



HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

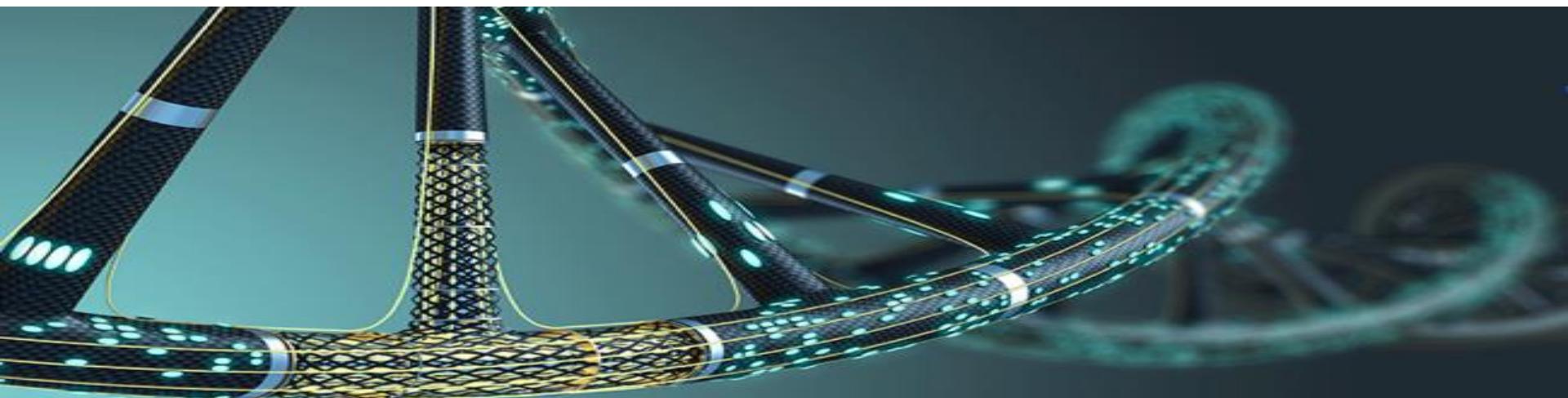
Машинное обучение в прогнозировании областей разрывов раковых геномов

Мария Попцова

Заведующая лабораторией биоинформатики

Факультет компьютерных наук

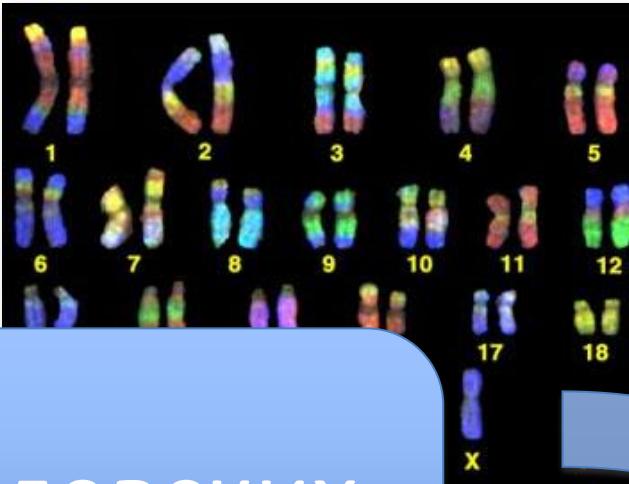
Департамент больших данных и информационного поиска



Как все начиналось?

ЧТО ТАКОЕ ГЕНОМ?

- Для биолога



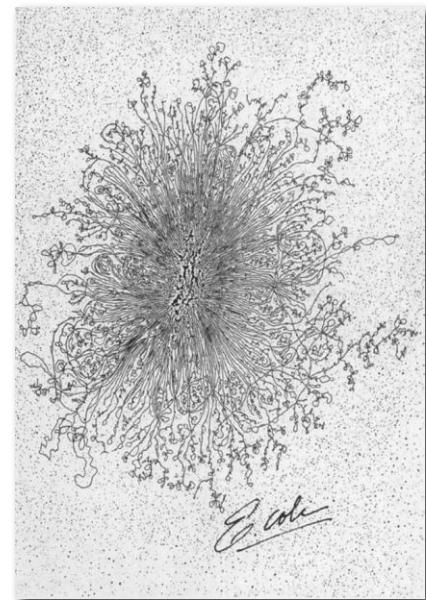
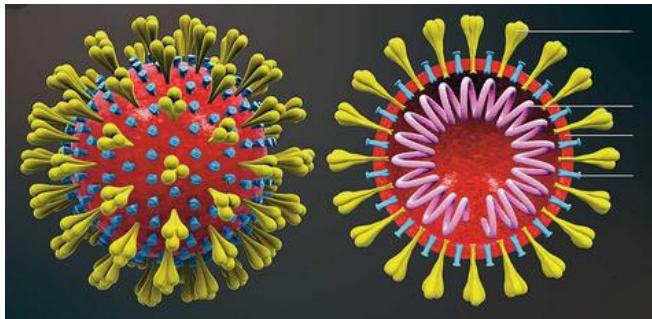
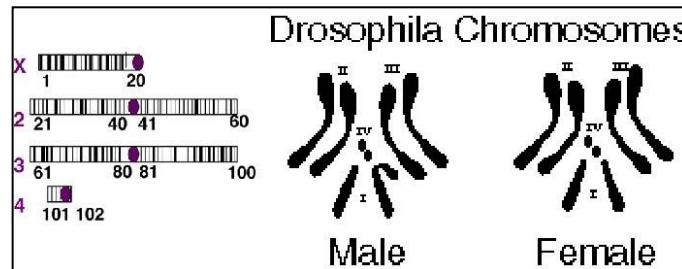
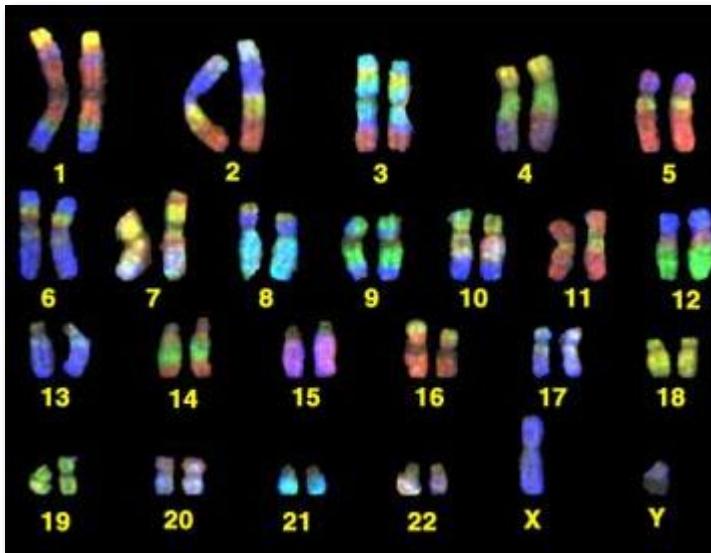
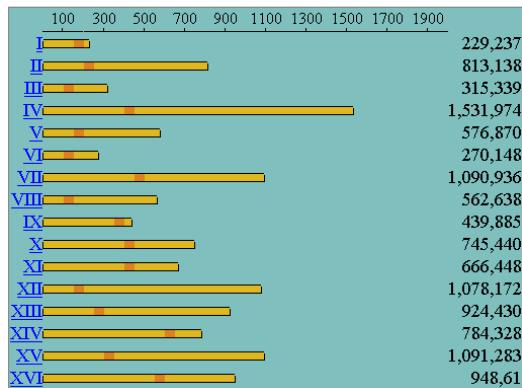
Ряд Нобелевских
премий 1950-1970

- Для биоинженеров

TAATTTAAAGGTTTCTGGCAAAACAAACGGCACAACAAATTATTAACCTAAAAAGCTTCTATA
AGGTAAAAAGCCAAATTTATGTGATCTGTGTTATCCCCCTAAATCGGAAAATGCTTAATGCC
GAAAATCCATGTTCTGCATGCCATCAATAAAAGCTCTACCATAAGCCAATATTAAAGCAGAA
ATCCCCAAATGTTCTTTAGTAAAGGGTAGTGTTAGCAAGAACCTCACCATTAAATAAGATGCAAATCT
TAAATTAAACCCAACTACGCAATTGTTAGCAAGAACCTCACCATTAAATAAGATGCAAATCT
TTGACATCCCCTACATCTCATAGCGGGAAATTAACCTACGCTTATTTTACTAGCTCTAC
TAACTATCCCTCCCTTCATCAATAGCAATAAAATATCATGTCCTAATGACTTTTATGGCCTAAT
CAACTCTTGTGTCAGCCTTAAATTTCTCTAAATAAAATTATACCAATCGGATAATCTT
TTTATCTTAAATATCATGGCATTAAGTCCTGAACATTAAAAATTGTTAAATTCCTAATTCCAA
TAAATAATCTTCTATTTAAATAGTGGAAATCAACCTTCACTAAACCTAATTAAATCTACAAG
ATCTGTTTATCTCTTAAATATCTCATCTTCGCAAGATTCTATCGCAAAATAAATTAA
CAAAAAAGATAAAAGATAAAAGCTTTAAAGCCCATCACAACACCCATTAGCAATCCATAATCTT
TGATAACAAAGTAGAAAGAGAACATAAACTTTAGTAGCATTAGAATTACCAACTCTCTTCTACAA
CAATTTGACCTTCTTCAAGATAATTACTTAAGCAACATCGCATAAAACTTCCCAGGAAGGA
ATCTTTACTCTGCAACCGTTAAATCAATGGATTCTTCTTAAATTTAATCTTCACTGGCATTCTTCTAC
CTCAACCGAAATATTACCAACTCTTCTGCTCTATAATTCTAATGCGCATTTCTTCTAC
TTATCAATGTTTTATCTCAAAATTCTAATAAAAGCCCGTAGCGCAATTAAACAGCTTATTAC
CCTTATTCTCAAACCTGTATAATTCTGAAAGAAATACCTCCACCAATCAACACAAGTTAACGATT
CTCACTATTCTGCAATACCCCTCCCTTAAATCTGCTTACTTTAAACTAAATCTCAACCCAGAA

Геномы разных организмов

Saccharomyces cerevisiae complete genome



Обитатели клетки-города

Нобелевская премия по химии 1954

Лайнус Полинг



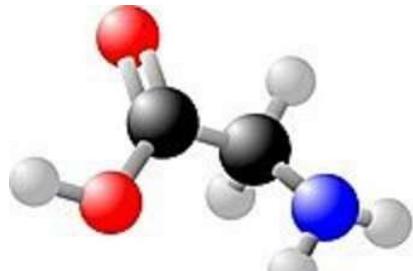
Белки - текст,
написанный на
алфавите
Из 20 букв



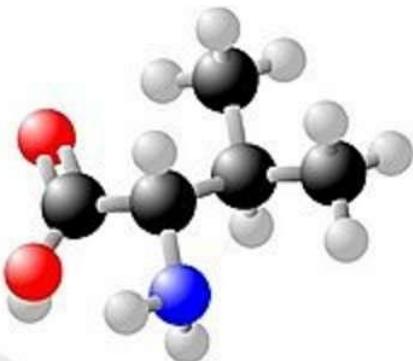
Альфа-субъединица АТФ-синтетазы

Вторичная структура – альфа-спираль и бета-лист

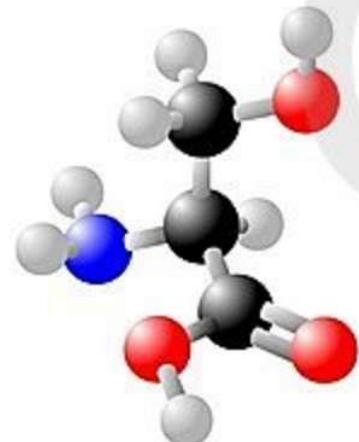
Белки сделаны из 20 аминокислот



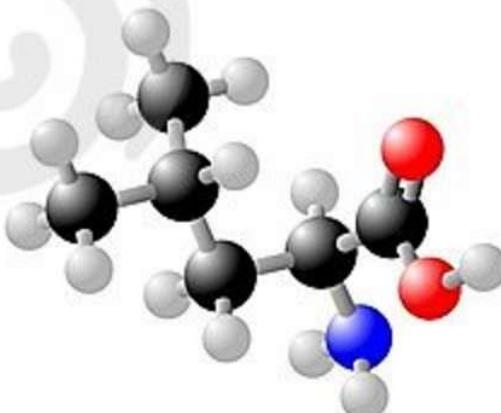
glycine



valine



serine

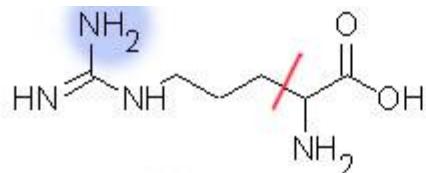


leucine

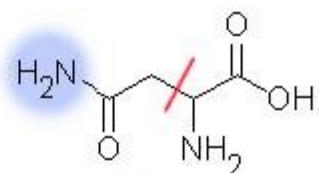
Over 100 amino acids exist in nature



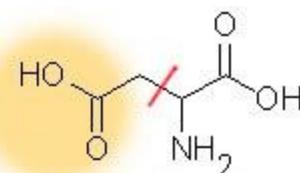
Аланин (Ala)



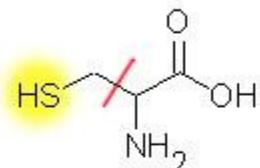
Аргинин (Arg)



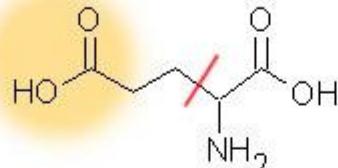
Аспарагин (Asn)



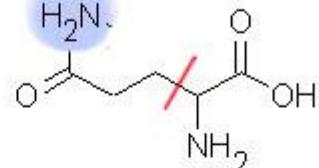
Аспарагиновая кислота (Asp)



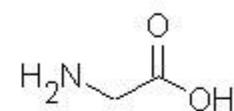
Цистеин (Cys)



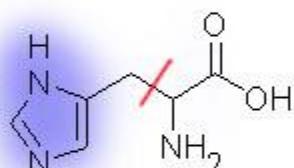
Глутаминовая кислота (Glu)



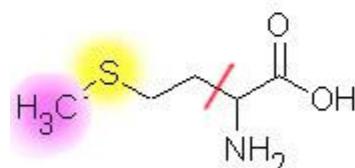
Глутамин (Gln)



Глицин (Gly)



Гистидин (His)



Метионин (Met)

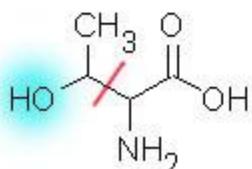
Белок - текст,
написанный на
алфавите
из 20 букв



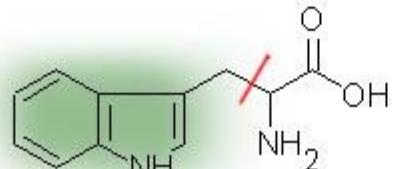
Фенилаланин (Phe)



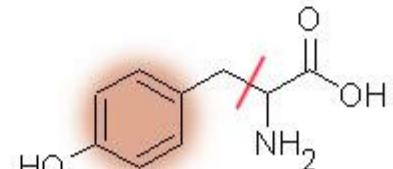
Пролин (Pro)



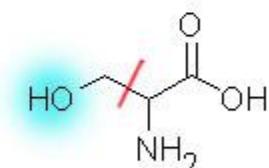
Тreonин (Thr)



Триптофан (Trp)



Тирозин (Tyr)

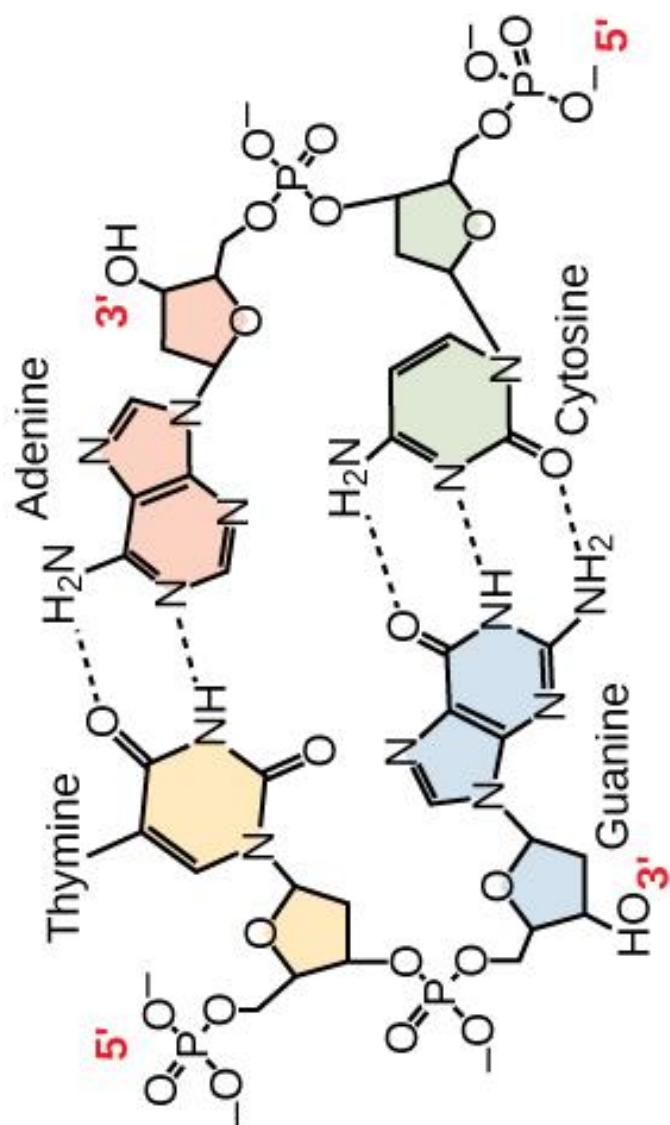
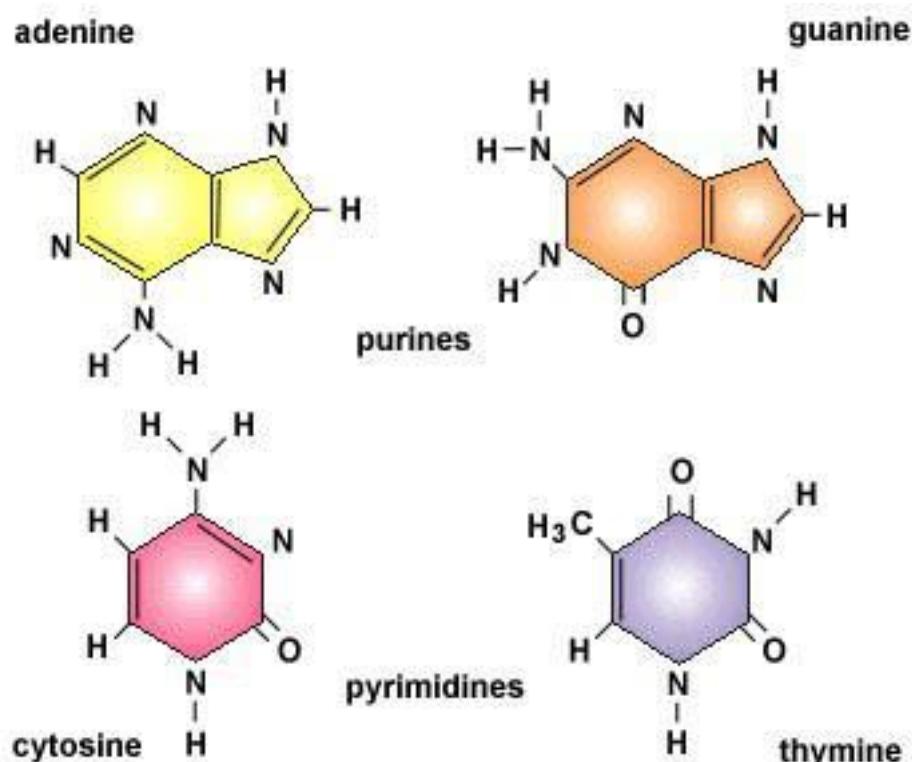


Серин (Ser)



Валин (Val)

ДНК состоит из 4 нуклеотидов



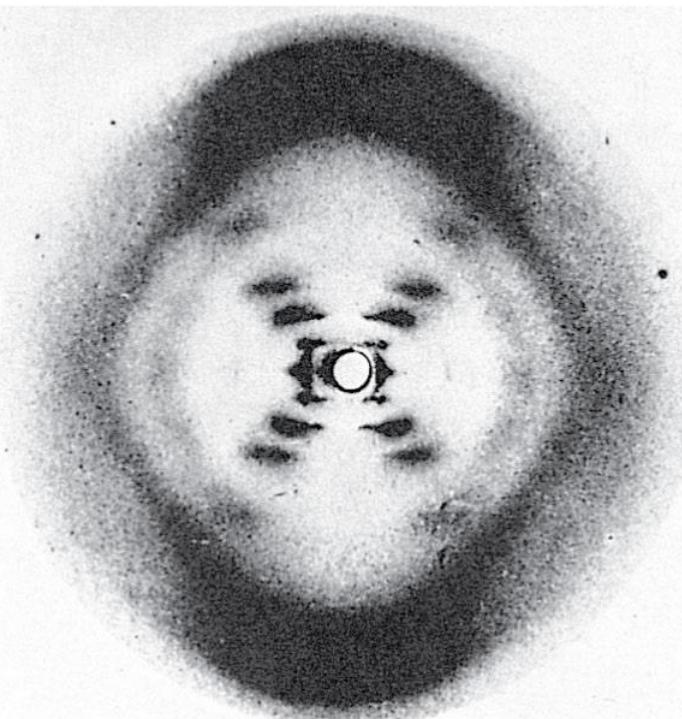
Good Quality DNA X-Ray images

- Морис Уилкинс и Розалинд Франклин



Rosalind Franklin

© 2011 Pearson Education, Inc.



Franklin's X-ray diffraction photograph of DNA

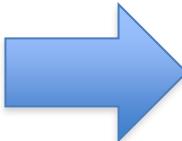
Nobel Prize in 1962



Геном - текст,
написанный на
алфавите
из 4 букв

Расшифровать генетический код

AIAAAATTAACTAUCUUCUAGATGGGAAUCUCAAAAGAAAAATAAATAAAAAAAAGAAAATAGGAAAT
AATTTCACCATATTATATAATCCCTATTTATAATGAAAACATAAATATATTGCCCTTATCGGGATA
TTAACAGCCTATAATGAATGGATTGAAAATACAGTTCAGTCCCATAAATTTTACTATCCCAATAAAC
AAGATTTGTTCAAACTGCTTATTTAATTAGCTTCACTATTTACATTGTTAAATATTCAATTTCAGC
CGATACACTTACATAAAATCTTGTGGAAACCCAAGTCGATTAACTCTAAGAACACTAGTATTTATA
GGAAAAAAACAACTAATGCATTCTCTATCCAATCTCCCTAATTACCTTTAAATTGAAATTGATTCA
TACCTAATAACTATAGCATTTACTATAAATTATCACTCTTTAAAGAATTCTTTGGATCTAGG
AATTCTATATTATACCATAAAAAATTTAAACATTTCTCTTTGGAAATATTAAACAAAAGACT
ATCCAATTGCTAAATATAAAAATAAAACTGACAAGACTAAAAACAAATTAAAGCAATTTCAGT
AAGGAATAAGAACTATTTACCTGAACTTAAAGACTTTTAATAGCATTTCATAATTAGAAAAGACT
TCTGTAACTGCTCCCATATTATAAGAATCTGTATCATAACTAGTAAATTAAATTCTCTTATA
ATATTAACAATACTTTAGACATGCTAGTTATCTTAGAAACATTGTCATGGCCAACTCA
TGATAAATTAGCTTGTGCCCCAAAACACGGCACAAAATTATTAACATTAAAAGCTTAACTATA
AGTAAAAAAAGCCTAATTATGTGTATCTGTTATTCCCTTAATCCTGGAAAATGCTTAATTGCC
GAAAATACCTCATGATTCTGCATGCCATCAATAAAAGCTTACCCATAAGCCCAATTATAAGCAGAAT
ATCCCCCAAATGTTCTTTAGTAAAGGGGTATGTTGATAAATTATGTCGGCACTGGGCCATT
TAAATAATACCCAACTACGCAATTGTTAGCAAGAACCTTACCAATTATAAAAGATGCAAATCT
TTGACATCCCTCATACTCCATACCGGAAATTATAAACTCCCATTTTTTACTAGCTTAC
TAACATCCCTCCCTCTCATAGCAATAAAATATCTGTCATGGCAATTGACTTTTATGCTTAAT
CACTATTTGTTGCTCAGCATCTTAAATTCTCTAAATAAAATTACCAATGGATTAACTCTT
TTTACTTCTAAATATCATGGCATTAAAGCTTGAACATTTTAAATTGTTAAATTCTCTAA
TAAATACTCTATCTTTAAATTAGTTGAAAATCCTAAACCTAATTAAATCTACAAAG
ATCTGATTTATCCTTAAATATTCTATACTTATCTGGCAAAAGATTCTATCGCAAAATAATTAA
CAAAAAAGATAAAAGATAAAAGCTTAAAGGCCATCTACAACACCCATTAAAGCAATCCATAATT
TGATAACAAAGTAGAAAAGAGAACATAAACTTTAGTAGCATTAGAATTACCAACTCTCTTCTCAA
CAATTTTGACCTATCTCAAGATAATTACTATTAAAGCAACATGTCGAAATAACTTCGCAGGAAGGA
ATCTTTACTCTGTCACCCGTTAAATCAATGGATTTCCTTAAATTAACTTCCACGTTATTG
CTCAACCGAAAATATTACCAACTCTTATCTGCTCTATAATTCTAATGGCATTATTCTATTAC
TTACATTTTTTCTCAAATATTCTAAATTAAAAGCCCGTACCGCAATTAAATAACAGCTTATTAC
CTTATTCTCAAACCTCTGTTATAATTCTGAAAGAAATACCTCCACCAATATCAACACAGTTAAAGCATT
CTCACTATTACTGCCAATACCCCTCCCTAAATCTGCTTACTTTAAACTAAATTCAAATCCAGAA



>gi|2400474|ref|YP_002960858.1| F0F1 ATP synthase alpha chain [Mycoplasma conjunctiva]
MTPKVVSLDYLVLAKGQYPWKEQQFYIKNKPNIQAVVIQAOQQDQAYLLFNNQKGSVIDELVELND
DKVKTSEFFGTIVDFSGTIIEPANQKVLNFLPHRSSAFATAASILGRKNLDTQLYTGLSIDLFPNPIGL
GORELIVGDRQTGKTHIGINAIINQRDTNIKCIYWSVGQKRQNLSFVLKALRENNALDNTIIFHAPSTSP
FEQYLIPYFAMAHAEILNSYDSDWLIVFDDLSKHSVYREVALLTNQPIGKEAFPSDFIFYTHSKLLERSGK
FVGRHSITALPILQTVDDITSLSISSLNVISITDGQIITSSALFAEGKIPAVNIGLVSRTGSSVQAKNVR
DVSKEISSIONYQKQIKLSKLDYDLNQQTSDLLFKGSQIEQFLQKGYSFVSPKVMGLGIKISWGLLK
NISEPSRWDIYTLIEVDPFAKRIYQKYNDNQFVDDKLARNYFATAVNQFLKVNNSKERLEIEFPEI
SEKISTSIAKLKEKIGAR
>gi|31544508|ref|NP_853086.1| F0F1 ATP synthase subunit alpha [Mycoplasma galliseptenae]
MAINLNEYSLLIKDKIKKYANKIISDQKGYIITIDGGIVRVSGLDDVLLNELVEFENGAYGIGALNLEPNS
VGWMLMSDYYDLKEGSSVKRTGKVIQAPVGDDGLLRVIDPILGPIDGKGEKLKNISGYAPIERLAYGMQR
KSVPQPLETGILAIIDSMLPIGKGQRELIIQDRQTGKTTIALDTIINQKGKVNVCIIYVAIGQKNSSSVAQIT
RLLEETGAMAYTTIVSATASELAALSYIAPFAGVTIGEEMMRQGKDVLIVYDDLSKHAVAYRALSLLLRR
PPGREAYPGDIFYLHSRLLERAGKLSDELGAGSITALPIIETQAGDISAYIPTNVISITDGQLFTTSLF
NSGQRPAIHVGLSRSVGSAAQQLSIKQVSGSLKLELAQYRELDTSFQSFSQSDLDAETKIVLHEGARVMEM
FKQPAQKPIDQTSCEAVLLFGIKNRFKWIPTDHIIKFKEFILDKIKQDQVYKKIEEKKAFFDDEIEKELTA
FFKDWDVKYTSTLVTDYNGSLYGDLEKE
>gi|12845263|ref|NP_073074.1| F0F1 ATP synthase subunit alpha [Mycoplasma genitalium]
MADKLNEYVALIKTEIKKYSKKIFNSEIYGQVISVADGIAKVGSLNELLNLNLIQFENNIQGIVLNLEQNT
VGIALFGDYSSLREGSTAKRTHSVMLKTPVGDMRIVNALGEAIDGRGDIKATEYDQIEKIAPGVVMRK
SVNQPLETGLTIDALFPQIKGQRELIVGDRQTGKTAIDTIINQKDKDVYCVVVAIGQKNSSSVAQIVH|
QLEVNDSMKVYTVWCATASDSDMVSLSPFTGITAEWLKKGDVLIVFDDLSKHAVAYRTLSLLLRP
PGREAFPGDWFYLSHSRLLERACKLNDENGGSITLPIIETQAGDISAYIPTNVISITDGQLFMVSSLFN
AGQRPAIQIGLSRSVGSAAQTKAIQQTGSLKLELAQYSELDSFSQGSDLDENTKKVLEHGKRVMEMI
KQPNKGPKYPSQVHEALFLFAINKAFIKFIPVDEIAFKQORITEEFNGSHPLFKELSNKEFTEDLESKT
AFKMLVKRKFISTLTDYDITKFGSIEELN
>gi|15402008|ref|YP_115568.1| F0F1 ATP synthase subunit alpha [Mycoplasma hyopneumoniae]
MNKDINIAAIKNEIENFEGKIQNHQD1GKVIIVGDDGVALVSGIEKVKFGLVEFENNVLGIALNLEQDLI
GVVIMASENSVFQGSIVRRTKSVISITVGQDQLLGRVWNALGIPIDGKAELDNSLKSAITNAPSIMDRKS
VDRGLKGTGAIADSMLPIGKGQRELIIQDRQTGKTTIAIDAILNQKGKVNVCVYVAIGQKNSSSVAQIVSL
LEKKGAFEYTTVILAGASELSPLQYLAQYSAAIAYWMNKKDVLIIYDDLSKHAIAVRTLSLLLRP
GREAFPGDIFYQH5YLLERSAQLSNDKGGGSITALPIIETQAGDISAYIPTNVISITDGQIFLRLDSLFNS
GQKPAIDIGLSRSVGSAAQTNLMKWASSSLKLNLAQYNELKAFQAQFGSDLGPSSQILIDRGNKIYEILK
QENQYPLTERQQIMLLILIRENLIDSLEQKSIPLFKSAFLKYCDSEPKFRDKIEKMDYNRVLEPNNLAGI
LKDITDIEKFQNLGNFV

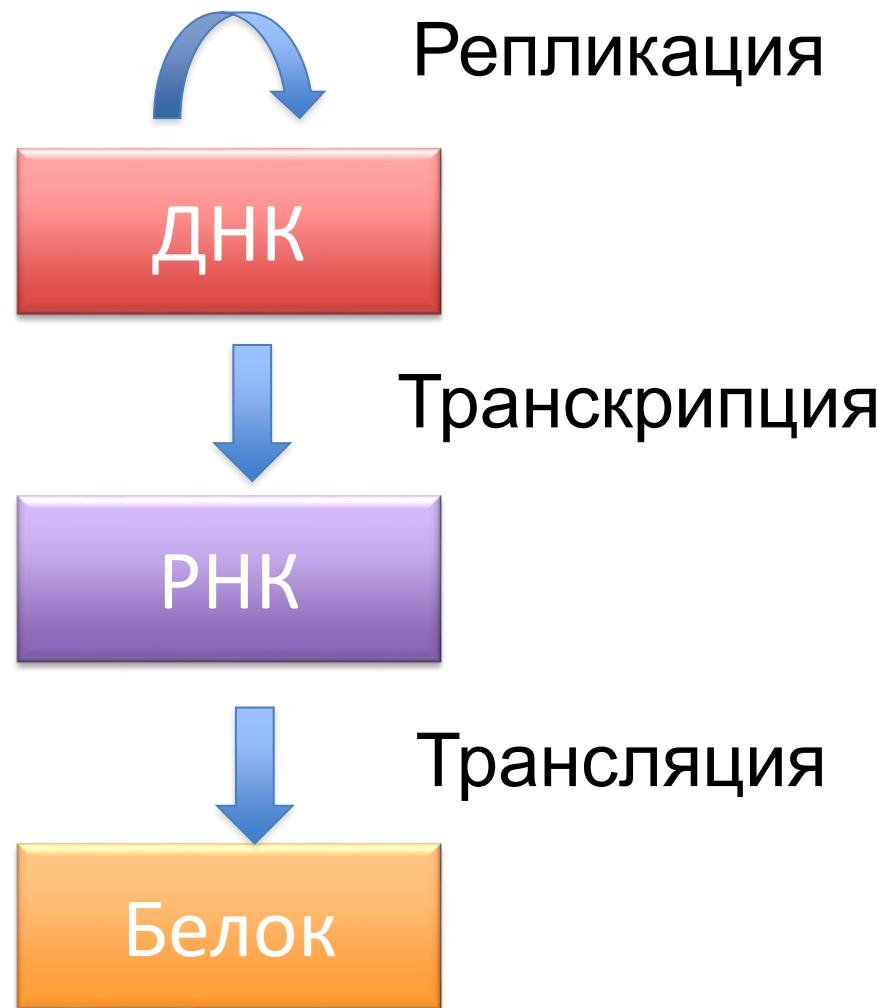
1968 - Нобелевская премия по физиологии и медицине

Ниренберг, Хорана,
Холли

	U	C	A	G	
U	UUU Phe UUC Phe	UCU Ser UCC Ser	UAU Tyr UAC Tyr	UGU Cys UGC Cys	U C
	UUA Leu UUG Leu	UCA Ser UCG Ser	UAA TER UAG TER	UGA TER UGG Trp	A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G



ИНФОРМАТИКА



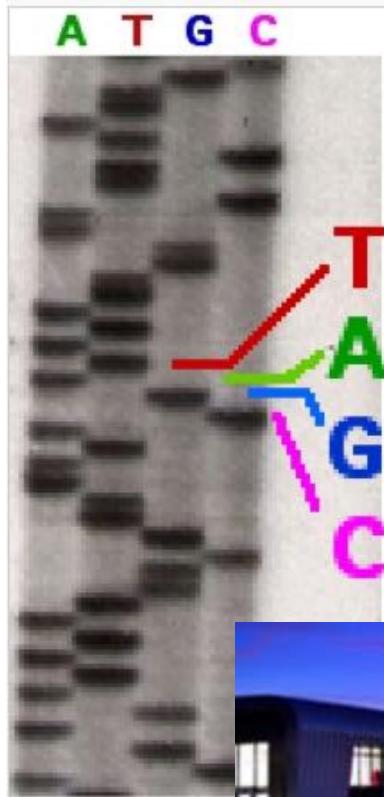
Рождение биоинформатики

Молекулярная эволюция – эволюция текстов

Ahus5	MASGIARGLAEERKSWRKNHPHGFVAKPETGQDGTV- NLMVWHCTIPGKAGTDWEGGFPLTMHPSEDYPSKPPKCKFPQGFFHPNVYP	89
OsUbc9	MSGGIARGLAEERKAWRKNHPHGFVAKPETMADGSA- NLMIWHCTIPGKQGTDWEGGYPLTLHPSEDYPSKPPKCKFPQGFFHPNVYP	89
PpUbc9	MSGGIARGLAEERKAWRKNHPHGFVARPETGADGAL- NLMVWQCTLPKGKVGTDWEGGFYPVAIHPSEDYPSKPPKCKFPQGFFHPNVYP	89
DdUbc9	MA-GISSARLSEERKNWRDHPYGFARPSTNTD GSL- NLVVWNCGIPGKTKTNWEGGVYPLIMEPTEDYPSKPPKCRFPKDFFHPNVYP	88
HsUbc9	MS-GIALSRLAERKAWRKDHPFGFVAVPTKNPDTGM- NLMNWECAIPGKKGTPWEGGLPKLRMLPKDDYPSSPPKCKFEPLLFHPNVYP	88
DrUbc9	MS-GIALSRLAERKAWRKDHPFGFVAVPMKNPDGM- NLMNWECAIPGKKGTPWEGGLPKLRMLFKDDYPSSPPKCKFEPLLFHPNVYP	88
DmUbc9	MS-GIAITRLGEERKAWRKDHPFGFVARPAKNPDTL- NLMIWECAIPGKKSTPWEAGGLYKLRLMIFKDDYPTSSPPKCKFEPLLFHPNVYP	88
SpHus5	MS-SLCKTRLQEERKQWRDHPFGFYAKPCKSSDGL- DLMNWKVGIPGKPKTWEAGGLYKLTMAPPPEEYPTTRPPKCRFTPLLFHPNVYP	88
ScUbc9	MS-SLCQLRLQEERKKWRKDHPFGFYAKPVKKADGSM- DLQKWEAGIPGKEGTNWAGGVYPIITVEYPNEYPSKPPKVKFPAGFYHPNVYP	88
PfUbc9	MS--IAKKRLAERAEWRKDHPAGFSAKYSPMSDGKGLDIMKWICKIPGKKGLWEGGEYPLTMEPTEDYPSKPPKCKFTTVLFHPNIYP	88
Ahus5	SGTVCLSILNEDYGRPAITVKQILVGQDLDTPNPAQTDGYHLFCQDPVEYKKRVKLISKQYPALV	160
OsUbc9	SGTVCLSILNEDSGGRPAITVKQILVGQDLDQPNPAQTDGYHIFIQDKPEYKRRVRVQAKQYPLL	160
PpUbc9	SGTVCLSILNEDSGGRPAITVKQILVGQELLDAQPNPAQTEAYQLFIQDPVEYKRRVRQQAKQYPPP	160
DdUbc9	SGTVCLSILNEEADWKPSVTIKTVLLGQDLDNPNPKSPAQQLPIHLFLTNKEEYDKKVKAAQSKVYPPPQ	159
HsUbc9	SGTVCLSILEEDKDWRPAITIKQILLGQELLNEPNIQDPAQAEAYTIYCQNRFVEYKRVRAQAKKFAPS-	158
DrUbc9	SGTVCLSILEEDKDWRPAITIKQILLGQELLNEPNIQDPAQAEAYTIYCQNRFVEYKRVRAQAKKFSPS-	158
DmUbc9	SGTVCLSLLDEEKDWRPAITIKQILLGQDLLNEPNIKDPAQAEAYTIYCQNRLBEYKRVRAQARAMAATE	159
SpHus5	SGTVCLSILNEEEGRWPATIKQILLGQDLDTPNIAQPAQTEAYTMFKDKVVEYKRVRAQARENAP--	157
ScUbc9	SGTICLISILNEDQDWRAITLKQIVLGVDLLDSPNPNSPAQEPAWRSFSRNKAEBYDKKVLLQAKQYSK--	157
PfUbc9	SGTVCLSILNEDDWKPSITIKQILLGQDLDNPNPNSPAQAEFPFLYQDRDSYEKKVKKQAIERFRPKD	159

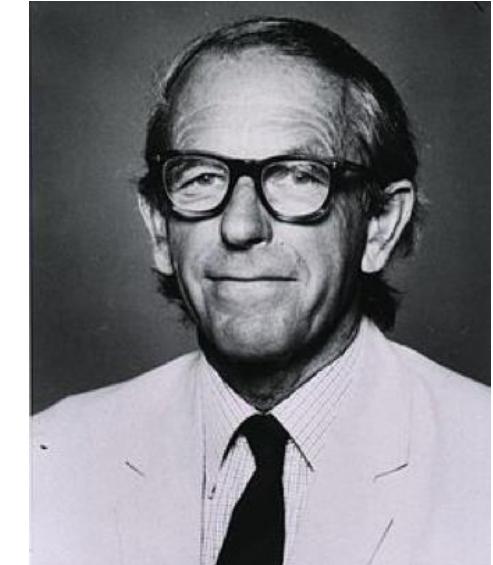
Метод секвенирования Сэнгера

1992 г Нобелевская премия по химии 1980 г.



- 1977 год

Начал с получения полной аминокислотной последовательности инсулина в 1951-52

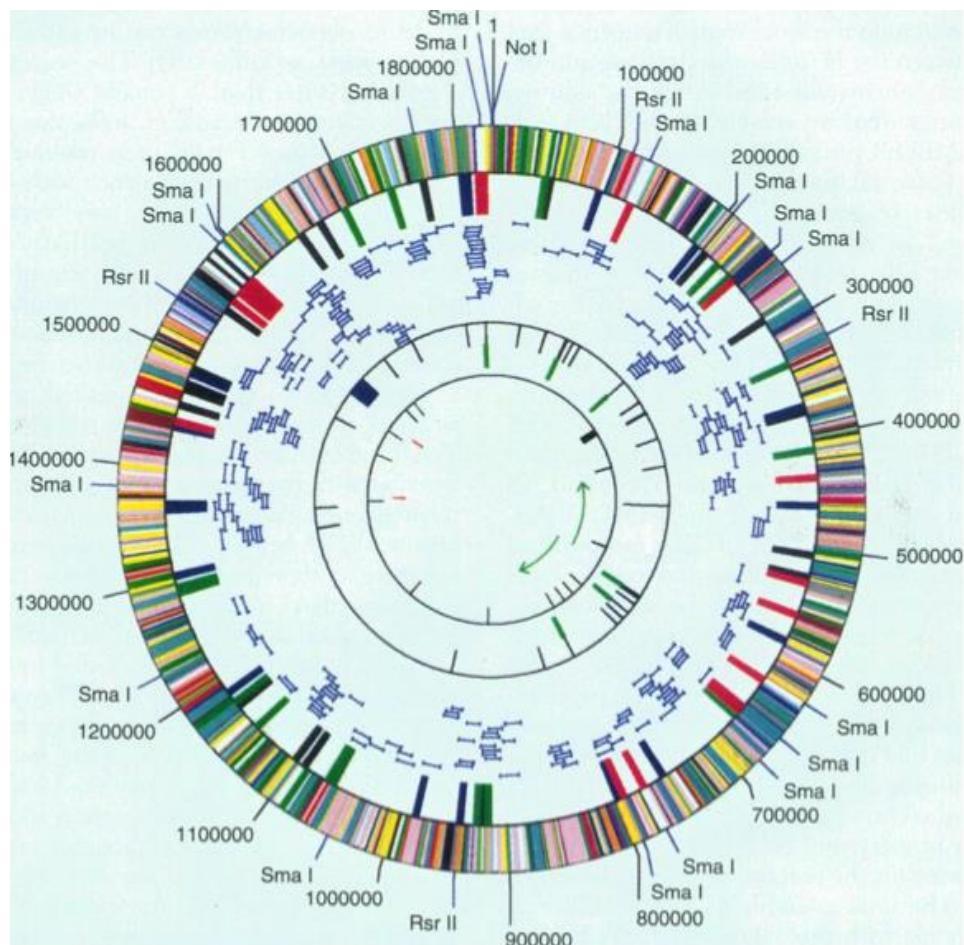


The Wellcome Trust Sanger Institute, 1992

Первый отсеквенированный геном

- 1995 г.

бактерия *Haemophilus influenzae* – гемофильная палочка или палочка Пфейфера



Проект генома человека

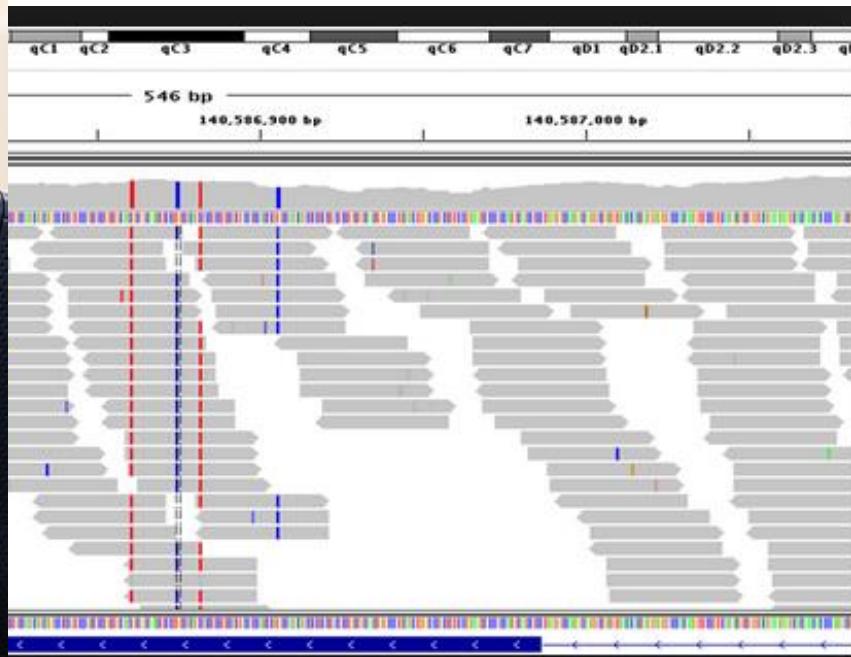
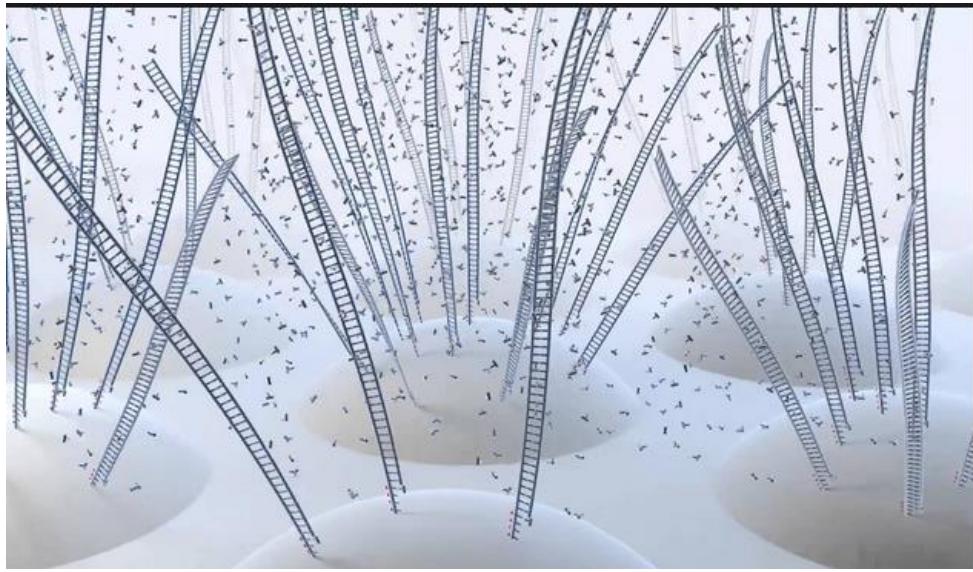
- Started in 1990
- Finished 2001/2003
- Sanger Sequencing



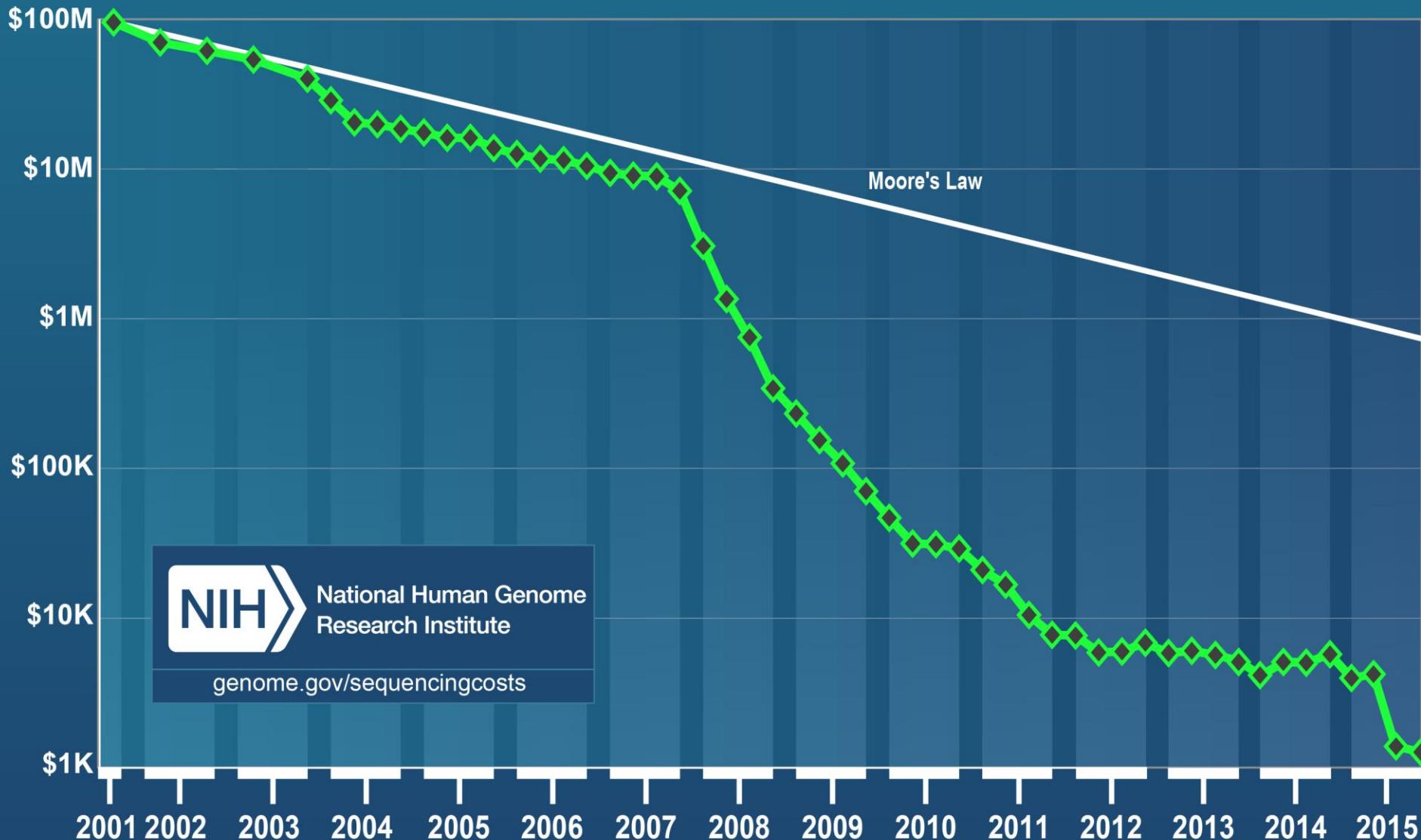


NOT IN USE
UNTIL FURTHER NOTICE
PLEASE PRE-RUN THIS
MACHINE FOR 30 MINUTES
BEFORE LOADING

Next Generation Sequencing,
or NextGenSeq,
or NGS



Cost per Genome



NextGen Sequencing Revolution

To print is more expensive than to sequence

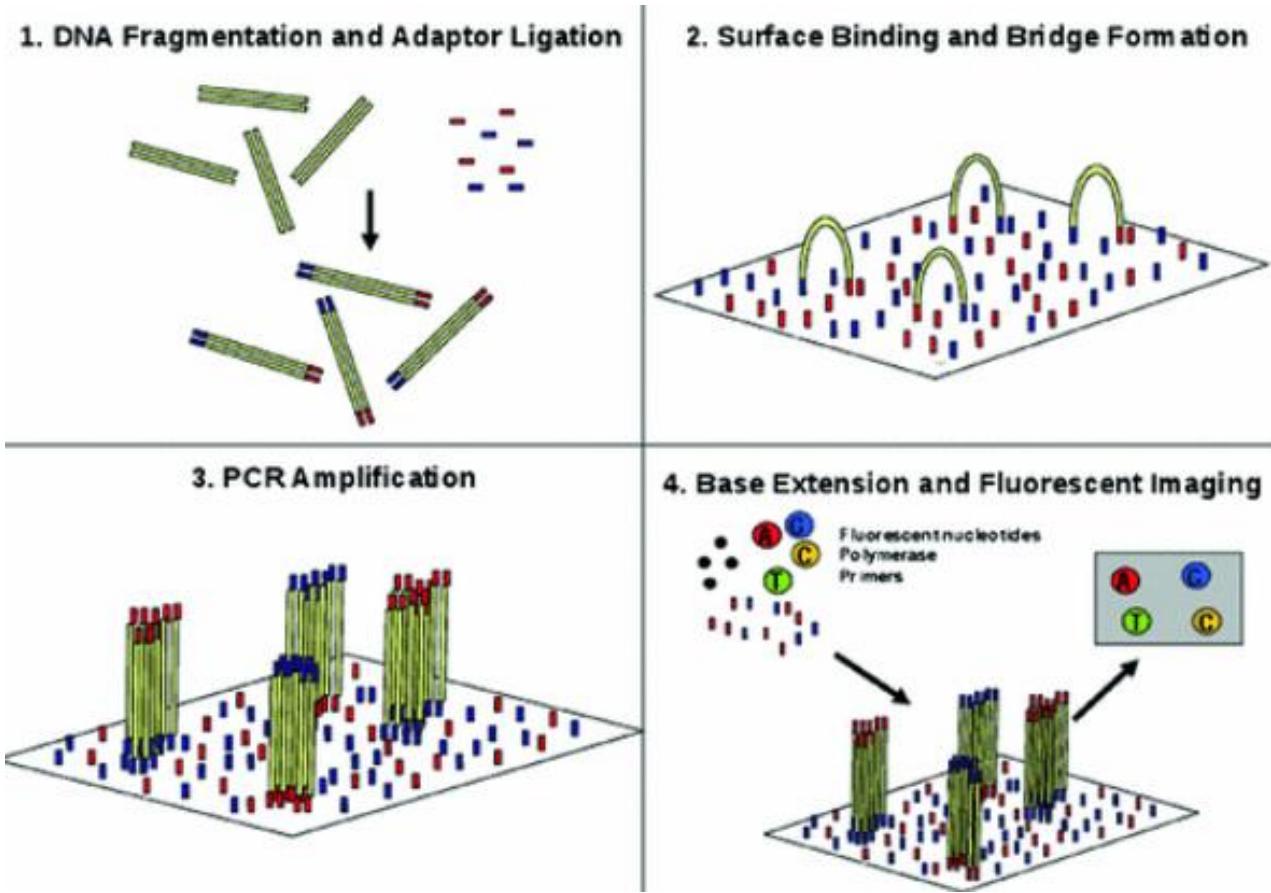


Human genome –
130 Volumes, double-sided,
4pt, ~ 43,000 chars per page

X-chromosome – 7 volumes
Y-chromosome – 1 volume

Next Generation Sequencing

- DNA-seq
- RNA-seq
- Chip-Seq
- MNase-Seq
- Hi-C



Massive parallel sequencing

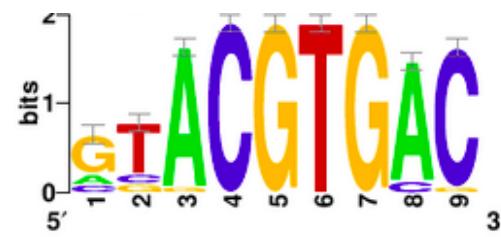
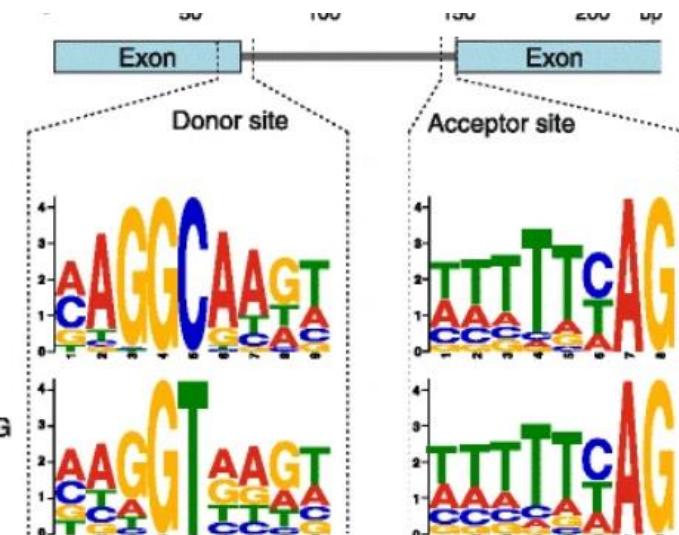
1 million to 43 billion short reads per instrument run

Слои геномной кодировки

- DNA sequence

DNA-seq, RNA-seq,
ChIP-seq

- Encoding into sequence: genes, exons, introns, promoters (some), enhancers (unclear), transcription factor binding sites



	U	C	A	G	
U	UUU Phe UUC Phe	UCU Ser UCC Ser	UAU Tyr UAC Tyr	UGU Cys UGC Cys	U
C	UUA Leu UUG Leu	UCA Ser UCG Ser	UAA TER UAG TER	UGA TER UGG Trp	C
A	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	A
G	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	G
	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

Legend:

- Hidrophobic - Imino (purple)
- Hidrophobic - Aliphatic (red)
- Polar - Neutral (green)
- Polar - Basic (yellow)
- Hidrophobic - Aromatic (pink)
- Polar - Acid (blue)

Эпигенетика

Levels of DNA Packaging in Eukaryotes

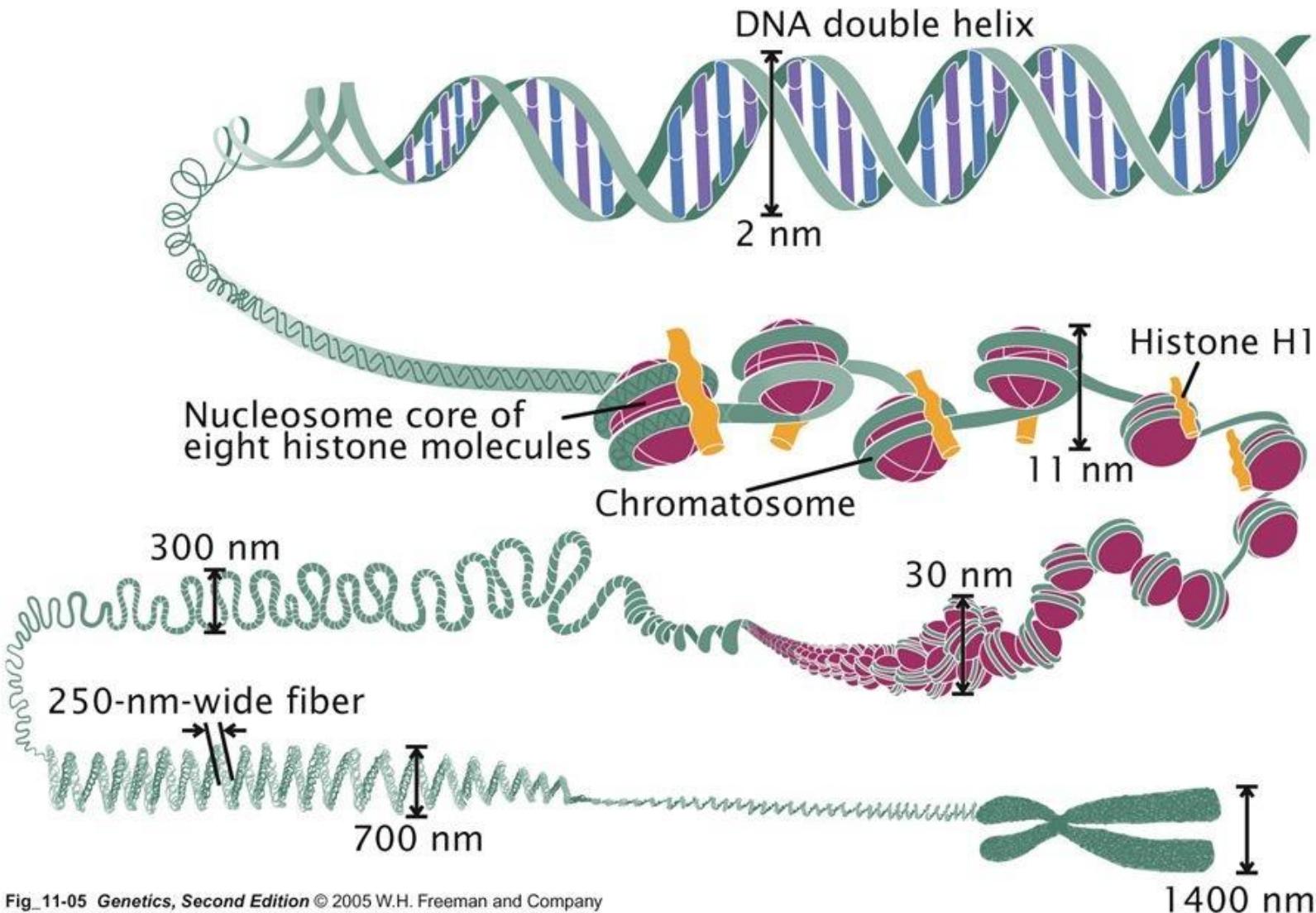
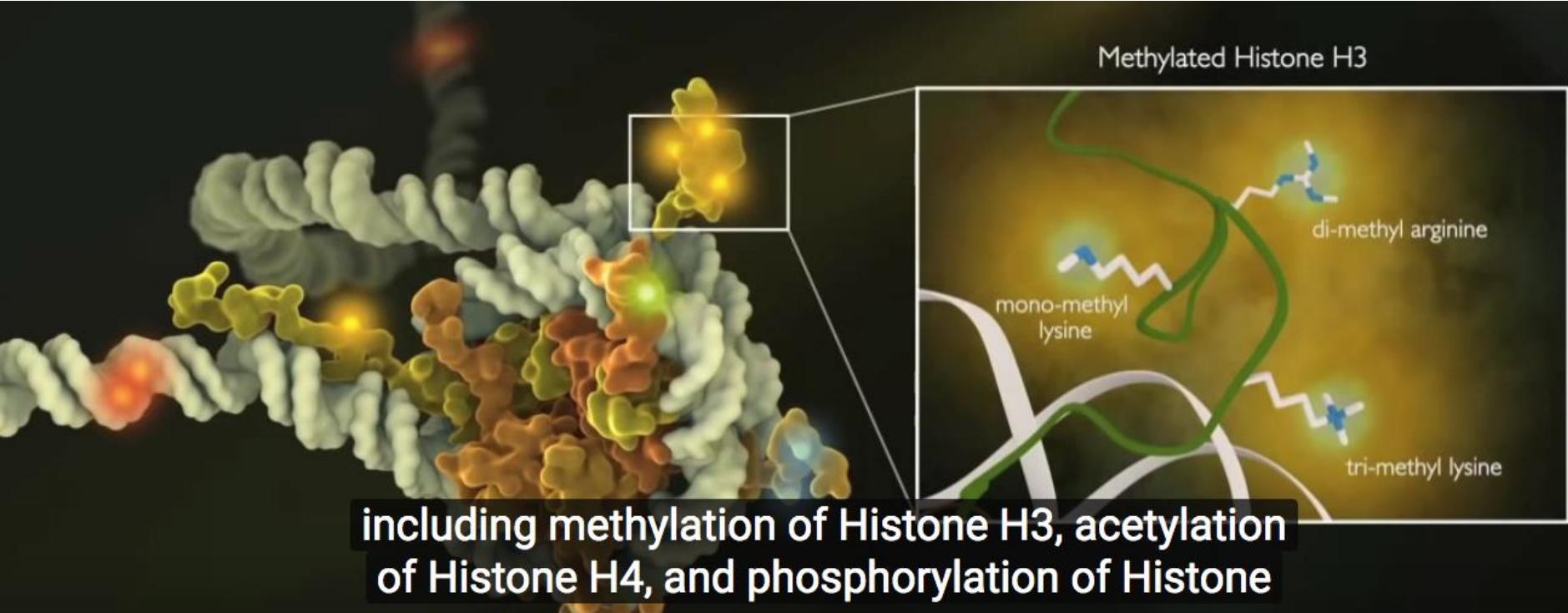


Fig. 11-05 *Genetics, Second Edition* © 2005 W.H. Freeman and Company

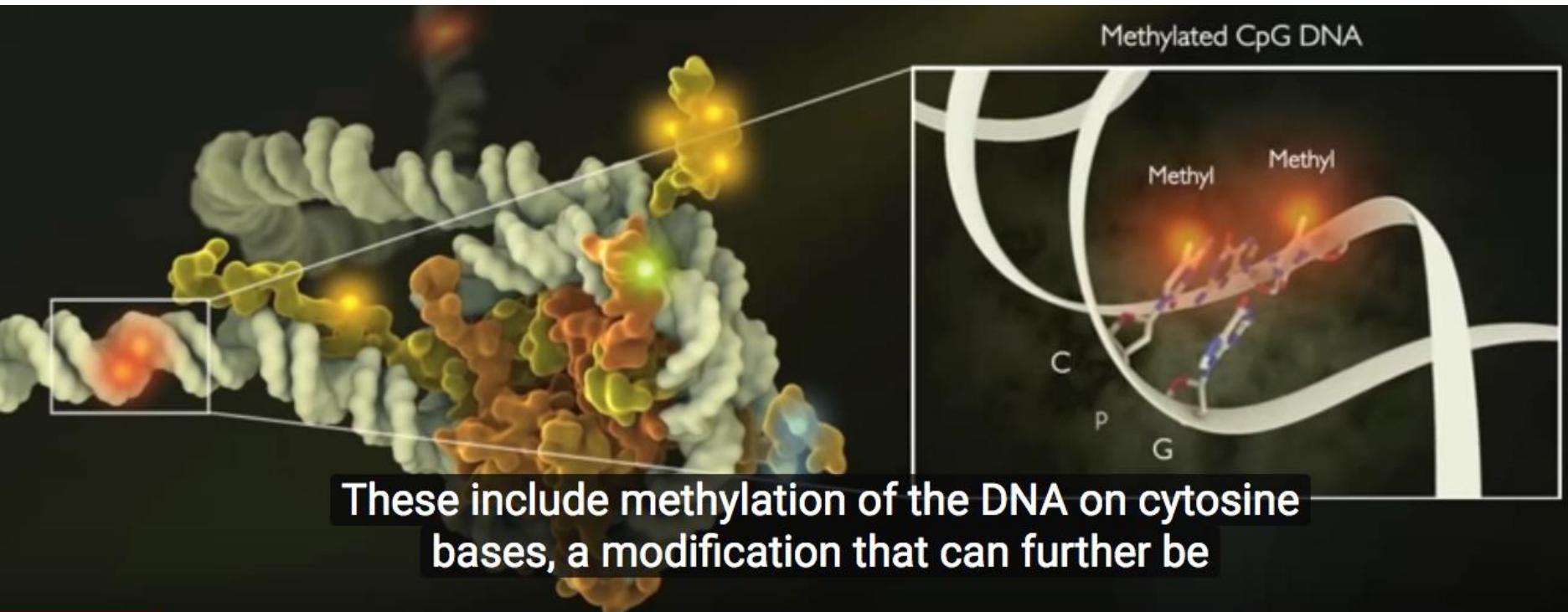
Слои геномной кодировки

- DNA marks: Epigenomics
 - Methylations, histone modifications, chromatine packaging

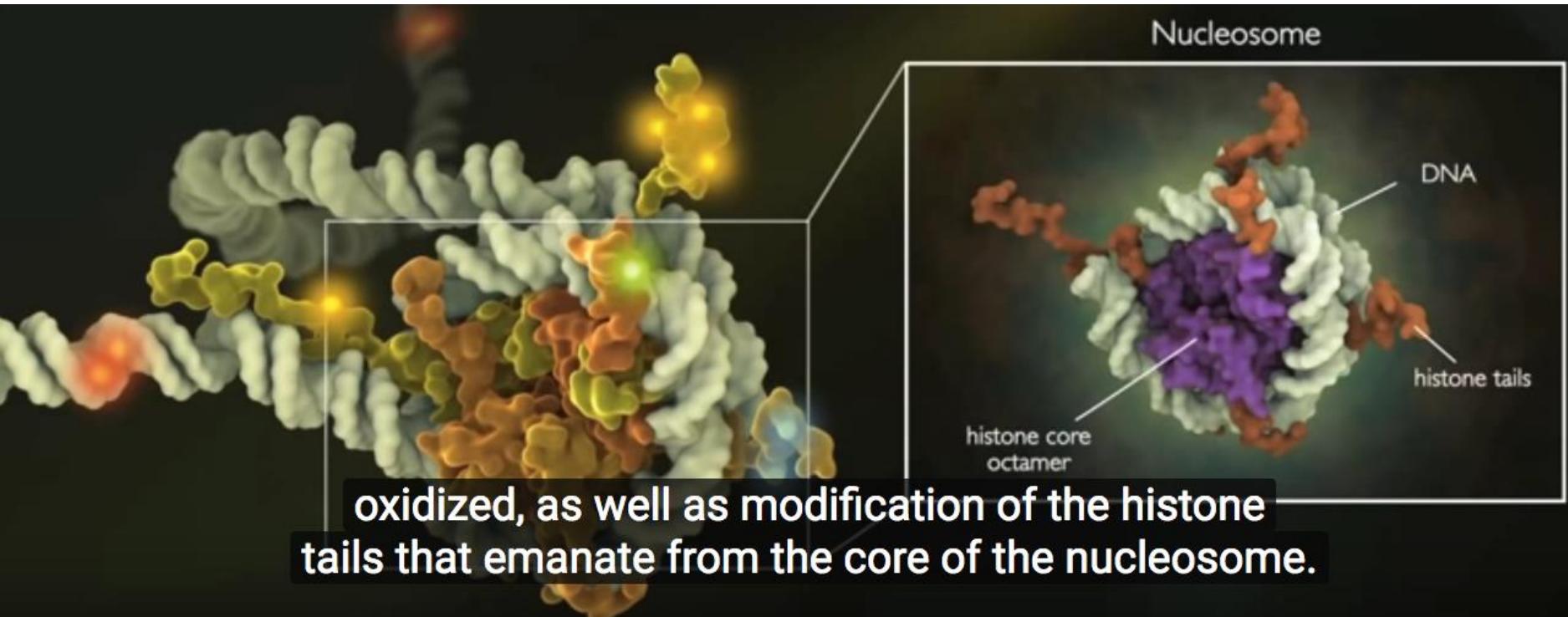
ChIP-seq, Bisulfite, MeDIP-seq,
MAD-seq, CLEVER-seq, Mnase-seq,
Hi-C, Hi-3C and more....



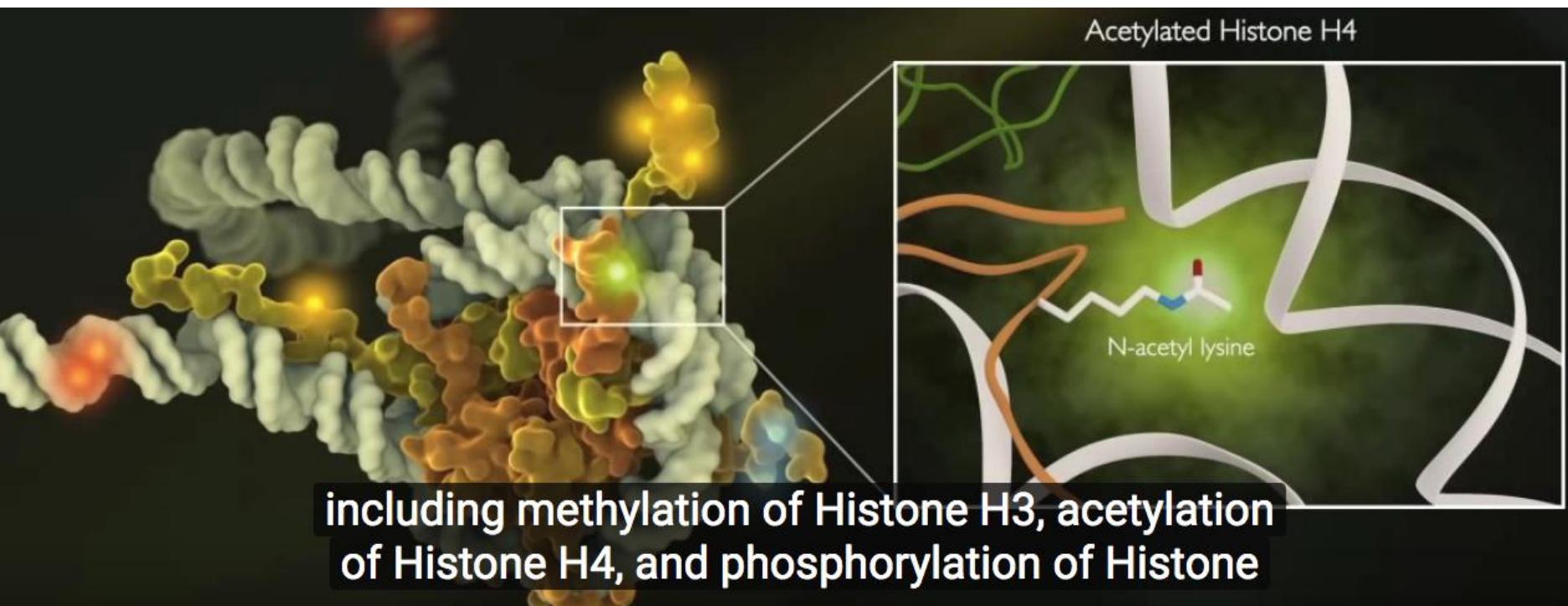
Второй слой геномной аннотации



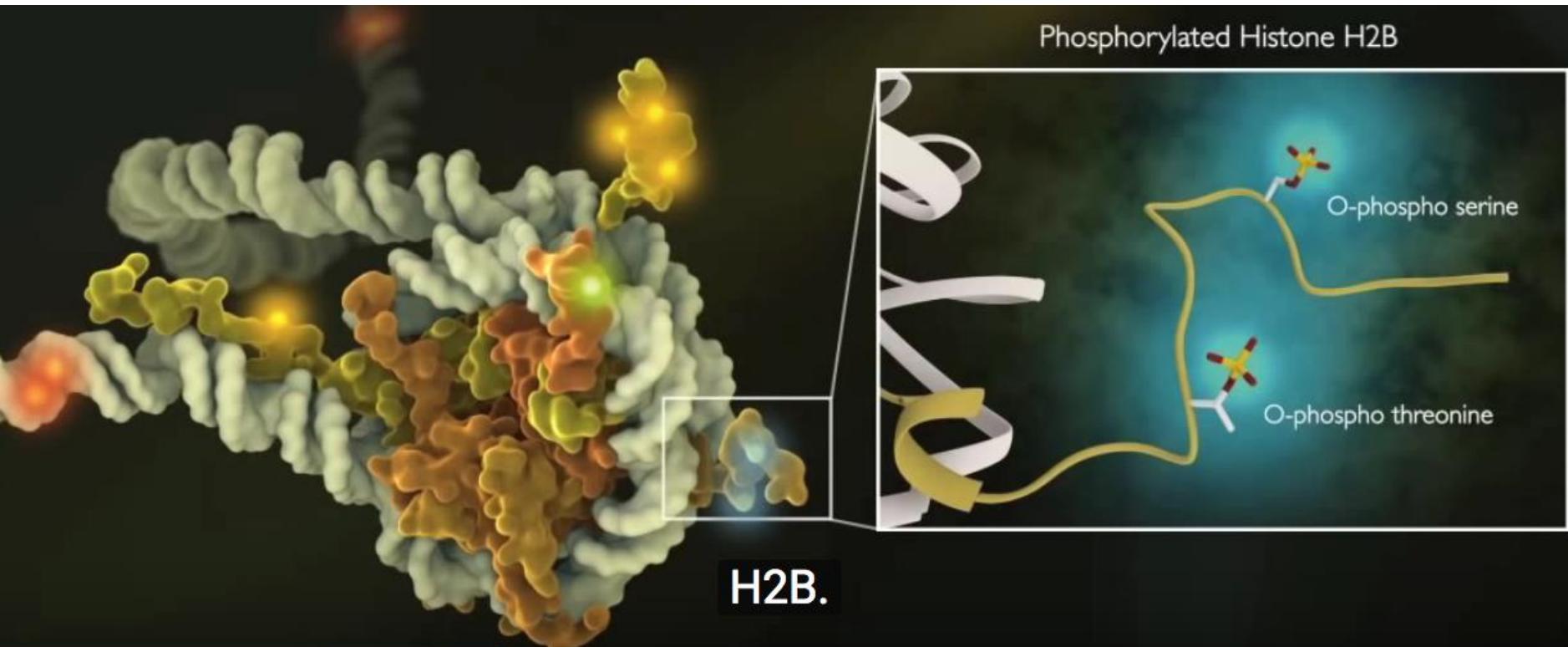
Второй слой геномной аннотации



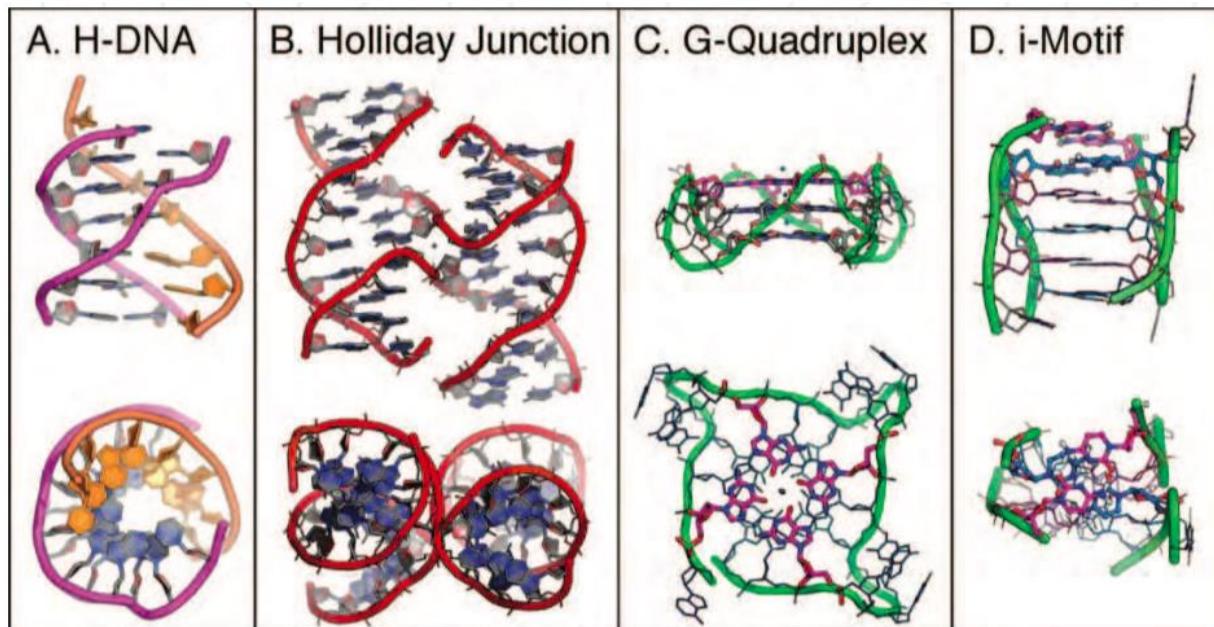
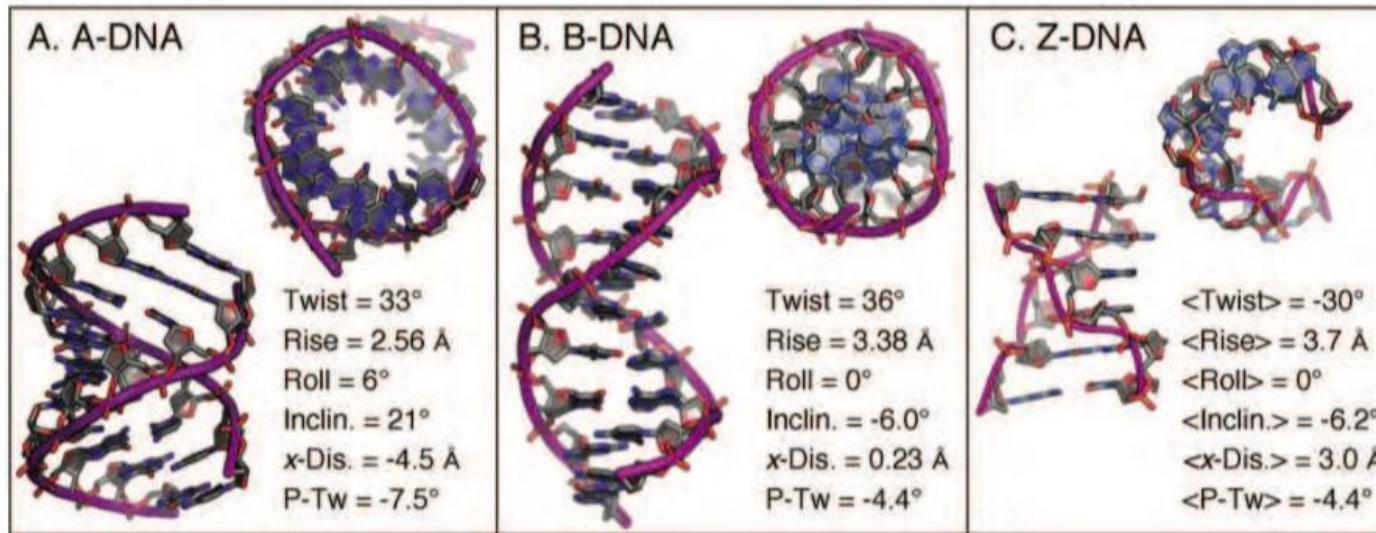
Второй слой геномной аннотации



Второй слой геномной аннотации



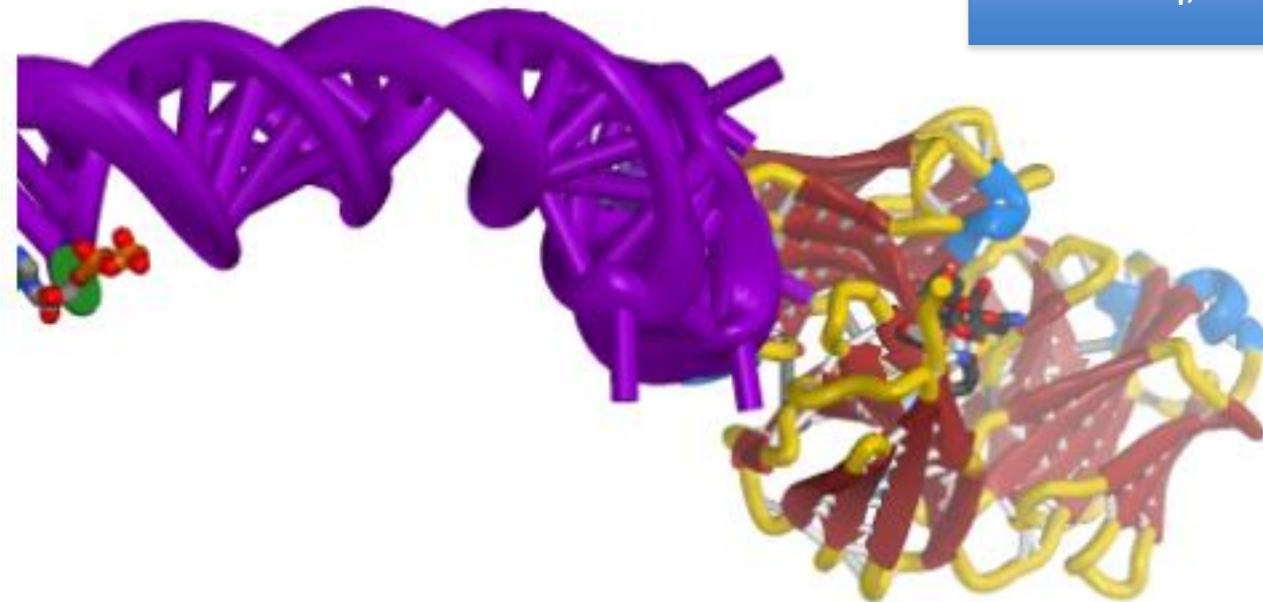
Третий слой – структуры ДНК



Третий слой геномной кодировки – структуры ДНК

- DNA structure
 - quadruplexes, triplexes, Z-DNA, stem-loops, cruciforms,

ChIP-seq, G4-seq, ssDNA-
Seq, KAS-seq



Накопление данных

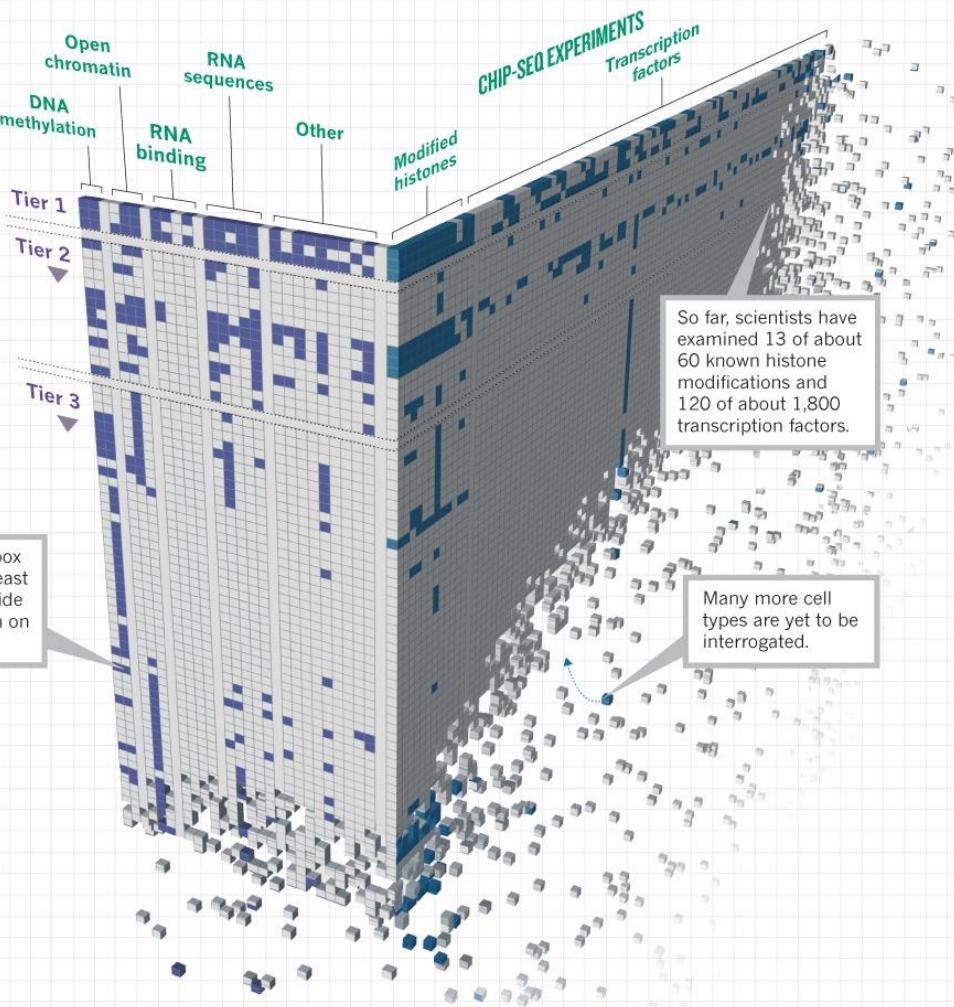
ENCODE: Encyclopedia of DNA Elements

MAKING A GENOME MANUAL

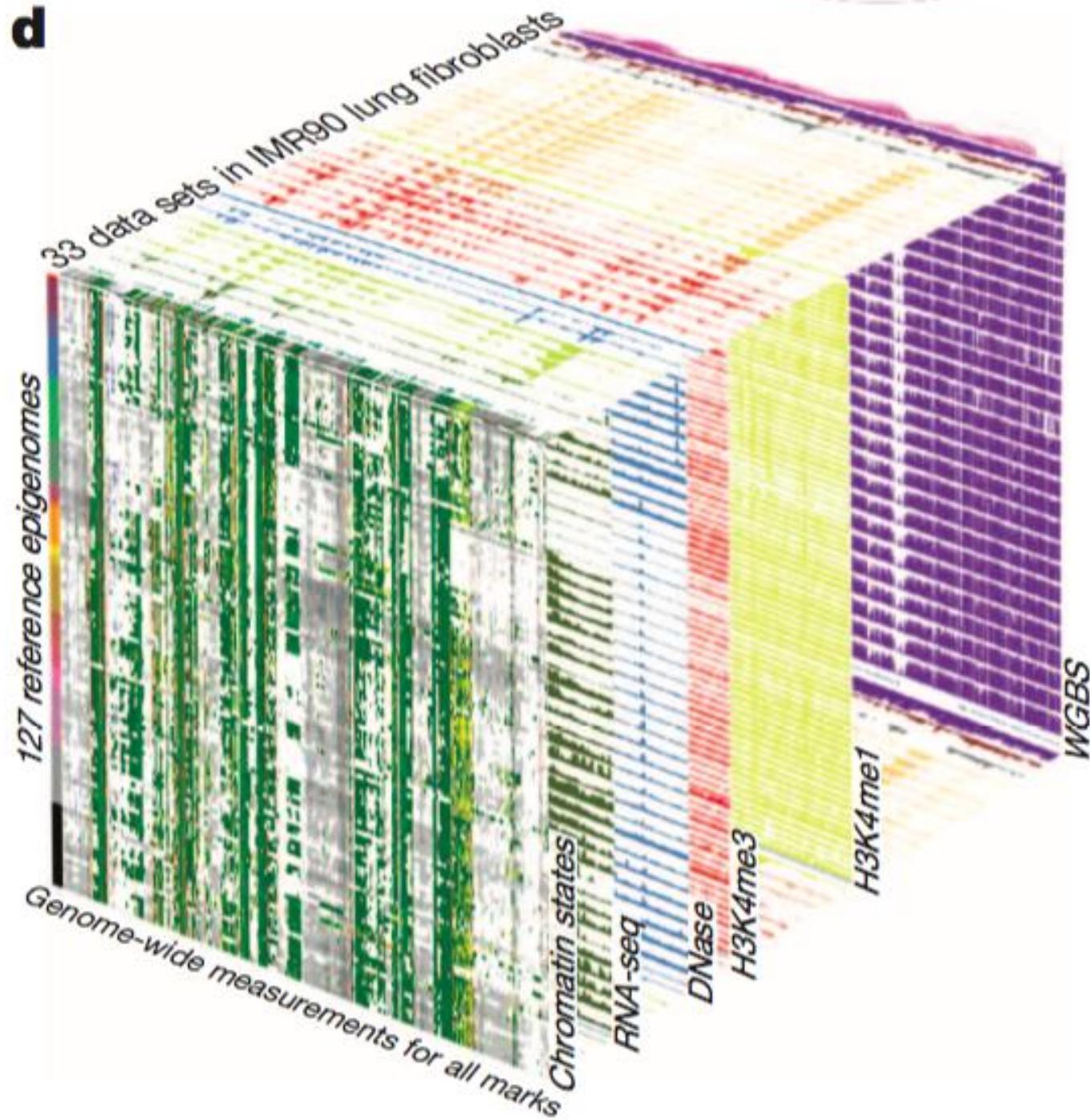
EXPERIMENTAL TARGETS
DNA methylation: regions layered with chemical methyl groups, which regulate gene expression.
Open chromatin: areas in which the DNA and proteins that make up chromatin are accessible to regulatory proteins.
RNA binding: positions where regulatory proteins attach to RNA.
RNA sequences: regions that are transcribed into RNA.
ChIP-seq: technique that reveals where proteins bind to DNA.
Modified histones: histone proteins, which package DNA into chromosomes, modified by chemical marks.
Transcription factors: proteins that bind to DNA and regulate transcription.

CELL LINES
Tiers 1 and 2: widely used cell lines that were given priority. Tier 3: all other cell types.

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.



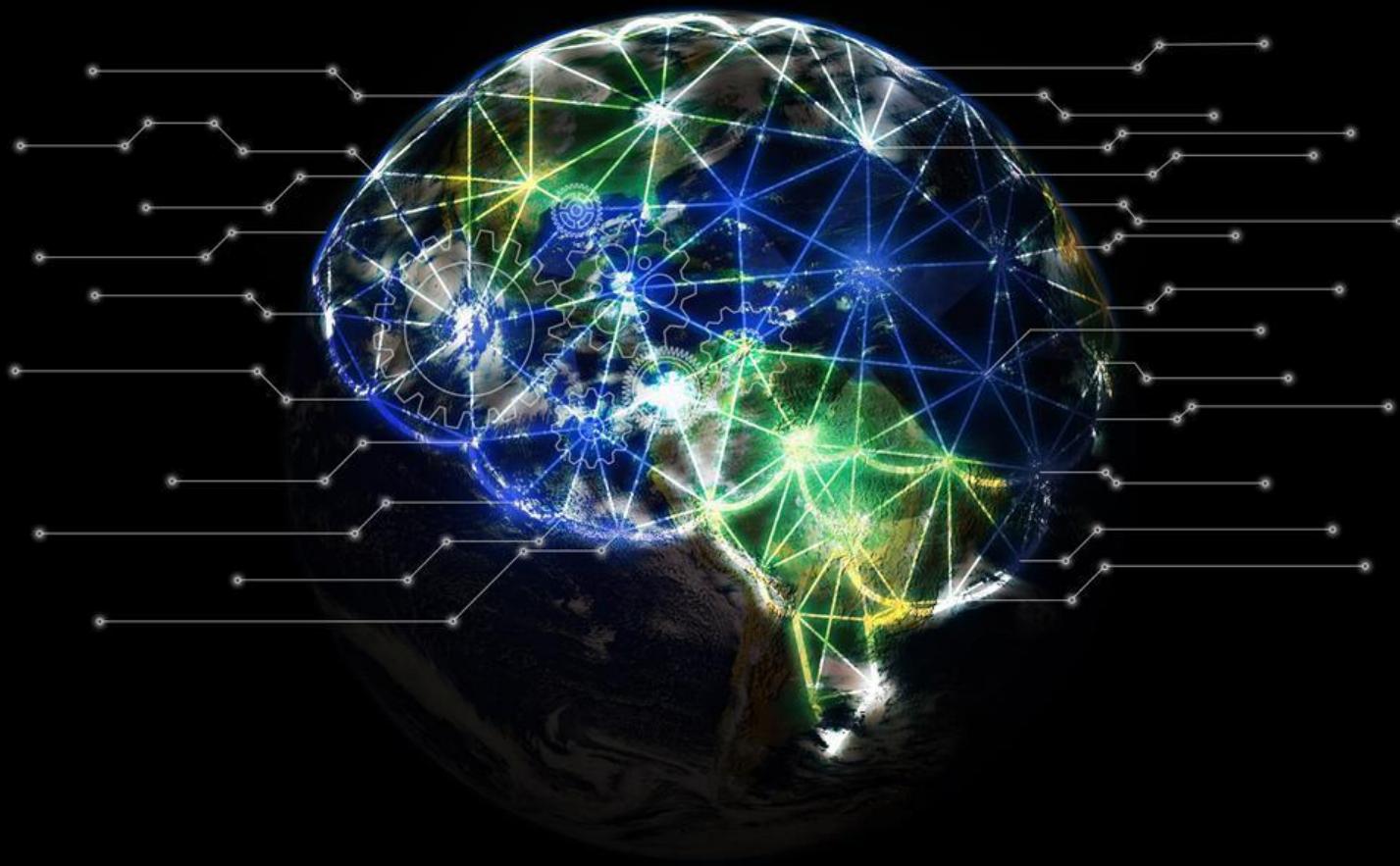
Relationship
of figure
panels
highlights
data set
dimensions



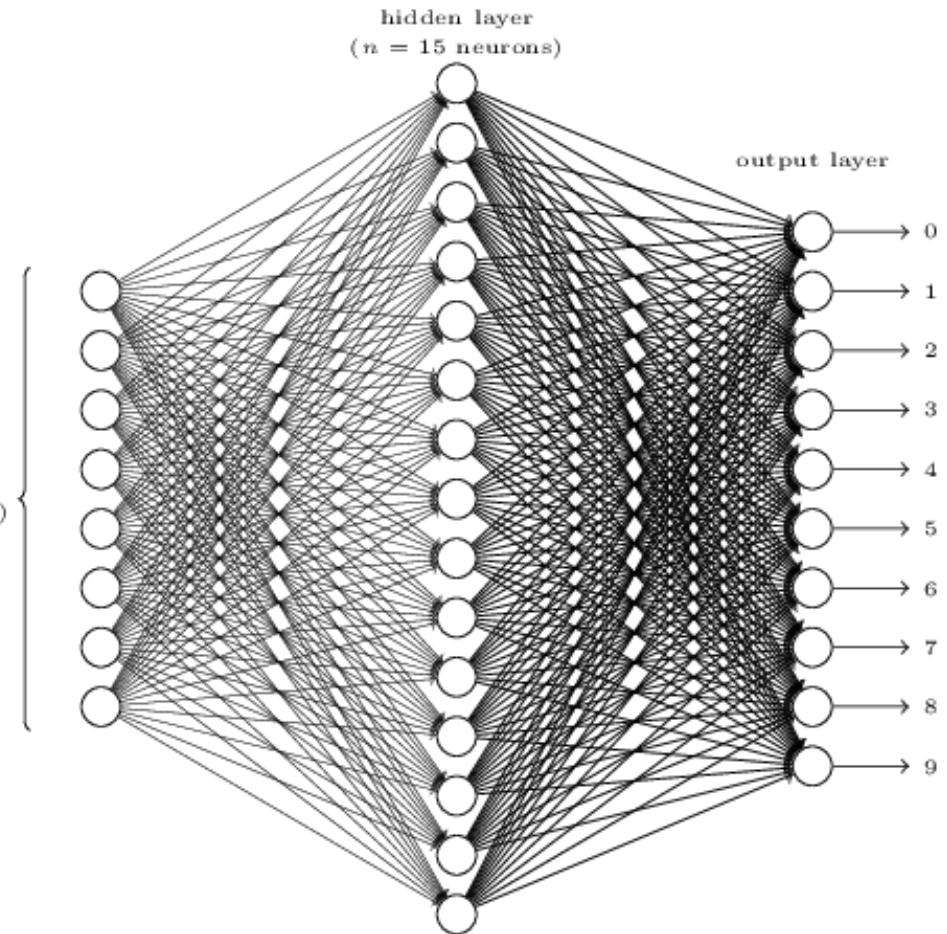
Что делать?



Машинное обучение

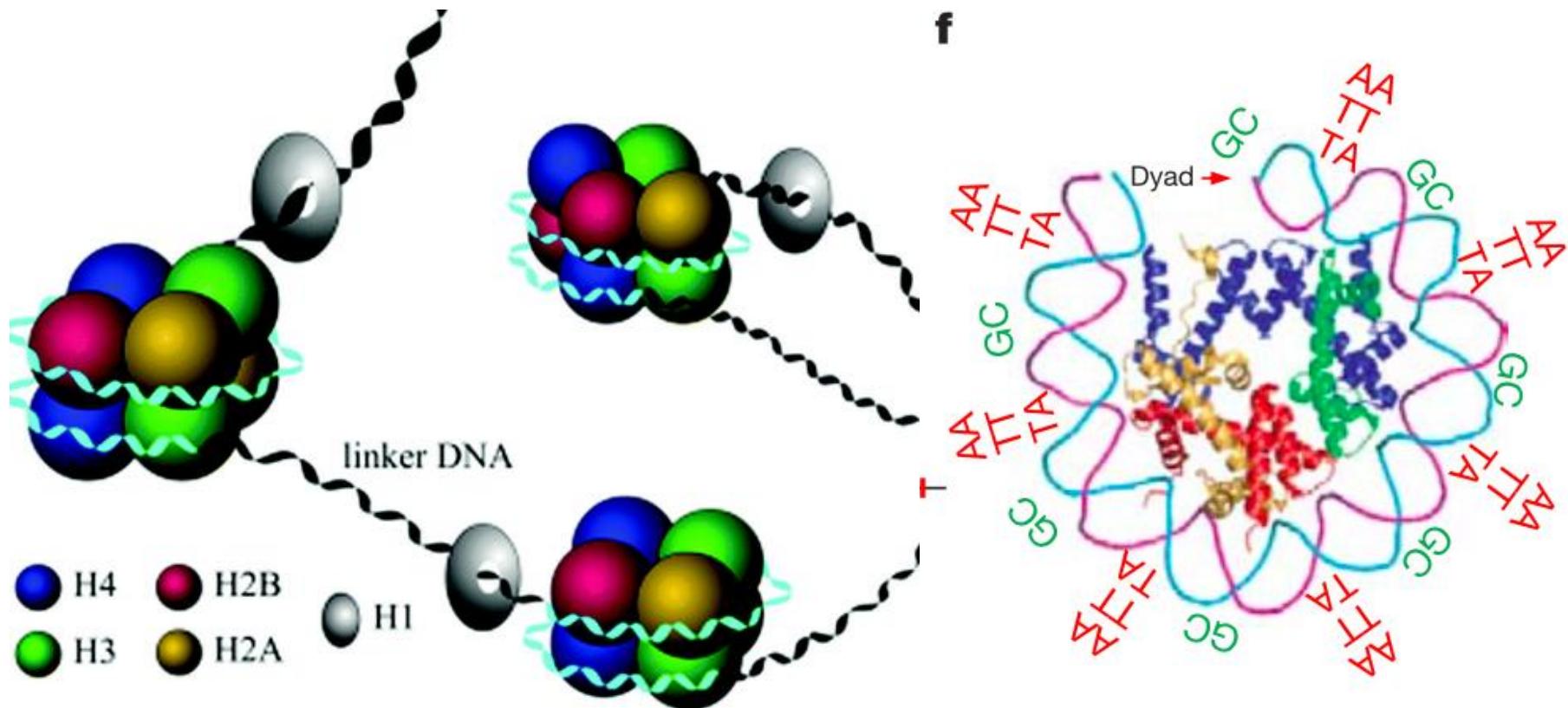


Нейронные сети



A genomic code for nucleosome positioning

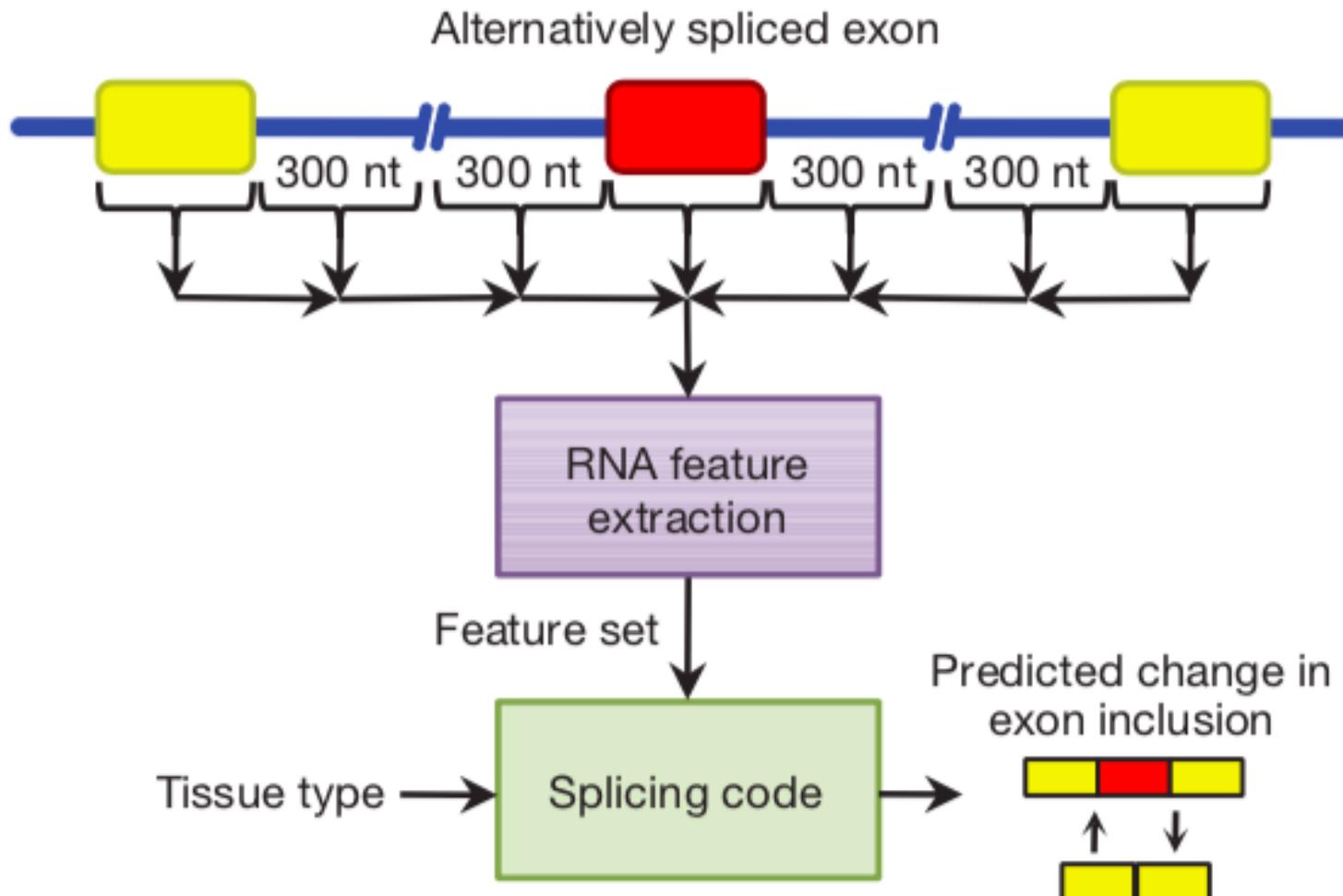
Eran Segal¹, Yvonne Fondufe-Mittendorf², Lingyi Chen², AnnChristine Thåström², Yair Field¹, Irene K. Moore², Ji-Ping Z. Wang³ & Jonathan Widom²



Nature 2006

Deciphering the splicing code

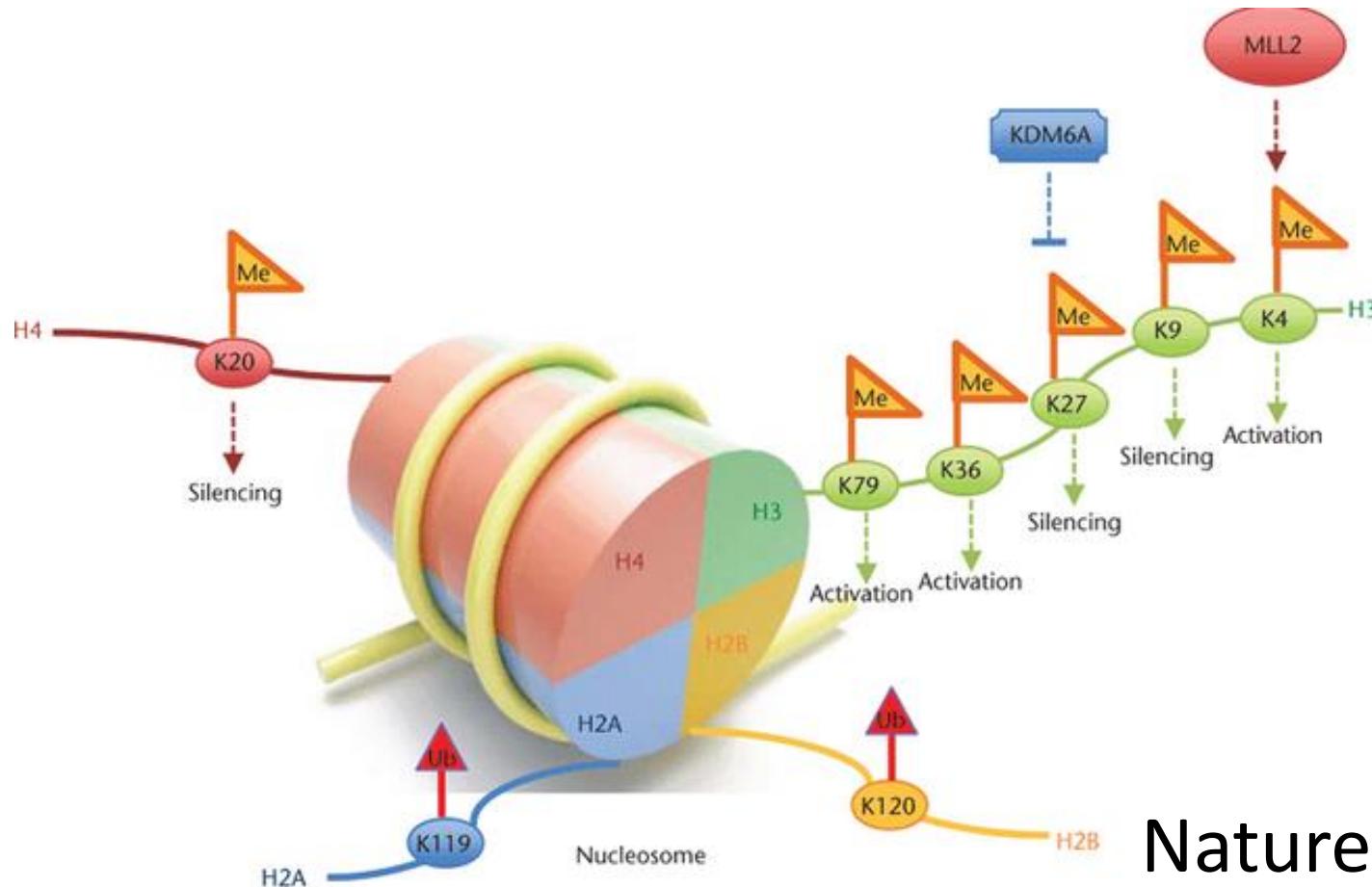
Yoseph Barash^{1,2,*}, John A. Calarco^{2*}, Weijun Gao¹, Qun Pan², Xinchen Wang^{1,2}, Ofer Shai¹, Benjamin J. Blencowe²
& Brendan J. Frey^{1,2,3}



Epigenetic code

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

Jason Ernst^{1,2} & Manolis Kellis^{1,2}

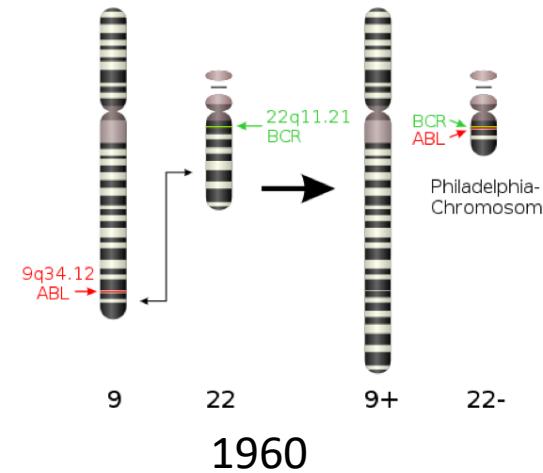
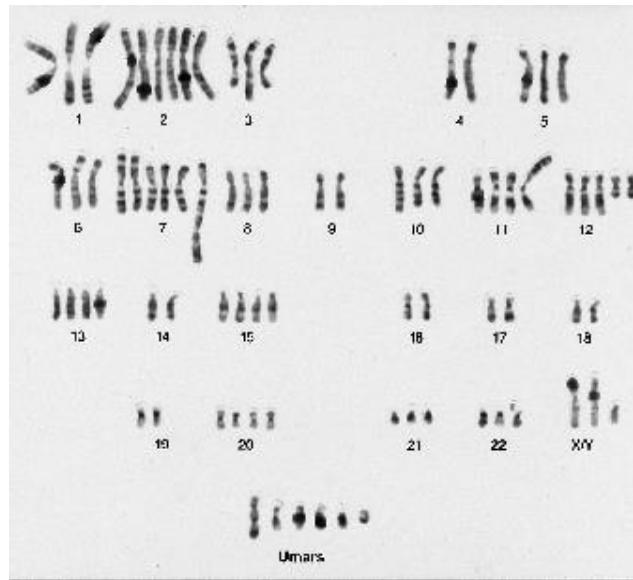
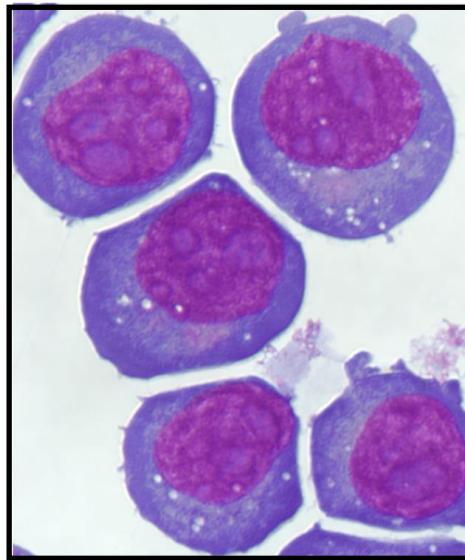


Nature 2010

PAK

РАК

Болезнь генома



‘Next Generation’ sequencing instruments are providing new opportunities for comprehensive analyses of cancer genomes

Lessons learned

➤ Heterogeneity within and across tumor types

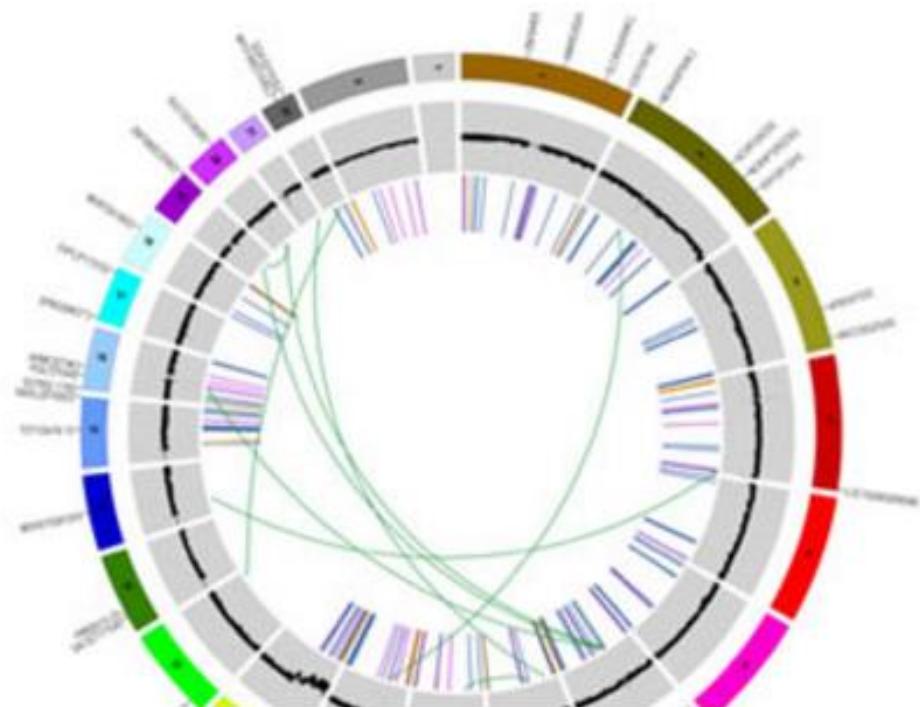
➤ Перефразируя начало Анны Карененой:

**Все нормальные геномы
похожи друг на друга,
каждый раковый геном
ненормален по-своему.**

➤Challenge in Treating Cancer

Every tumor is different

Every cancer patient is different

[Home](#)[About Cancer Genomics](#)[Cancers Selected for Study](#)[Research Highlights](#)[Publications](#)

Program Overview

Explore how The Cancer Genome Atlas works, the components of the TCGA Research Network and TCGA's place in the cancer genomics field in the Program Overview.

[Learn More ▶](#)



TCGA's Study
of Esophageal
Cancer



Researcher
Studies Own
Cancer



Cancers
Selected for
Study



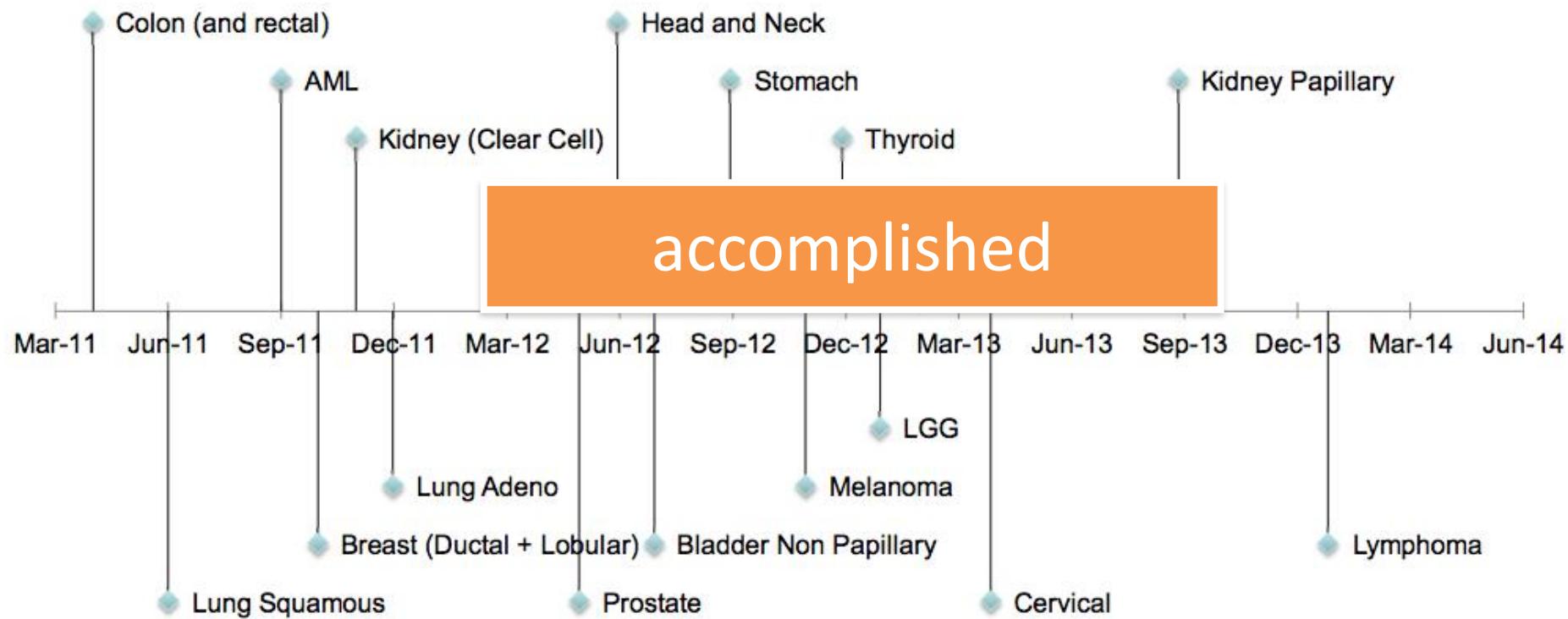
About TCGA

Active Tumor Projects

THE CANCER GENOME ATLAS



Timeline to Completion of Comprehensive Analysis for Each Tumor Project



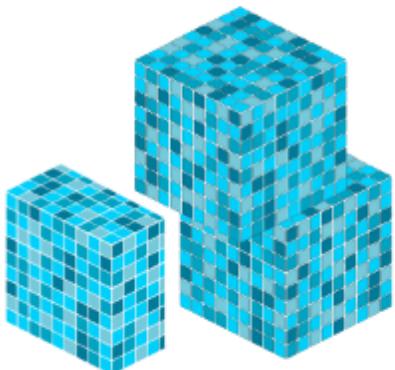
TCGA BY THE NUMBERS

TCGA produced over

2.5

PETABYTES

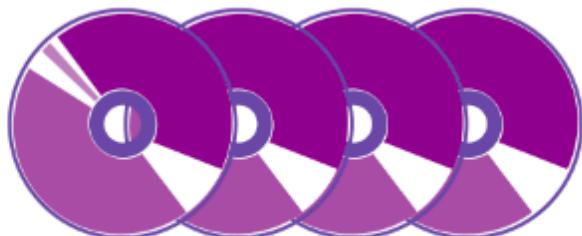
of data



To put this into perspective, **1 petabyte** of data is equal to

212,000

DVDs



TCGA data describes



33

DIFFERENT
TUMOR TYPES

...including

10

RARE
CANCERS

...based on paired tumor and normal tissue sets collected from



11,000

PATIENTS

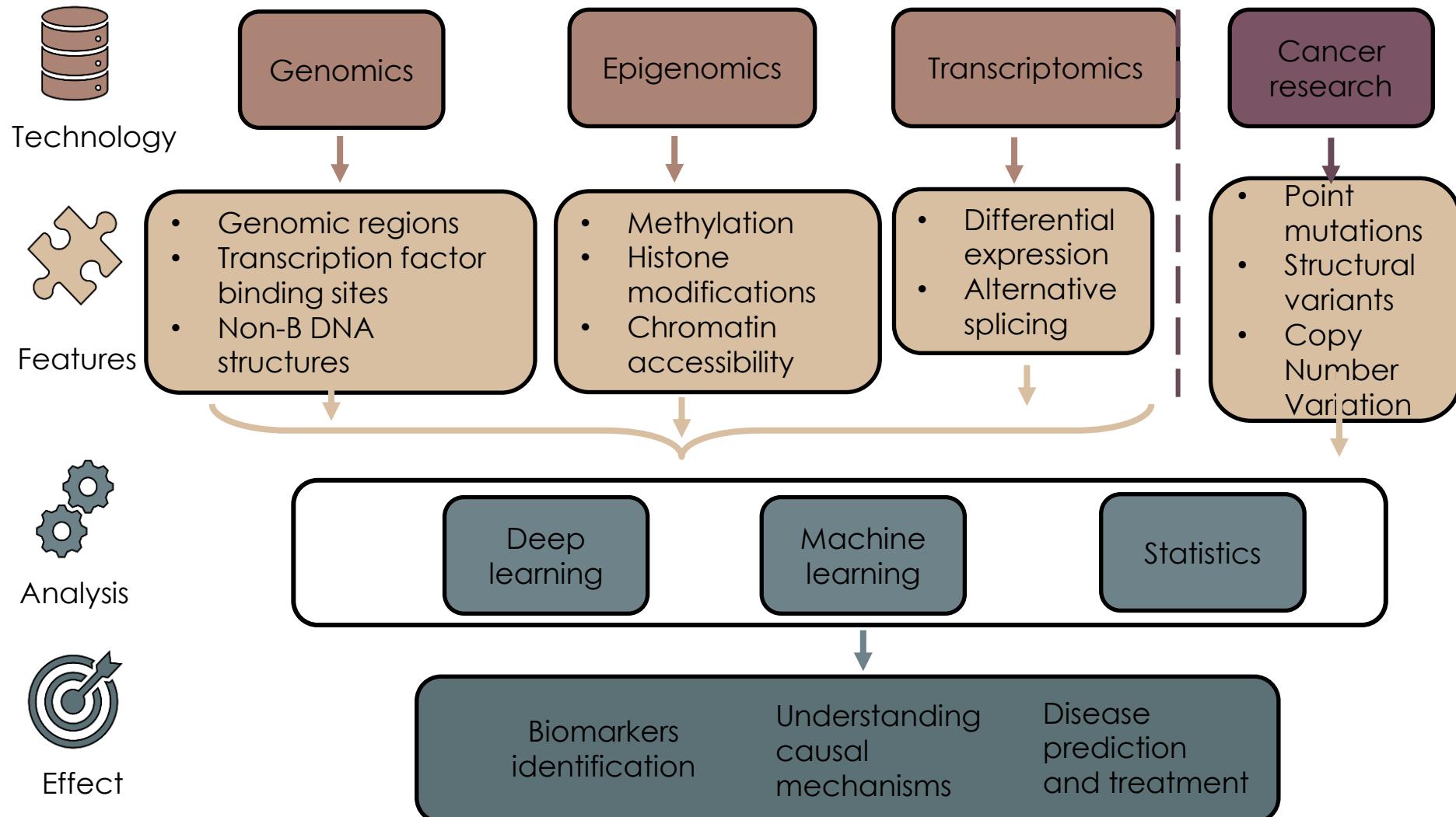
...using

7

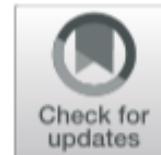
DIFFERENT
DATA TYPES



Omics approach



Tissue-specific impact of stem-loops and quadruplexes on cancer breakpoints formation



Kseniya Cheloshkina and Maria Poptsova* 

BMC Cancer (2019) 19:434

487,425 breakpoints from 2234 samples covering 10 cancer types

stem-loop- based model best explains the blood, brain, liver, and prostate cancer breakpoint hotspot profiles

quadruplex- based model has higher performance for the bone, breast, ovary, pancreatic, and skin cancer.

For the overall cancer profile and uterus cancer the joint model shows the highest performance.

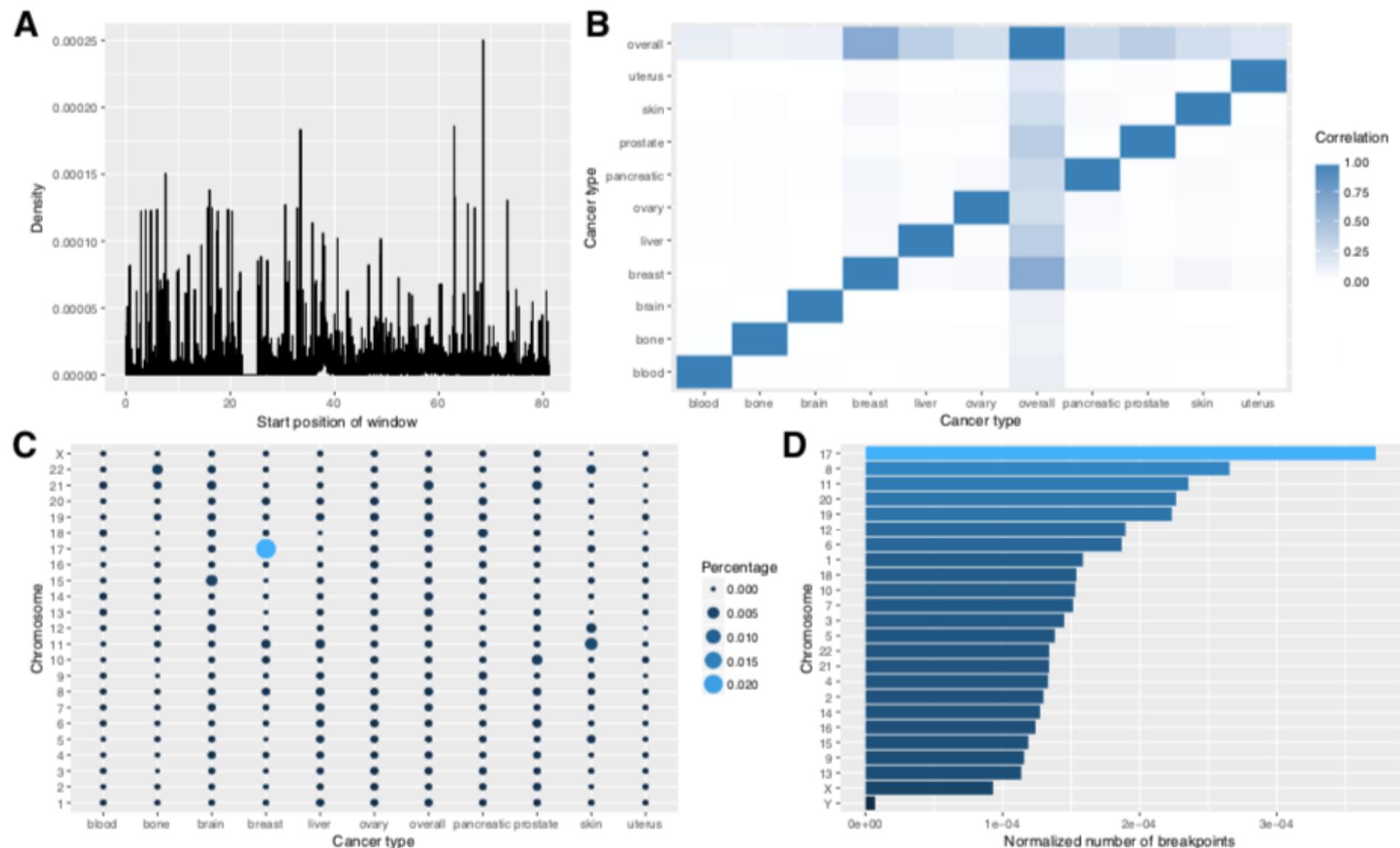


Fig. 1 a Number of breakpoint hotspots by chromosome and type of cancer for 0.1% labeling type. b Correlation between different cancer profiles. c Breakpoint density in general cancer profile across chromosome 17 (in Mb). d Number of breakpoints normalized by chromosome length

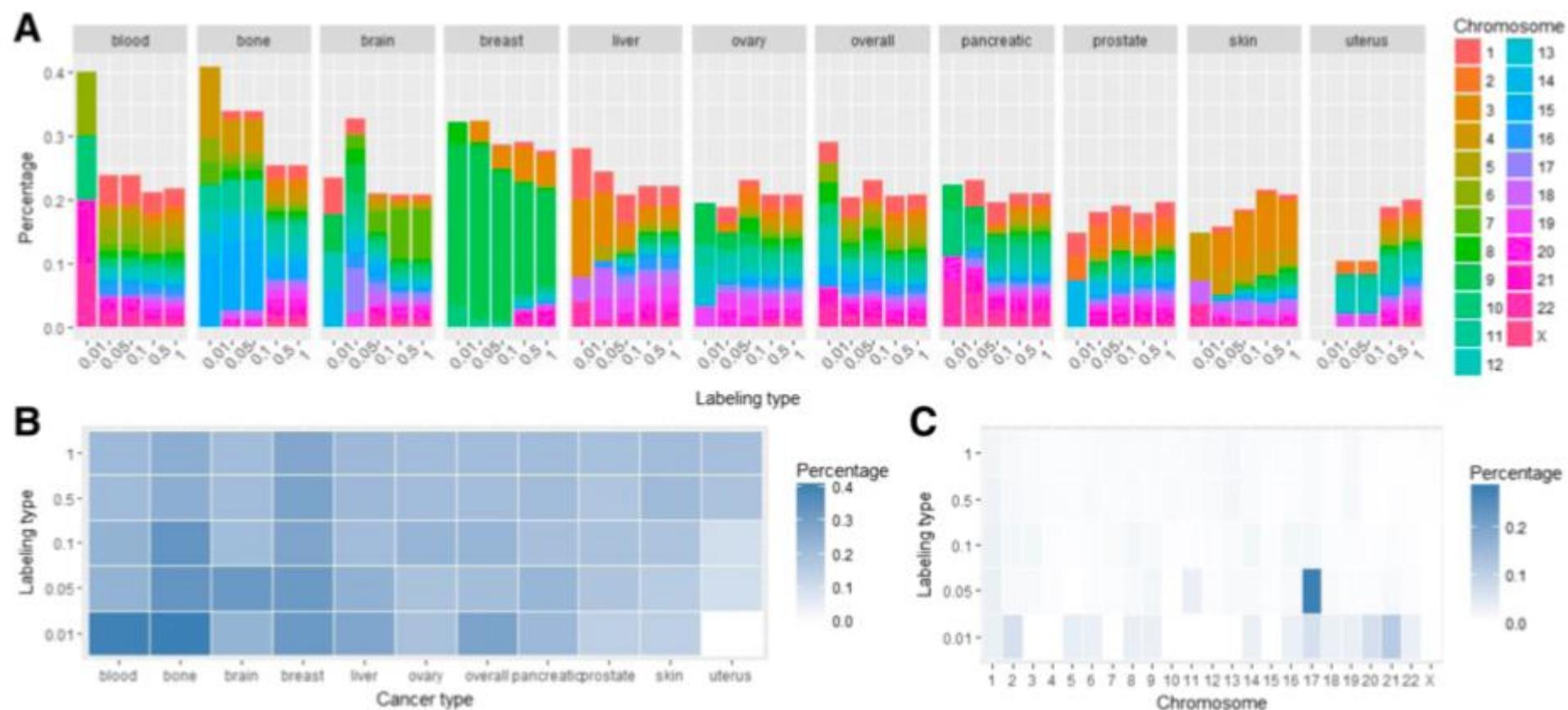
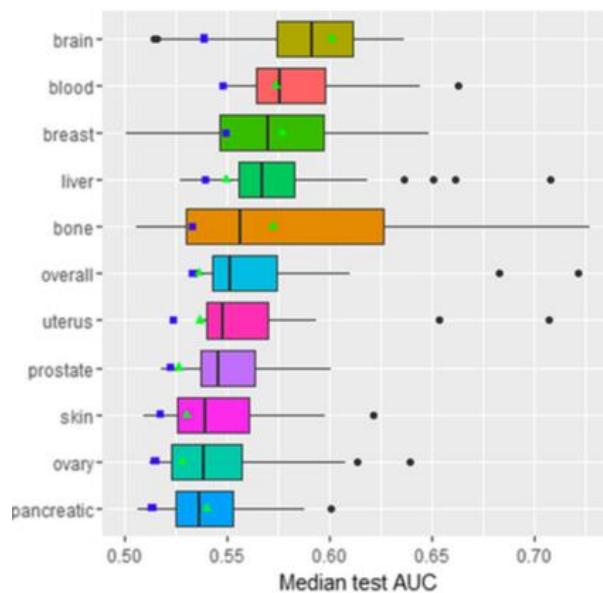
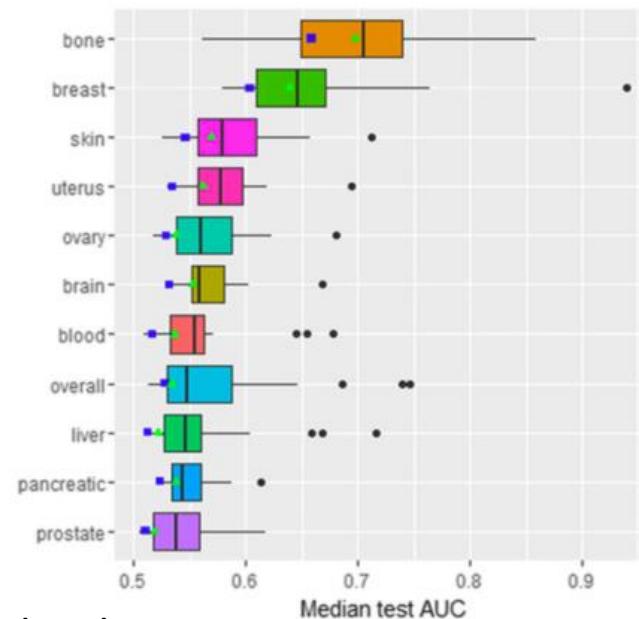


Fig. 2 a Percentage of breakpoint hotspots intersecting with whole gene regions in different cancer types for 5 different hotspot labeling types. b Heatmap of the percentage of the breakpoint hotspots' intersections with whole gene regions across different cancer types. c Heatmap of the percentages of all breakpoint hotspots' intersections in all cancer types stratified by different chromosomes

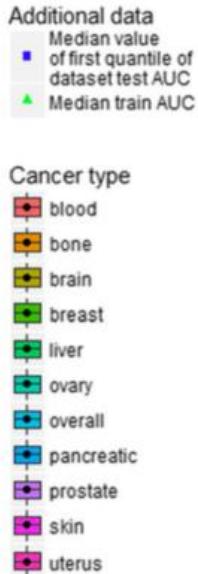
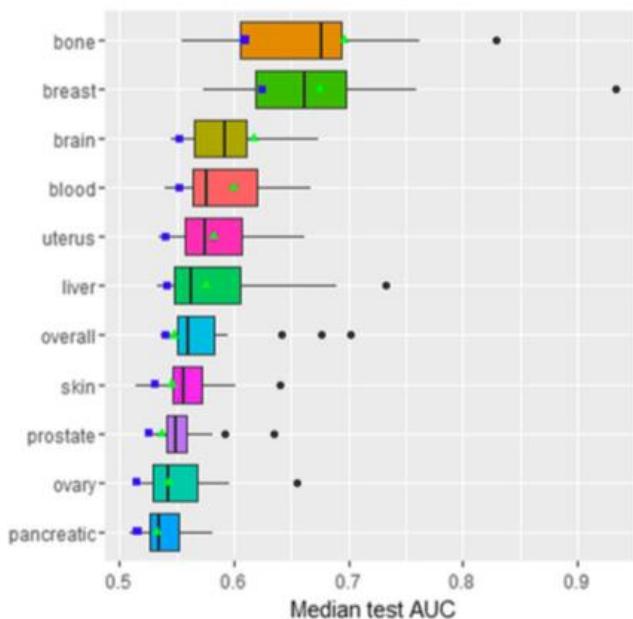
Stem-loops



Quadruplexes



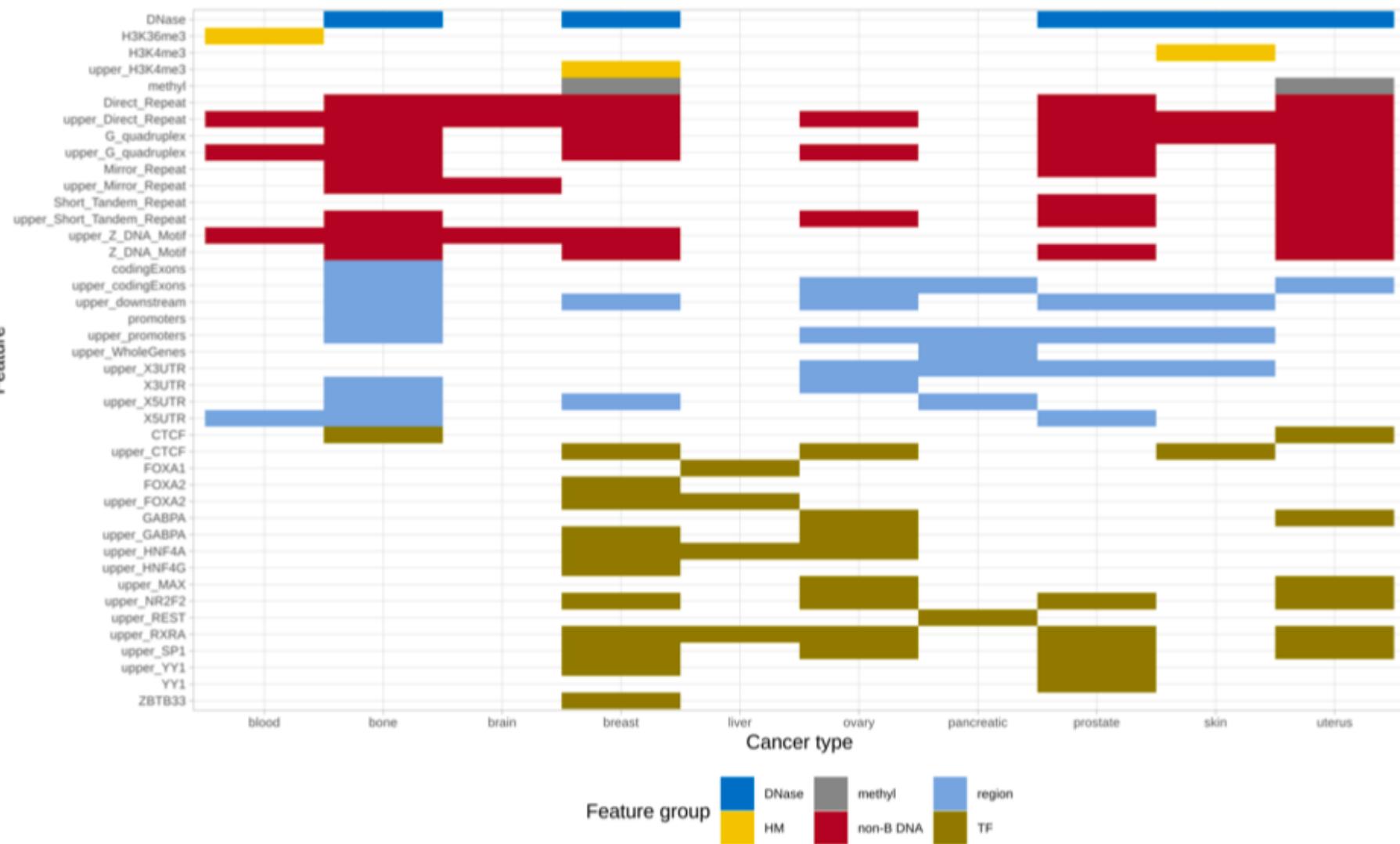
Stem-loops + quadruplexes

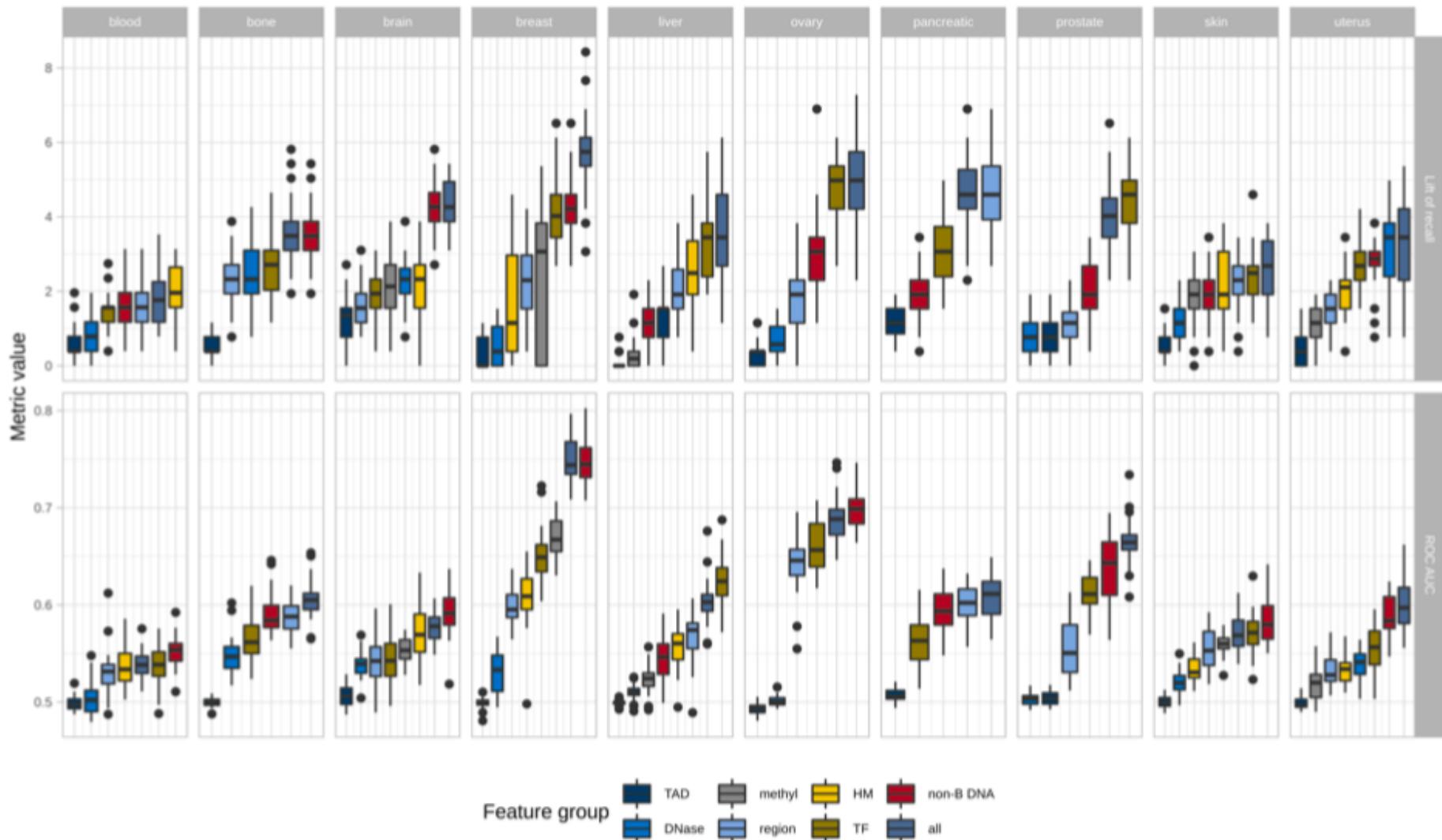


Comprehensive analysis of cancer breakpoint hotspots with machine learning models

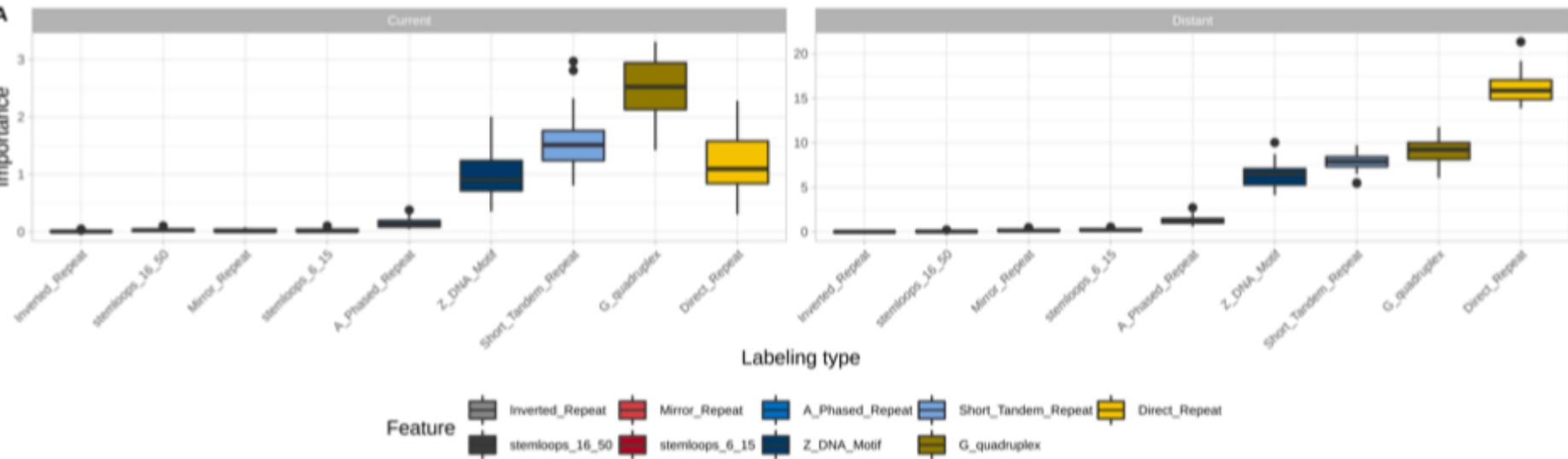
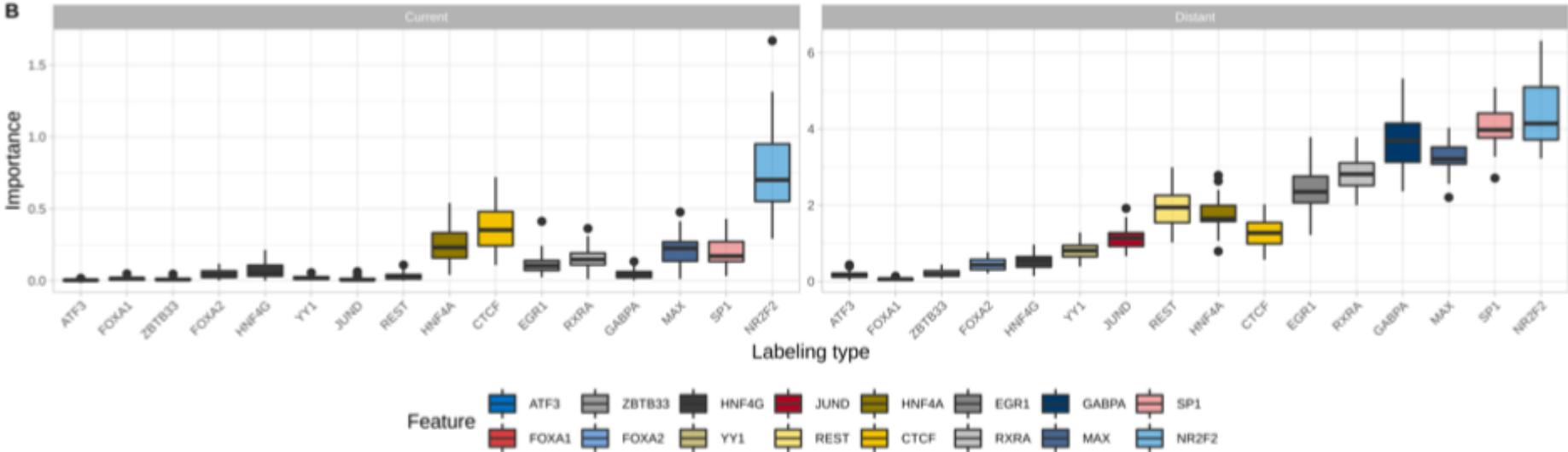
- non-B DNA structures
- genomic regions
- chromatin structure
- transcription factor binding sites
- epigenetic markers

Вклад разных групп





Contribution of different feature groups separately and all together.

A**B**

Conclusions

- Non-B DNA structures are the major contributors followed by transcription factor binding sites and genomic regions.
- Contributions of individual features inside the groups revealed G-quadruplexes and repeats from non-B DNA group, CTCF, GABPA, RXRA, SP1, MAX and NR2F2 from transcription factors group and promoter and 5'UTR regions.



Рэй Курцвейл

Технический директор Google

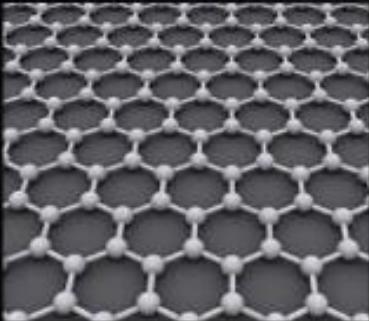
The future of the world: a forecast up to the year 2099

the **DAWN** *of the* **SINGULARITY**

The following predictions were made by Ray Kurzweil
in his book *The Singularity is Near*.

Kurzweil, now the Director of Engineering at Google,
had made 147 predictions since the 1990's and has maintained
an astonishing 86% accuracy rate.

2019-2029



Three-dimensional nanotube lattices are the dominant computing substrate



The digital world makes paper books and documents almost completely obsolete



Total power of all computers is comparable to total brainpower of the human race



Creative AI is now capable of making complex art and music



Autonomous vehicles now dominate our roads



Humans begin to develop deep relationships with AI



Computers are embedded everywhere in our environment (furniture, jewelry, walls, etc)

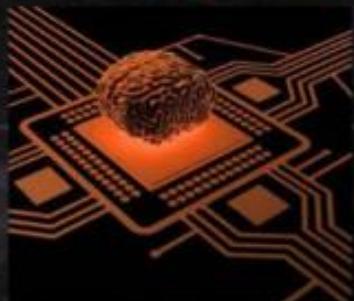


Language translation machines are now routinely used in conversation

2029-2039



The manufacturing, agricultural and transportation sectors of the economy are almost entirely automated



Computers are now capable of autonomously learning and creating new knowledge



A \$1,000 USD personal computer is now 1,000x more powerful than the human brain



Advanced brain mapping leads to hundreds of distinct subregions in the brain being identified



Artificial intelligences claim to be conscious and openly petition for recognition of this fact



VR eyeglasses and headphones are replaced with computer implants

- 2020 – Персональные компьютеры достигнут вычислительной мощности, сравнимой с человеческим мозгом.
- ...
- 2028 – Солнечная энергия станет настолько дешевой и распространенной, что будет удовлетворять всей суммарной энергетической потребности человечества.
- ...
- 2033 – Самоуправляемые автомобили заполнят дороги.

- 2036 – Используя подход к биологии, как к программированию, человечеству впервые удастся запрограммировать клетки для лечения болезней, а использование 3D-принтеров позволит выращивать новые ткани и органы.
- 2037 – Гигантский прорыв в понимании тайны человеческого мозга. Некоторые из алгоритмов будут расшифрованы и включены в нейронные сети компьютеров.

Нобелевская премия по физиологии и медицине

- 203X год - за открытие алгоритмов работы генома
- 203X год – за программирование клеток

**МИР
БЕЗ
РАКА**

https://www.hse.ru/ma/adbm/ Getting Started

Национальный исследовательский университет «Высшая школа экономики» → Образовательные программы магистратуры (Москва) → Факультет компьютерных наук → Магистерская программа «Анализ данных в биологии и медицине»

Магистерская программа

Анализ данных в биологии и медицине

Магистратура предназначена для бакалавров в области прикладной математики и/или анализа данных и специалистов, работающих в области наук о жизни и желающих приобрести или усовершенствовать навыки математического моделирования и анализа медико-биологических данных.

Программа нацелена на подготовку будущих лидеров биоинформационических исследований, обладающих компетенциями для разработки и применения на практике вычислительных методов с целью решения задач в различных областях биологии и медицины.

[Задать вопрос о программе](#)

О программе ▾ Абитуриентам ▾ Студентам ▾

Академический руководитель
Гельфанд Михаил Сергеевич

Научно-учебная лаборатория биоинформатики

← → ⌂ ⌂ https://cs.hse.ru/big-data/bioinform/ ⌂ ... ⌂ ☆ ⌂ НУЛ биоинформатики

Most Visited Getting Started

- [О лаборатории >](#)
- [Проекты >](#)
- [Публикации >](#)
- [Сотрудники >](#)
- [Семинары по биоинформатике >](#)
- [Научные семинары и конференции >](#)
- [Семинар по методам машинного обучения в биоинформатике >](#)

Руководитель лаборатории



Попцова Мария
Сергеевна
тел.: +7(962) 909-51-53
mpoptsova@hse.ru

Менеджер

**Лаборатория объявляет набор
стажёров-исследователей!**

Создана в 2018 году для развития направления биоинформатики на ФКН. Благодаря революции в высокопоточных технологиях секвенирования биоинформатика стала наукой о больших данных геномики. Основными направлениями образовательной и научной деятельности лаборатории являются фундаментальные исследования роли вторичных структур ДНК в функционировании генома, организации хроматина и ДНК-белковых взаимодействий. Проходя стажировку в лаборатории, студенты смогут погрузиться в активную научную деятельность на ранних стадиях обучения. Лаборатория сотрудничает с ведущими лабораториями биоинформатики мирового уровня.

ФКН провел первую международную школу по машинному обучению в биоинформатике





Спасибо за внимание

