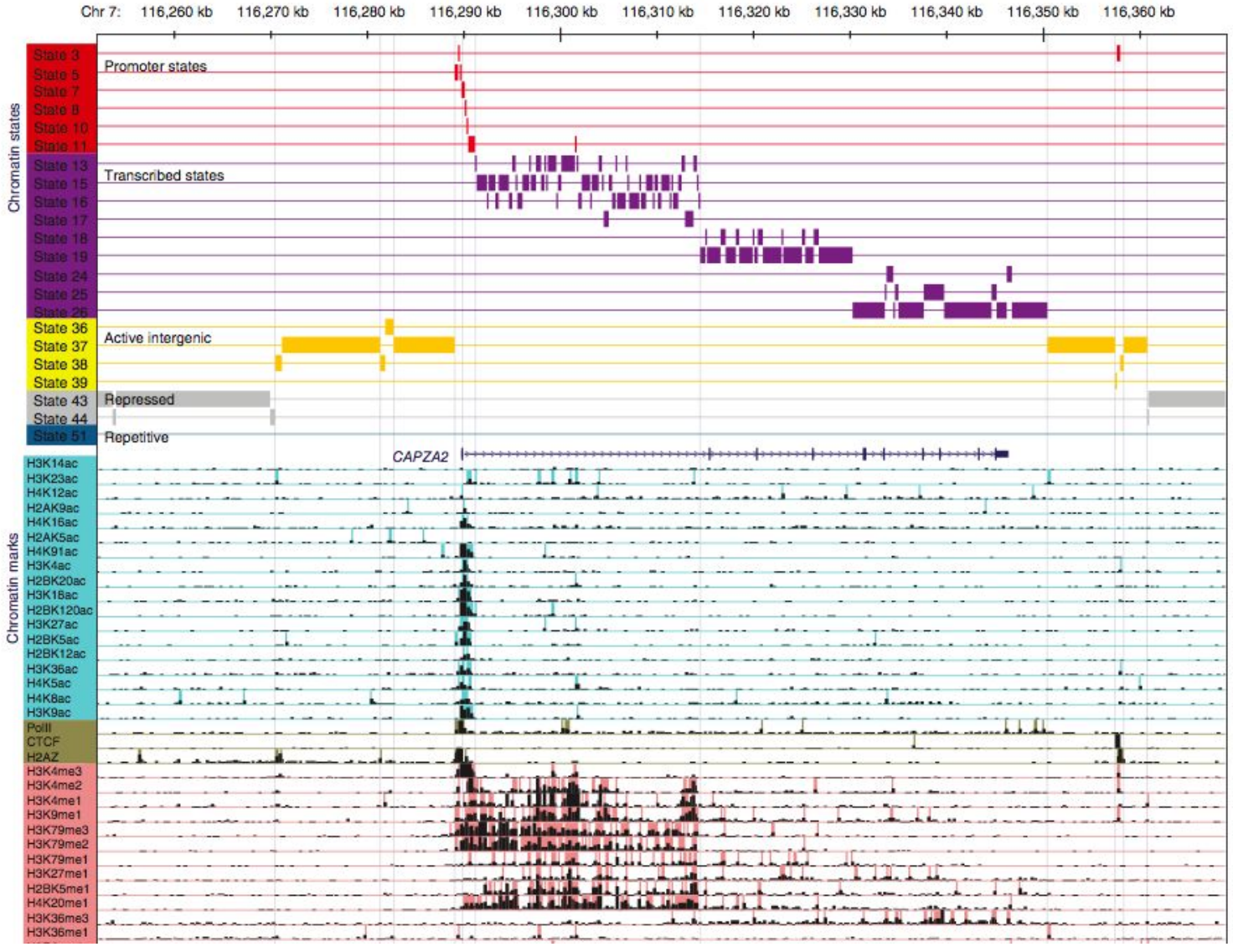
Eran Segal[1], Yvonne Fondufe-Mittendorf[2], Lingyi Chen[2], AnnChristine Thåström[2], Yair Field[1], Irene K. Moore[2], Ji-Ping Z. Wang[3] & Jonathan Widom[2]

- Genome intrinsic organizatio can explain ,50% of the in viv nucleosome positions
- Probabilistic nucleosome–DN interaction model - built on dinucleotide distrubution
- Thermodynamic model for predicting nucleosome positi genome-wide.

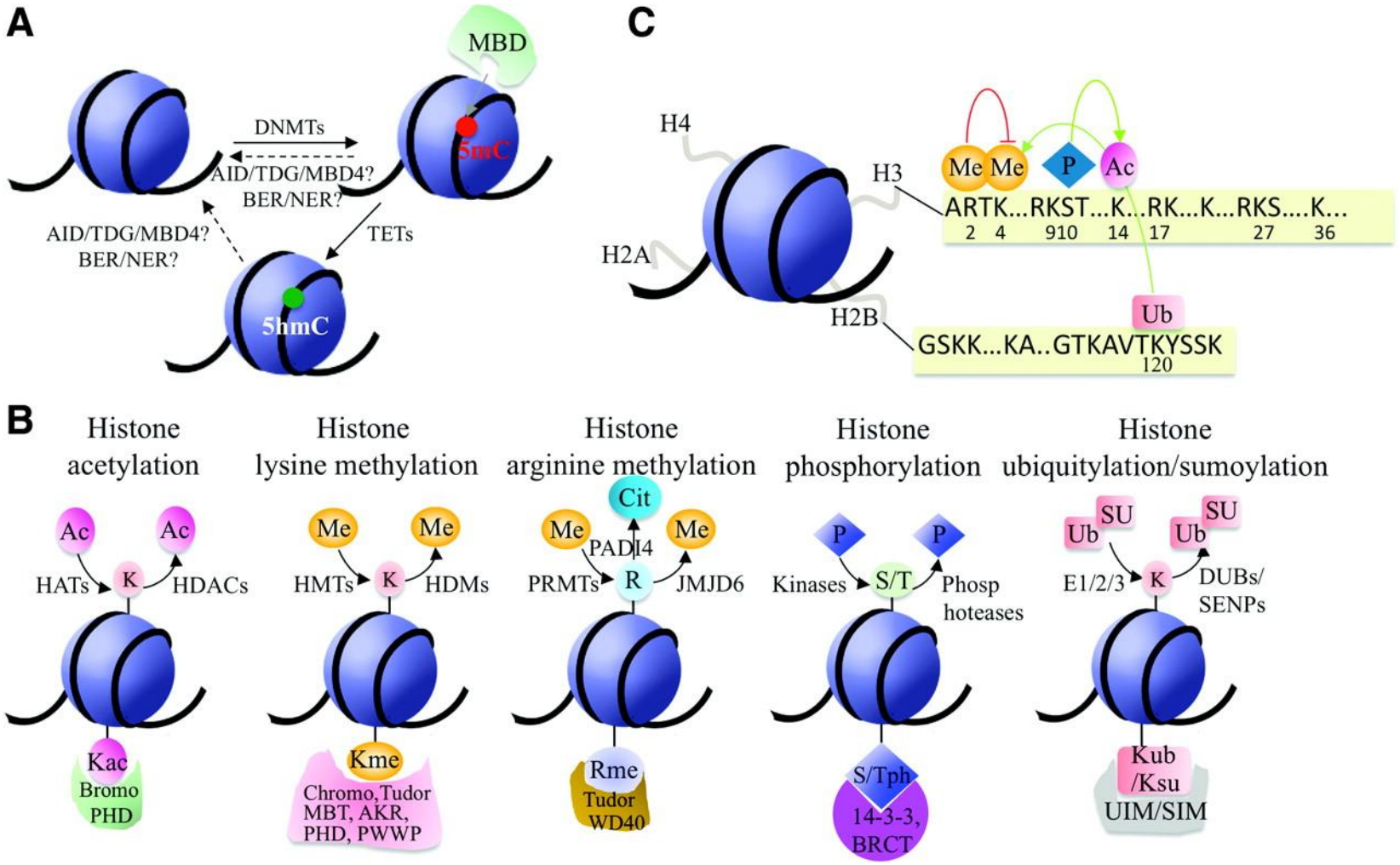


Discovery and characterization of chromatin states for systematic annotation of the human genome

Jason Ernst[1,2] & Manolis Kellis[1,2]



Schematic overview of epigenetic regulatory mechanisms.

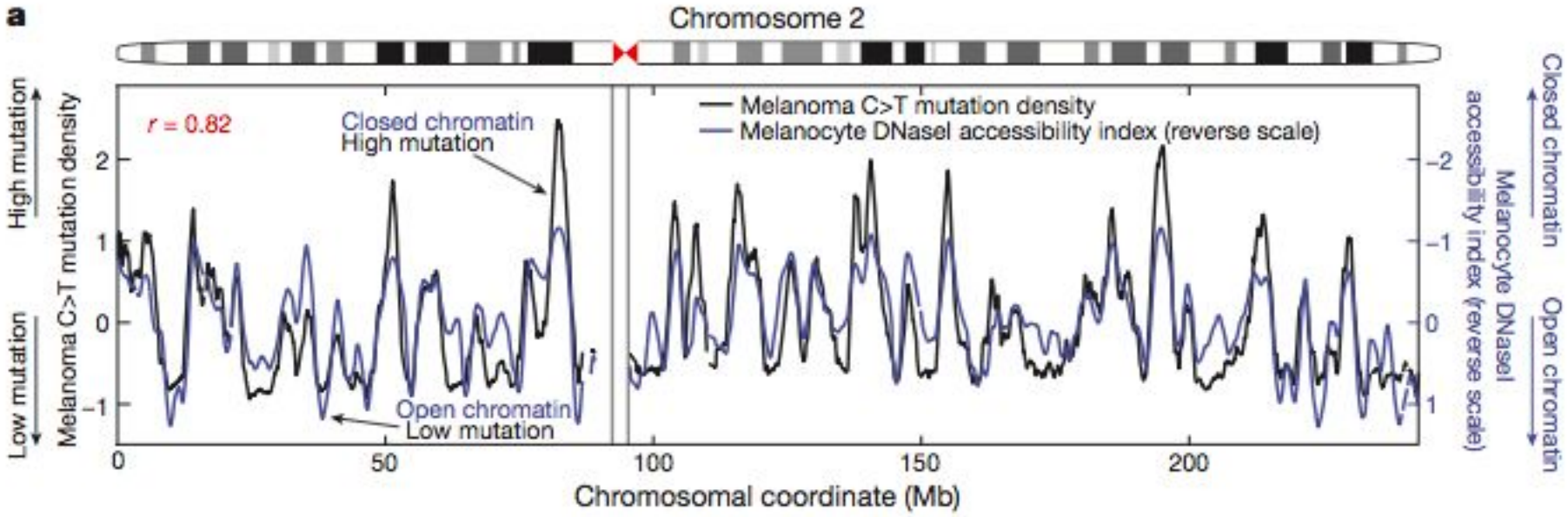


Yonggang Zhou et al. Circ Res. 2011;109:1067-1081

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

Paz Polak[1,2]*, Rosa Karlić[3]*, Amnon Koren[2,4], Robert Thurman[5], Richard Sandstrom[5], Michael S. Lawrence[2], Alex Reynolds[5], Eric Rynes[5], Kristian Vlahovičeka[3,6], John A. Stamatoyannopoulos[5,7] & Shamil R. Sunyaev[1,2]

Random Forest model predicts cancer mutation densities from epigenomic mark ups

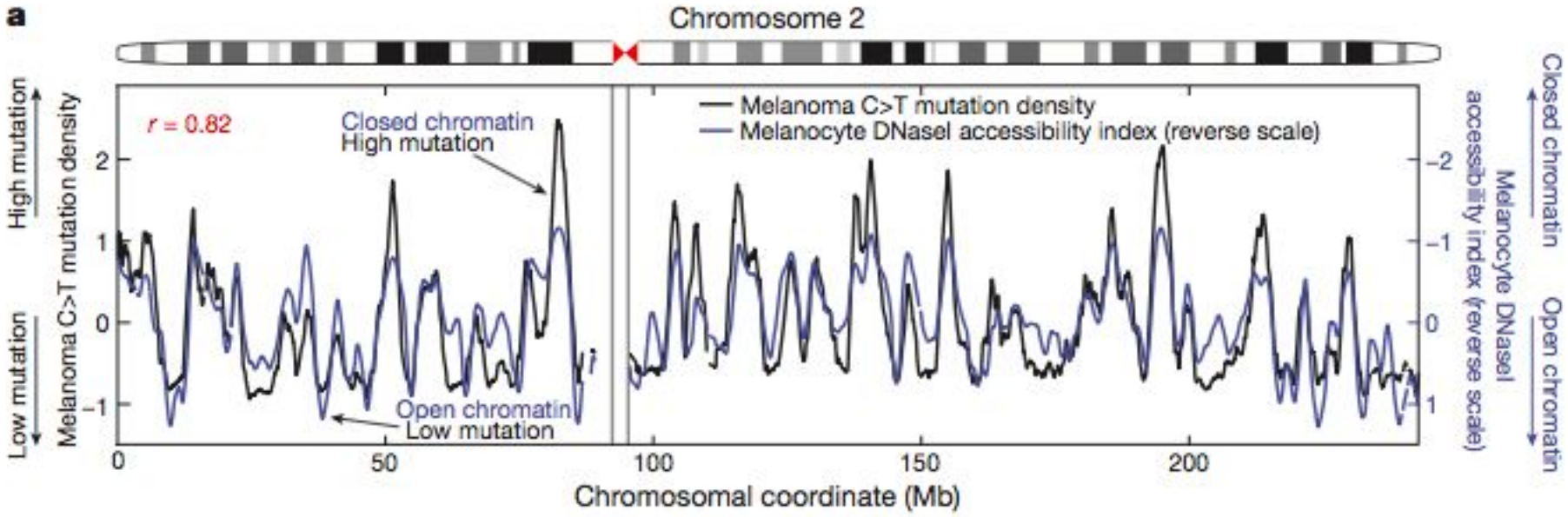


Cell-of-origin chromatin organization shapes the mutational landscape of cancer

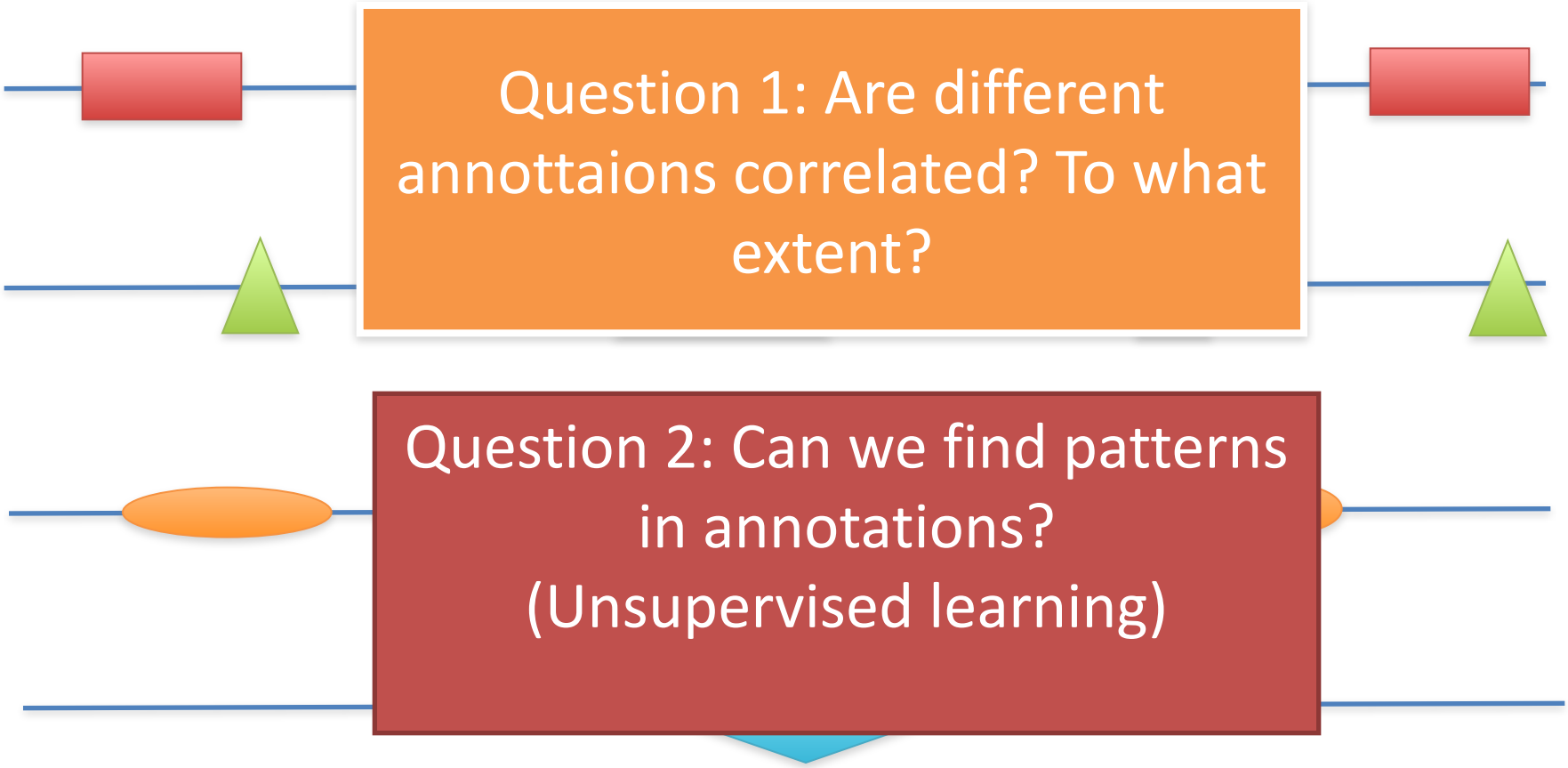
Paz Polak[1,2]*, Rosa Karlić[3]*, Amnon Koren[2,4], Robert Thurman[5], Richard Sandstrom[5], Michael S. Lawrence[2], Alex Reynolds[5], Eric Rynes[5], Kristian Vlahovičeka[3,6], John A. Stamatoyannopoulos[5,7] & Shamil R. Sunyaev[1,2]

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

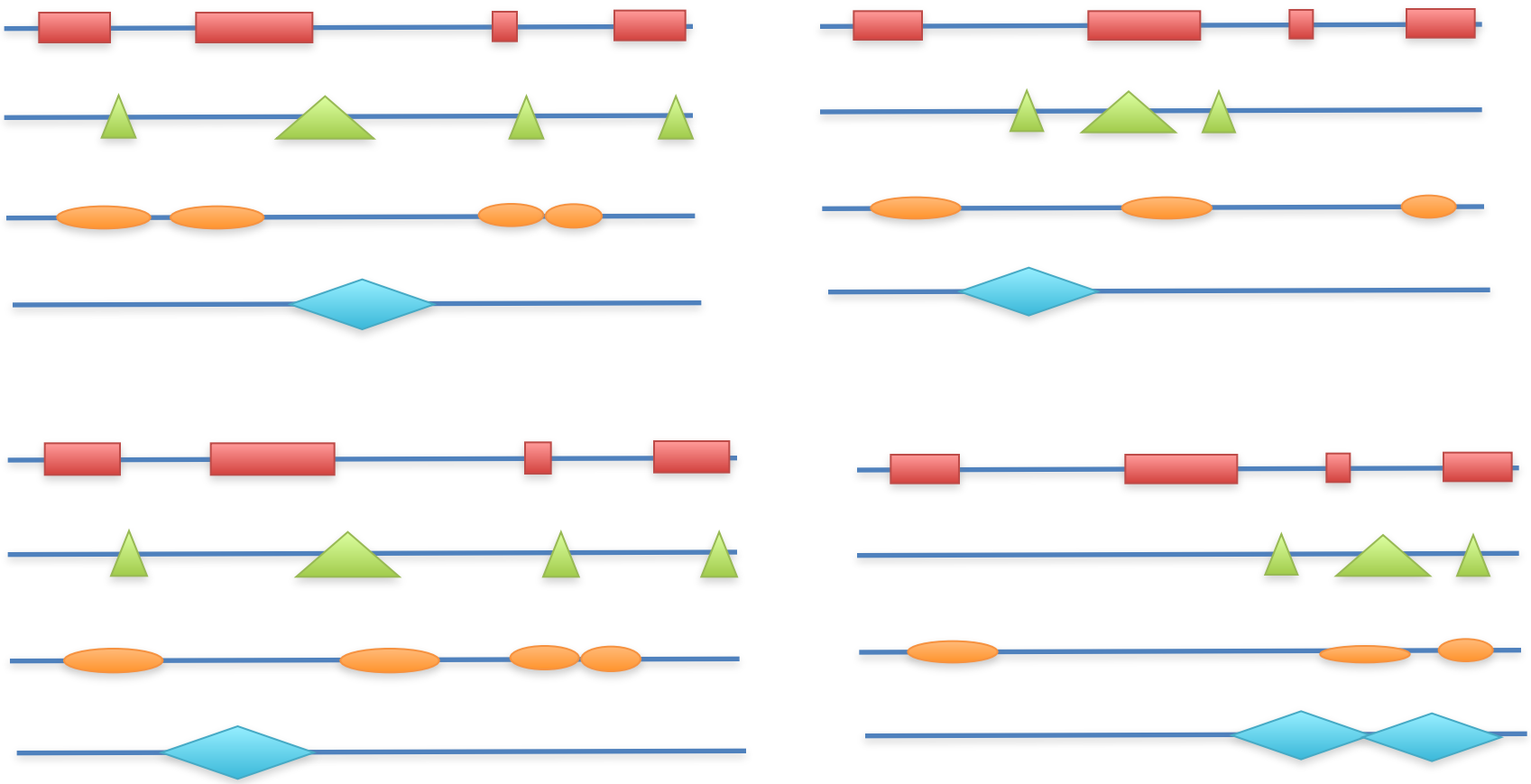
360 | NATURE | VOL 518 | 19 FEBRUARY 2015



We have many experimental genome-wide annotations



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