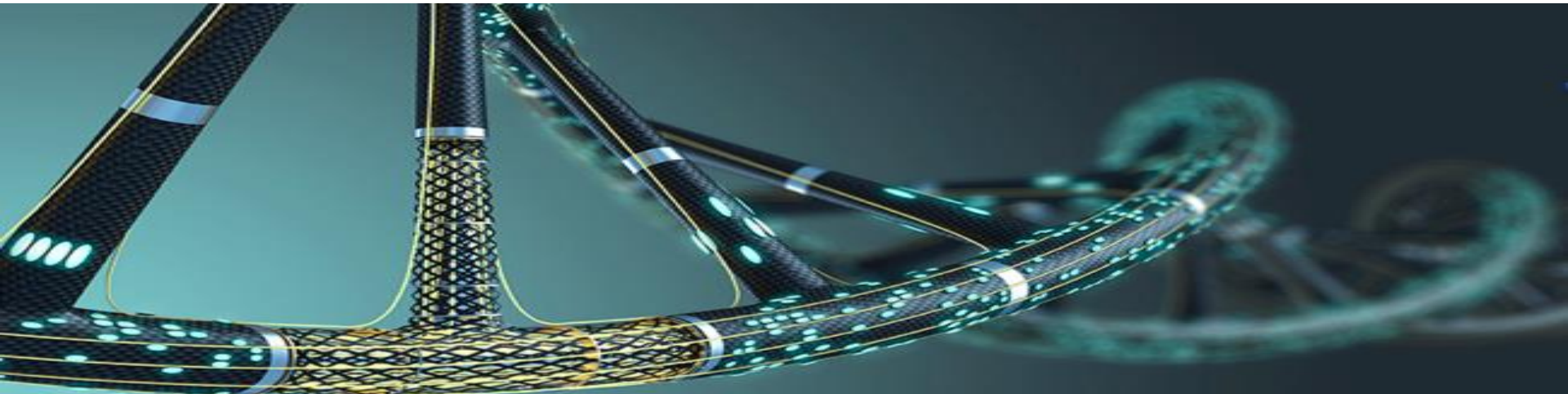
HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

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?

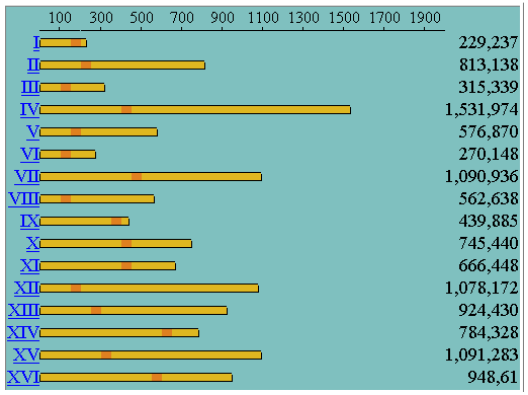
ЧТО ТАКОЕ ГЕНОМ?

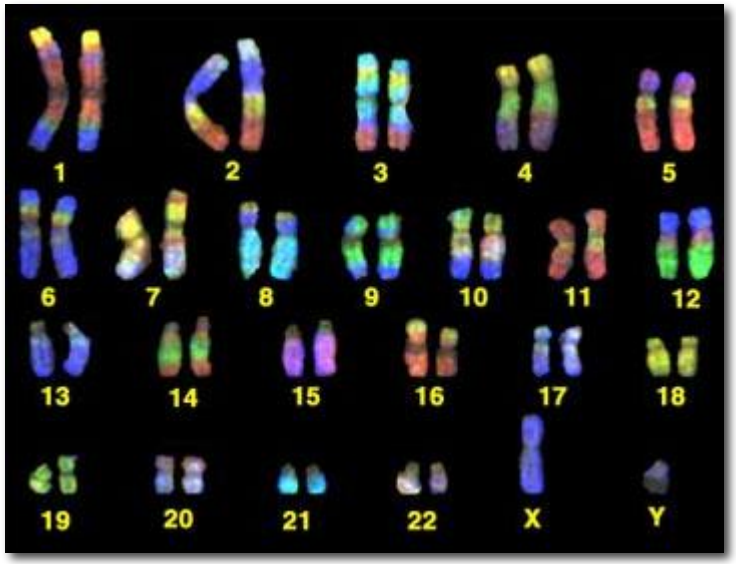
- Для биолога
- Для биои

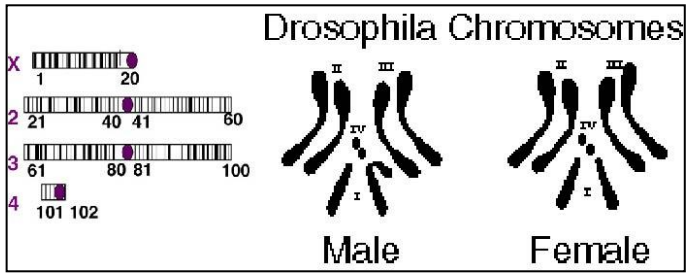
Ряд Нобелевских премий 1950-1970

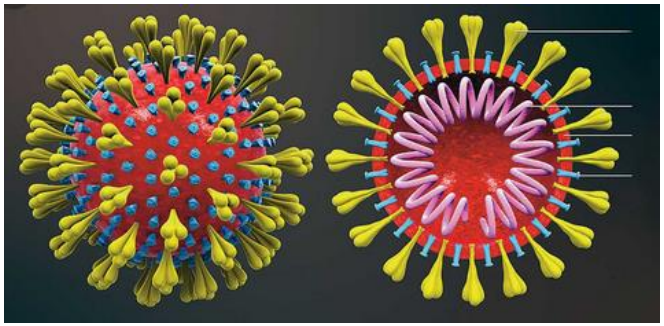
Геномы разных организимов

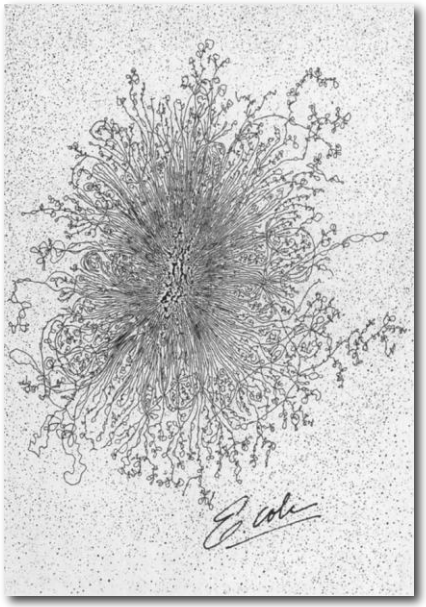
***Saccharomyces cerevisiae* complete genome**











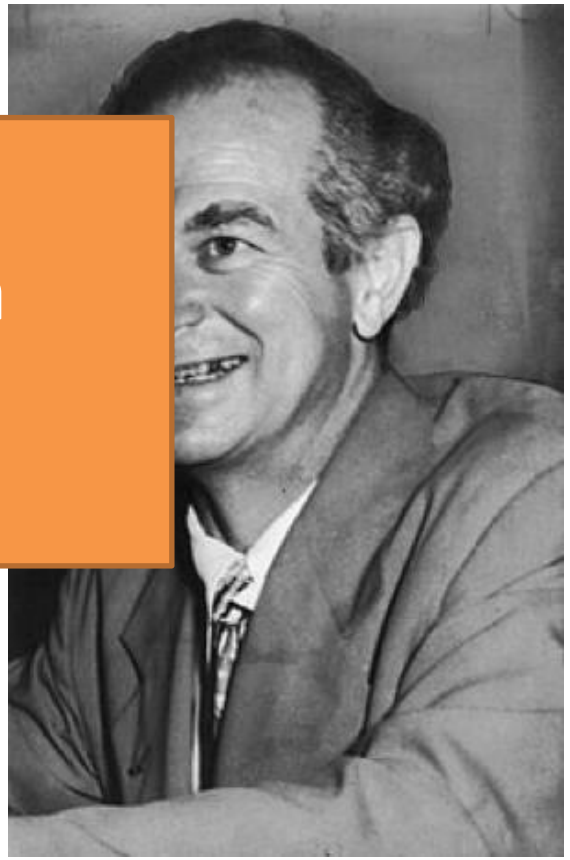
Обитатели клетки-города

Нобелевская премия по химии 1954

Лайнус Полинг

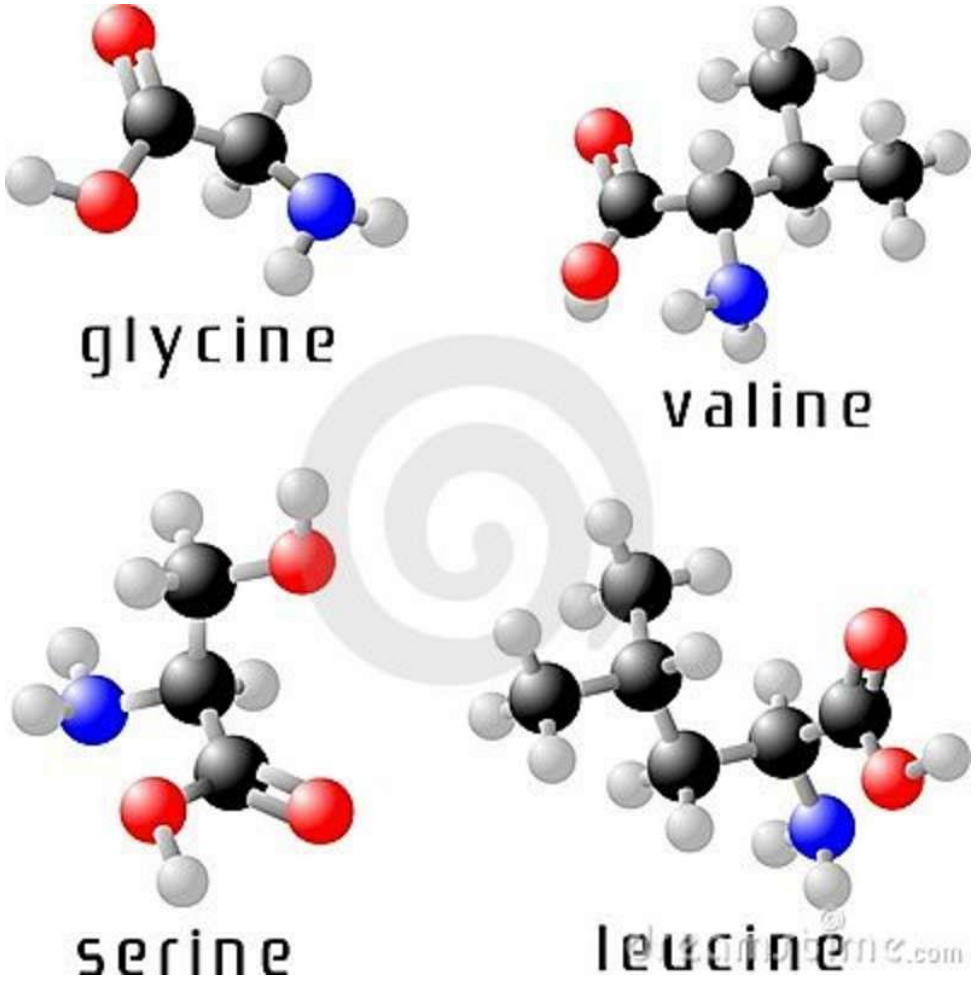
Альфа-субъединица АТФ-синтетазы

Белки - текст, написанный на алфавите
Из 20 букв

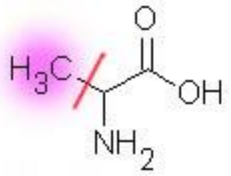


Вторичная структура – альфа-спираль и бета-лист

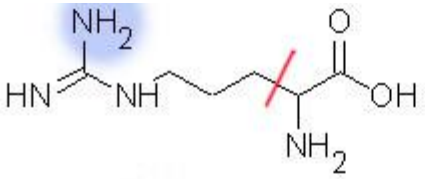
Белки сделаны из 20 аминокислот



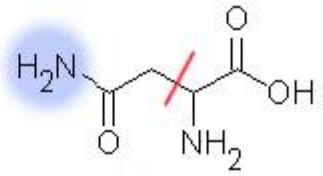
Over 100 amino acids exist in nature



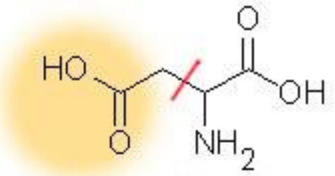
Аланин (Ala)



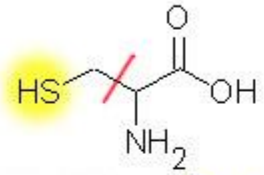
Аргинин (Arg)



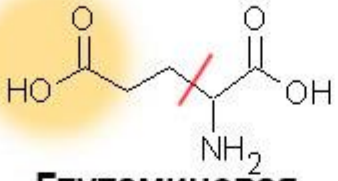
Аспарагин (Asn)



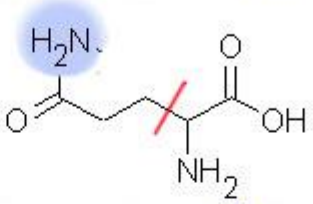
Аспарагиновая кислота (Asp)



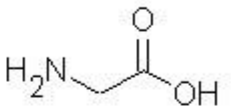
Цистеин (Cys)



Глутаминовая кислота (Glu)



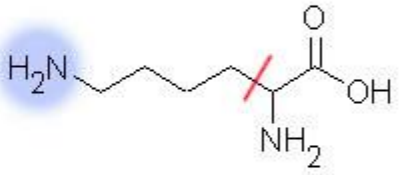
Глутамин (Gln)



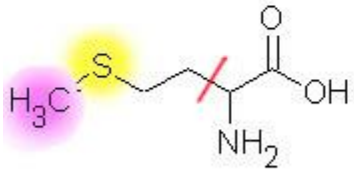
Глицин (Gly)

Белок - текст, написанный на алфавите из 20 букв

Гистидин (His)

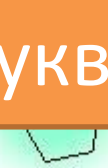


Лизин (Lys)



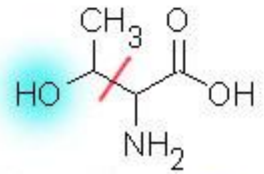
Метионин (Met)

Фенилаланин (Phe)

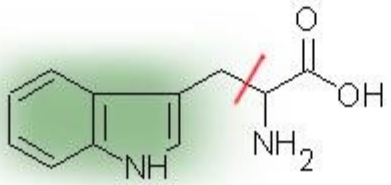


Пролин (Pro)

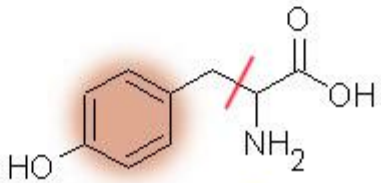
Серин (Ser)



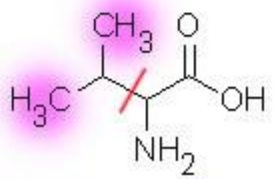
Треонин (Thr)



Триптофан (Trp)

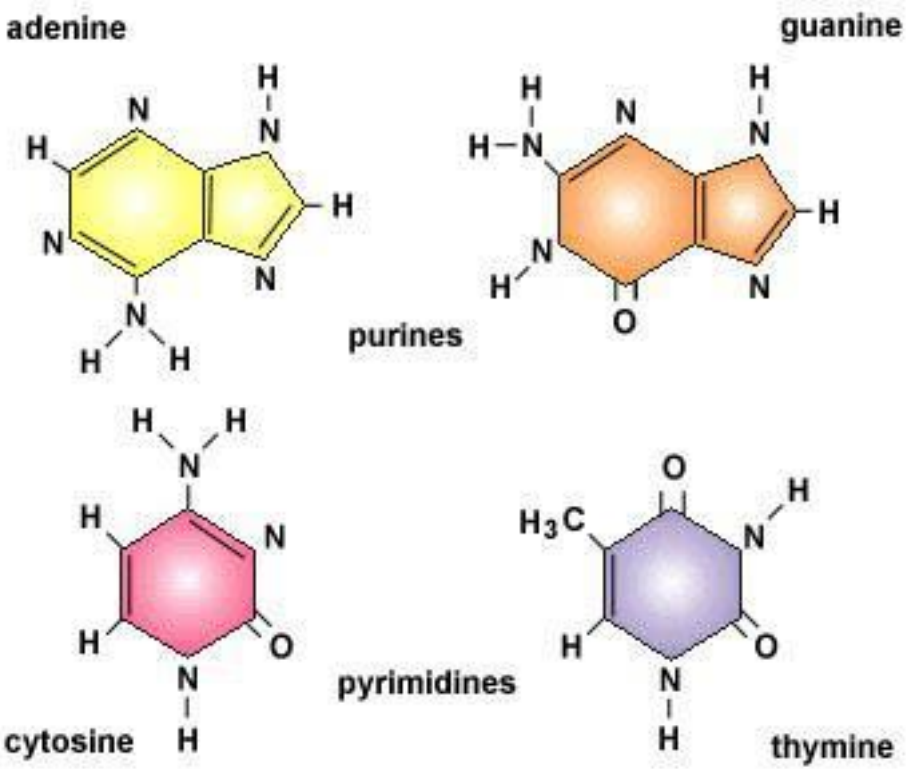


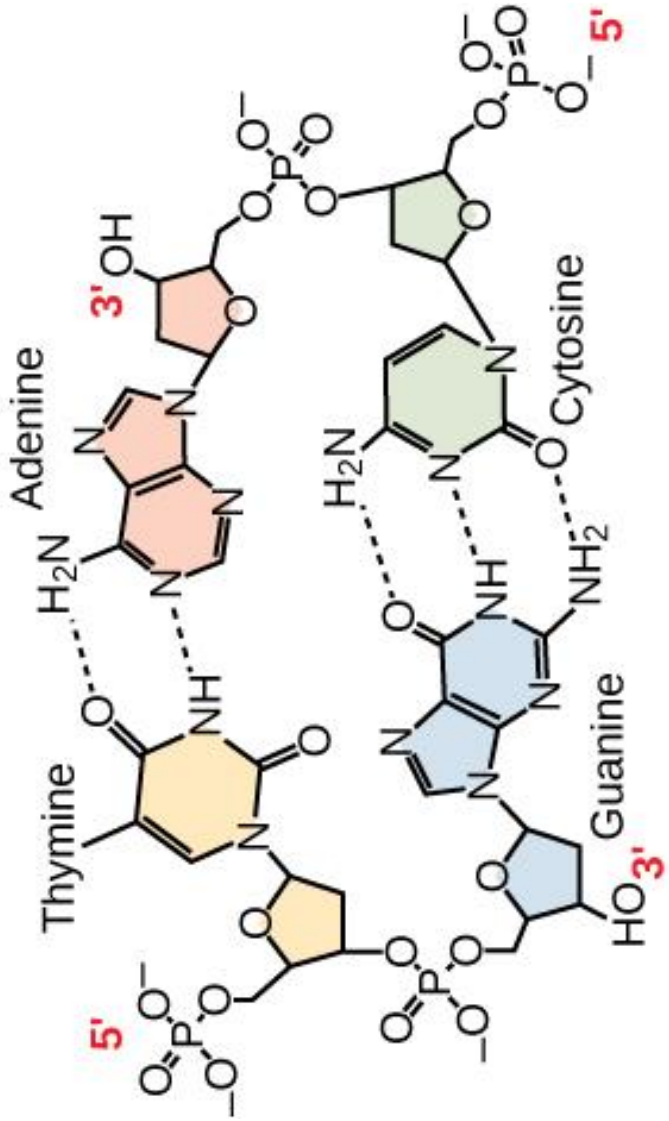
Тирозин (Tyr)



Валин (Val)

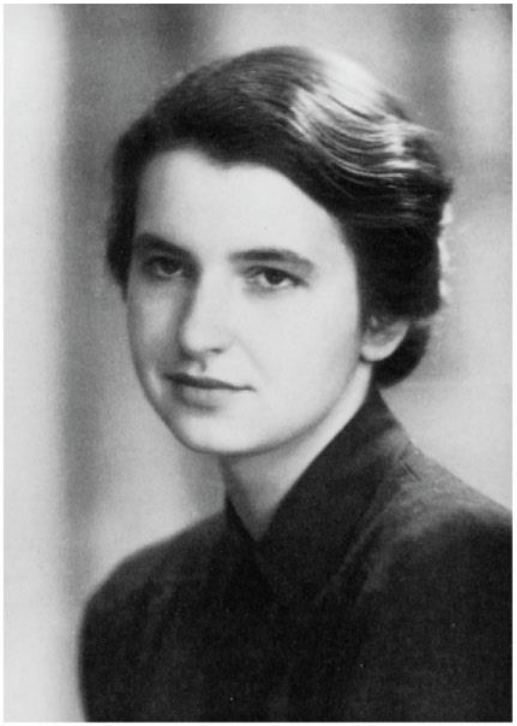
ДНК состоит из 4 нуклеотидов



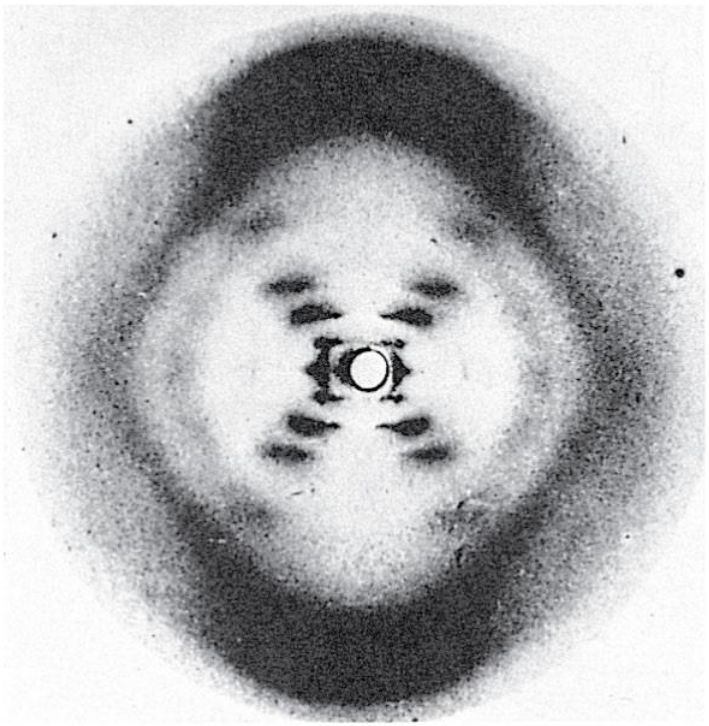


Good Quality DNA X-Ray images

- Морис Уилкинс и Розалинд Франклин
- .

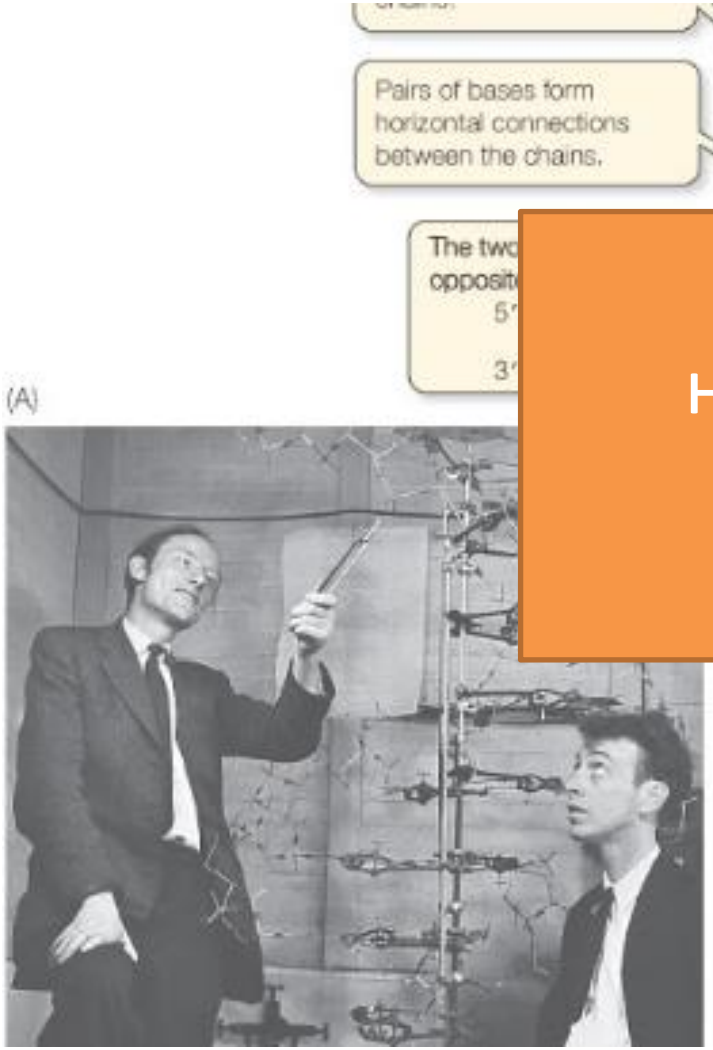


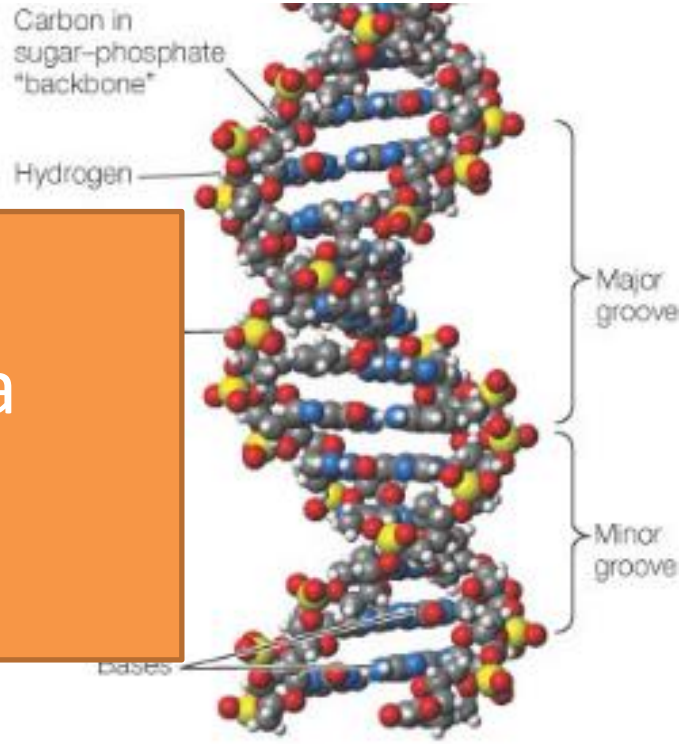
Rosalind Franklin



Franklin's X-ray diffraction photograph of DNA

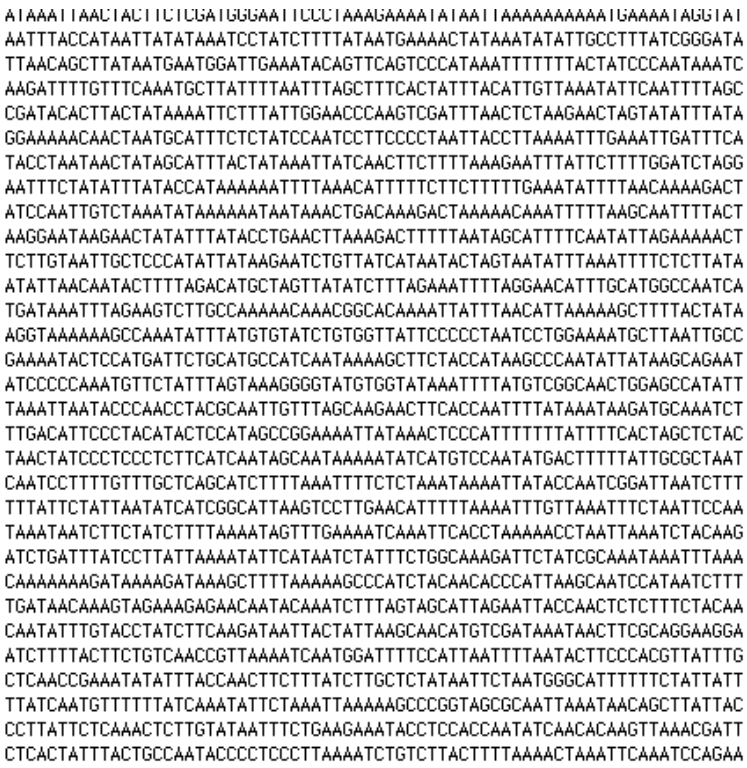
Nobel Prize in 1962

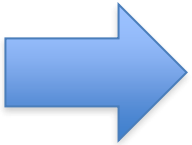




Геном - текст, написанный на алфавите из 4 букв

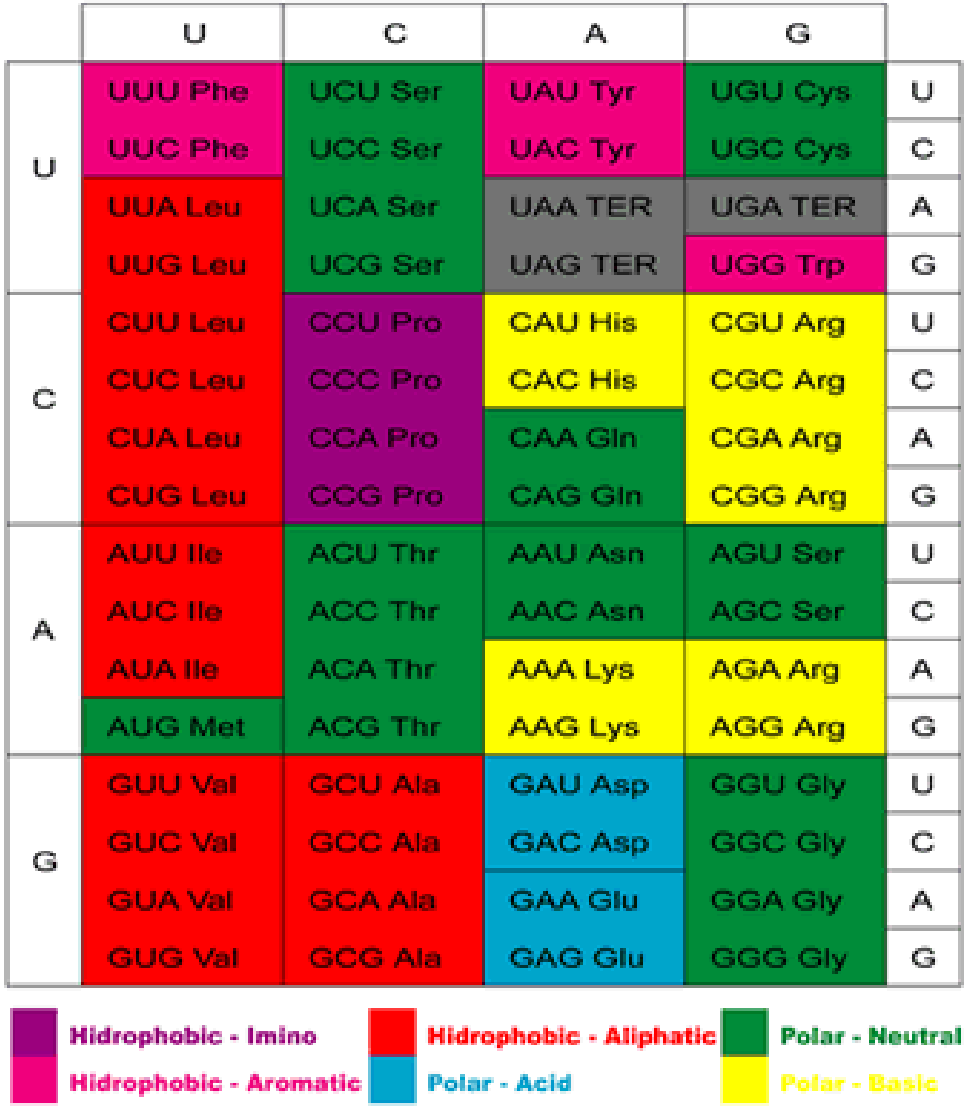
Расшифровать генетический код





1968 - Нобелевская премия по физиологии и медицине

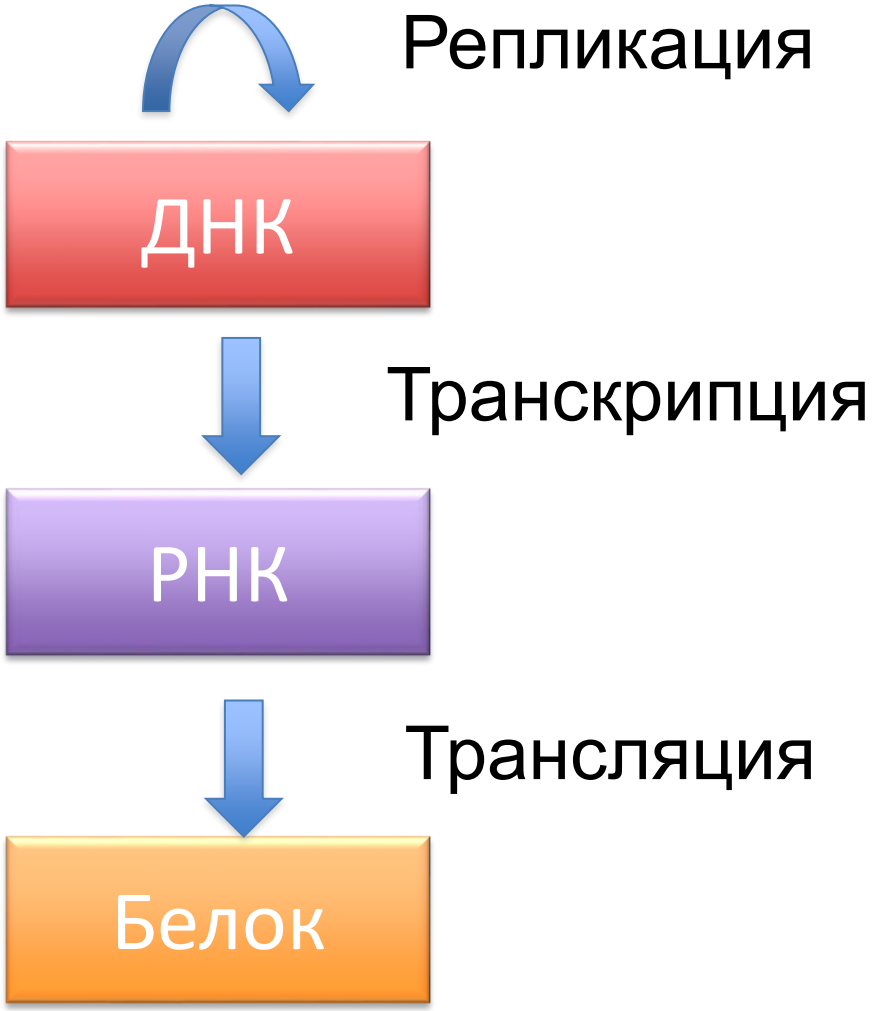
Ниренберг, Хорана, Холли



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

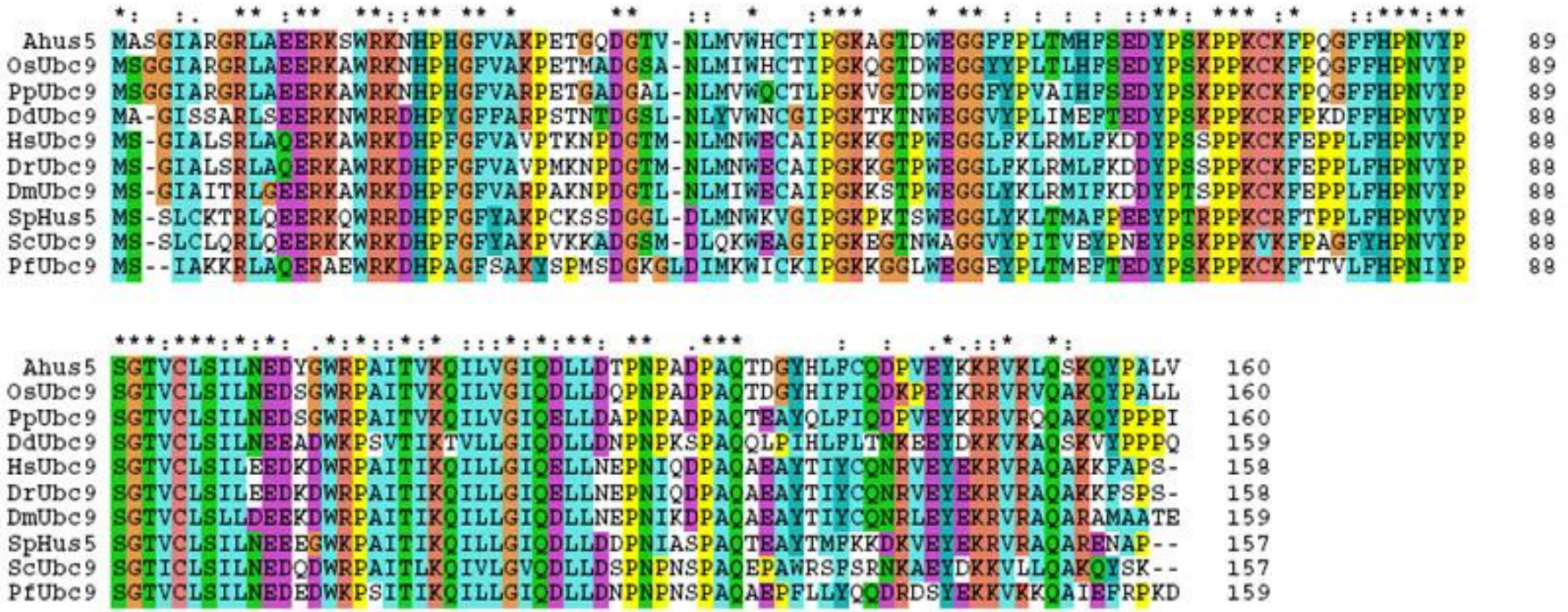
Hidrophobic - Imino
Hidrophobic - Aromatic
Hidrophobic - Aliphatic
Polar - Acid
Polar - Neutral
Polar - Basic

ИНФОРМАТИКА



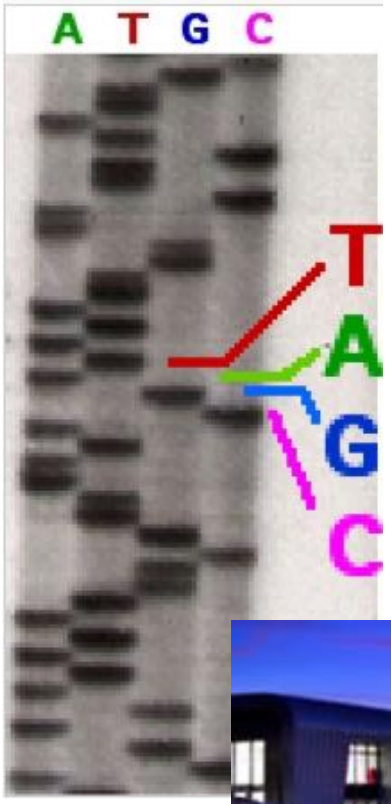
Рождение биоинформатики

Молекулярная эволюция – эволюция текстов



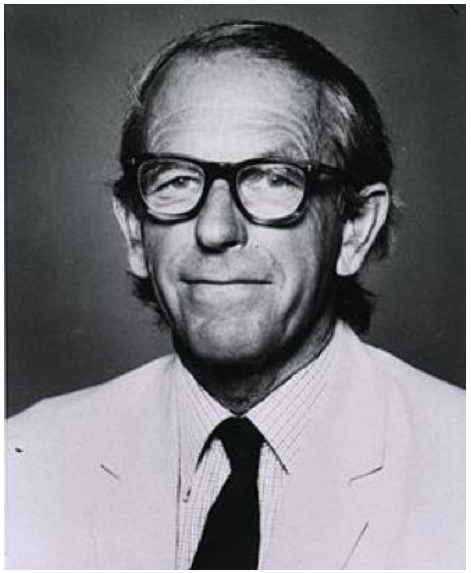
Метод секвенирования Сэнгера

1992 г Нобелевская премия по химии 1980 г.



- 1977 год

Начал с получения полной аминокислотной последовательности инсулина в 1951-52

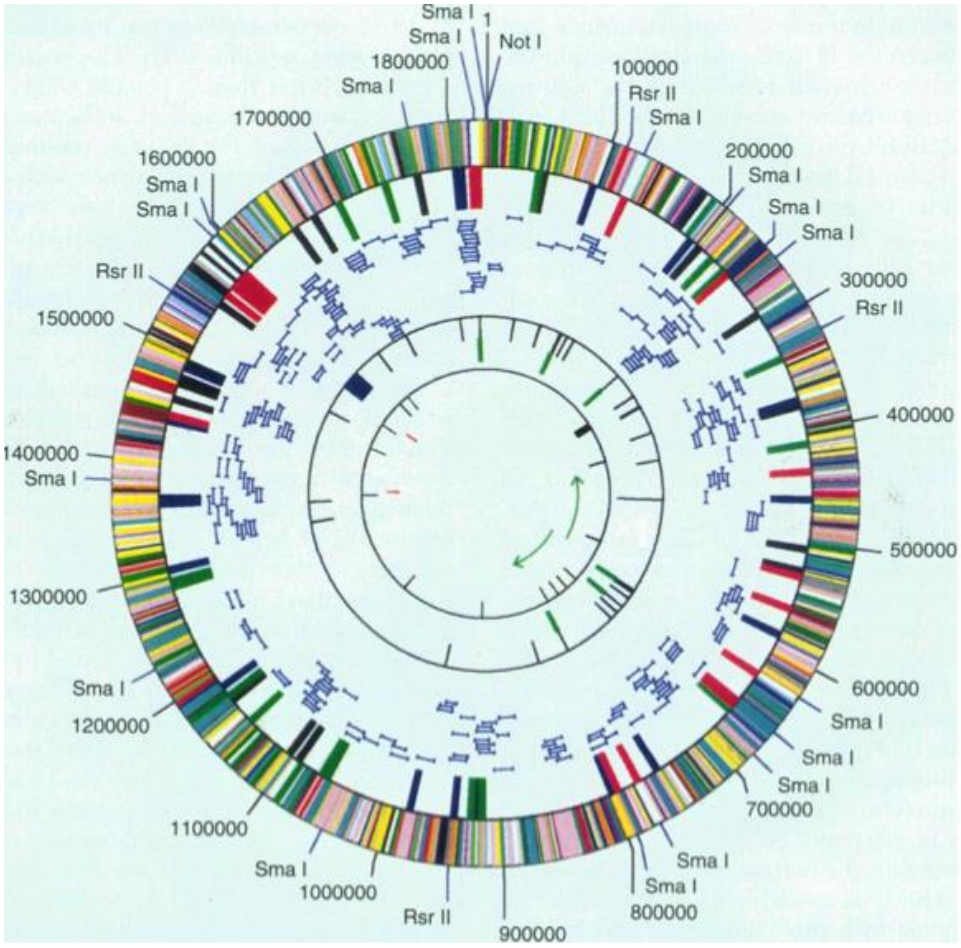


The Wellcome Trust Sanger Institute, 1992

Первый отсеквенированный геном

- 1995 г.

бактерия *Haemophilus influenzae –гемофильная палочка или палочка Пфейфера*

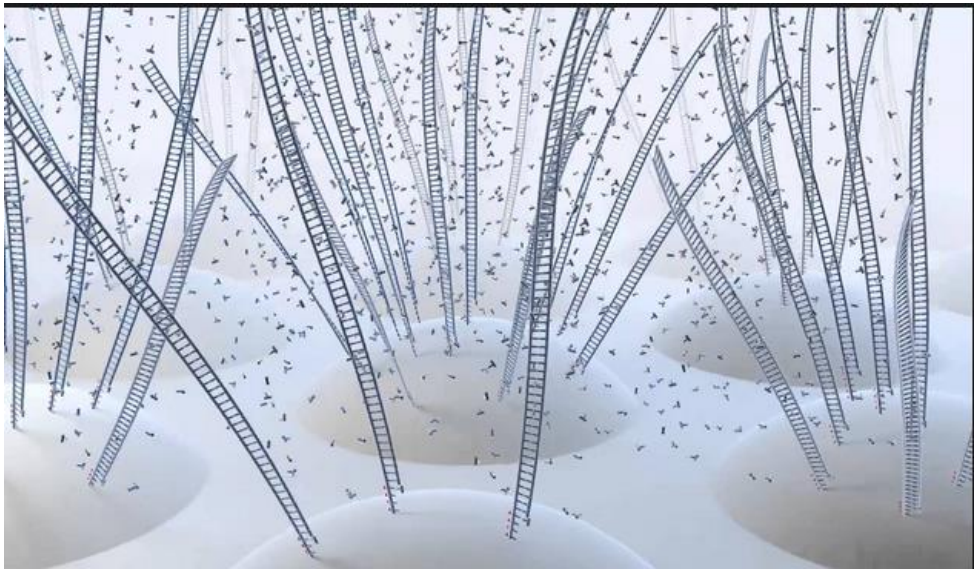


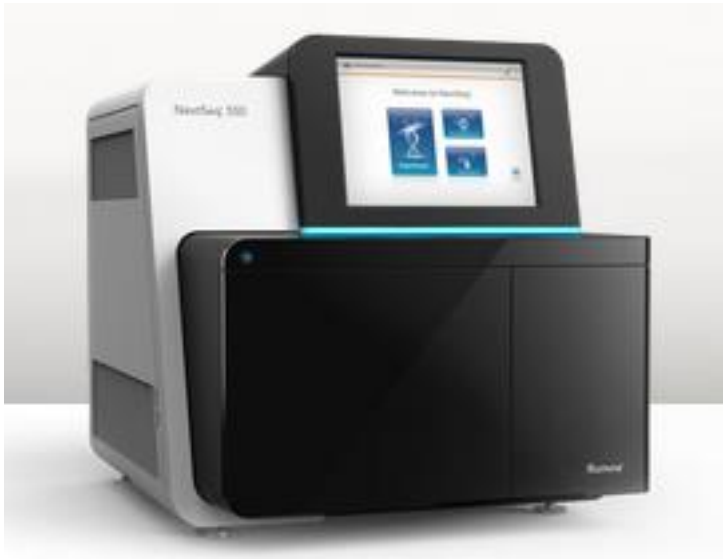
Проект генома человека

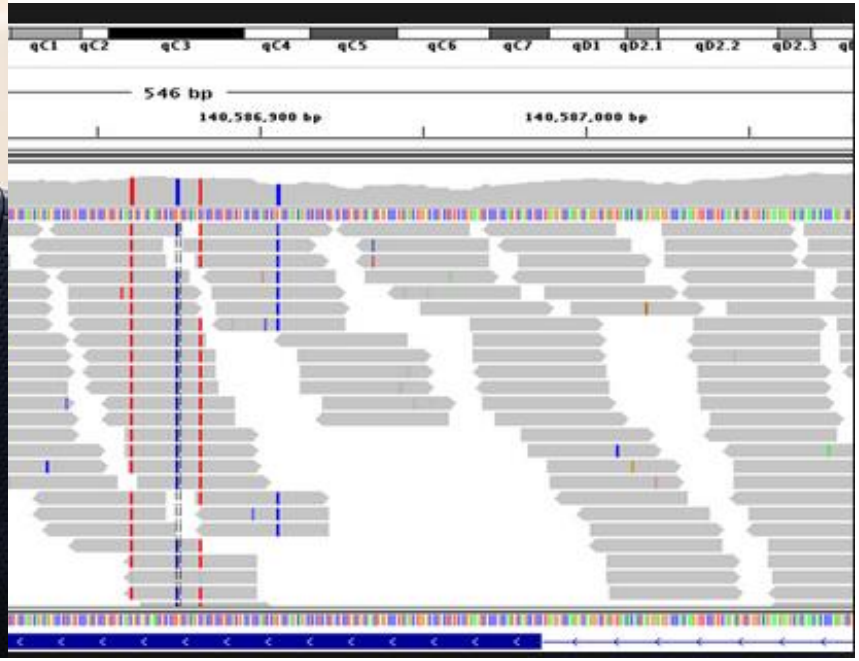
- Started in 1990
- Finished 2001/2003
- Sanger Sequencing


UNTIL FURTHER NOTICE
PLEASE PRE-RUN THIS
MACHINE FOR 30 MINUTES
BEFORE LOADING

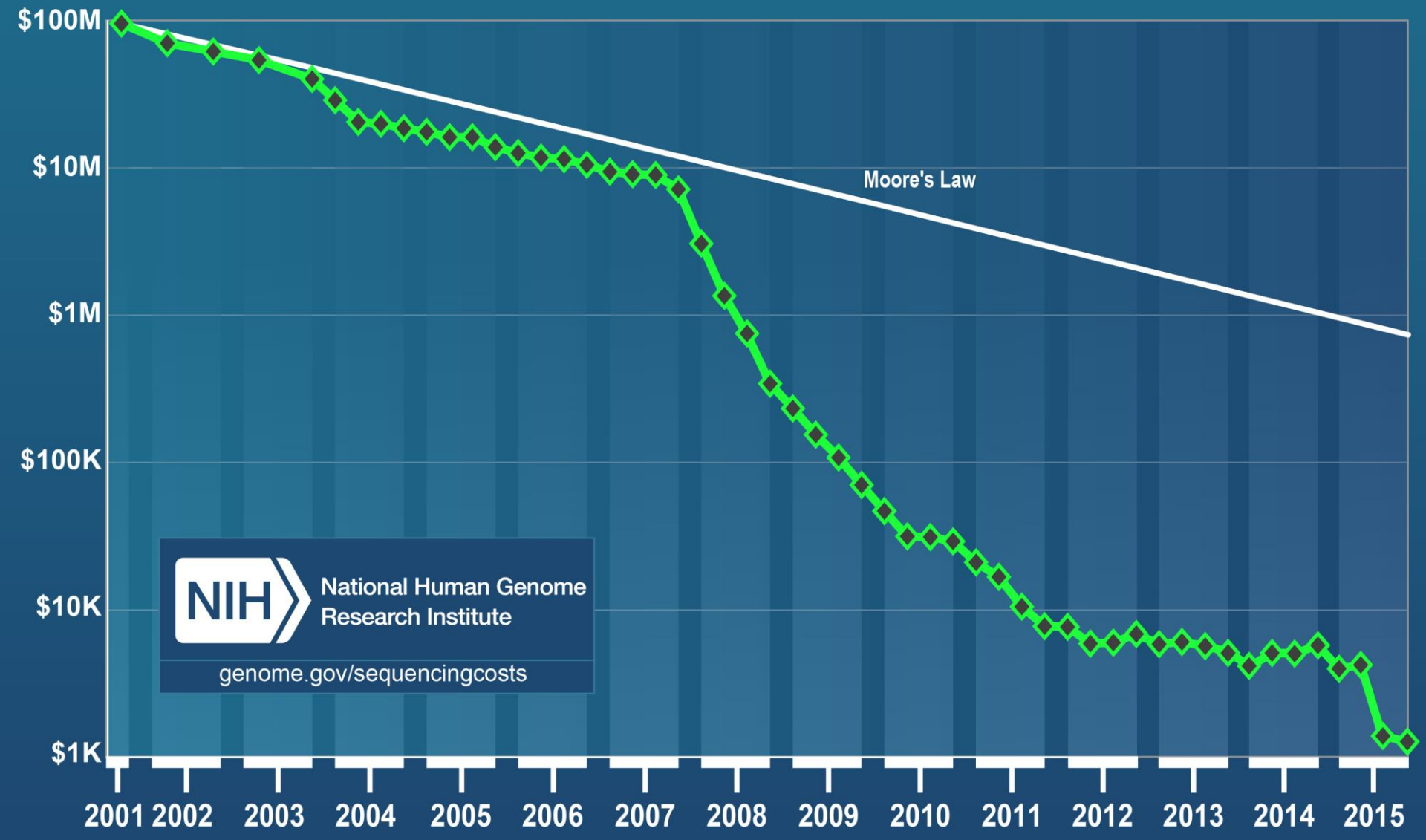
Next Generation Sequencing, or NextGenSeq, or NGS






qC1 qC2 qC3 qC4 qC5 qC6 qC7 qD1 qD2.1 qD2.2 qD2.3
546 bp
140,586,900 bp
140,587,000 bp

Cost per Genome



NextGen Sequencing Revolution

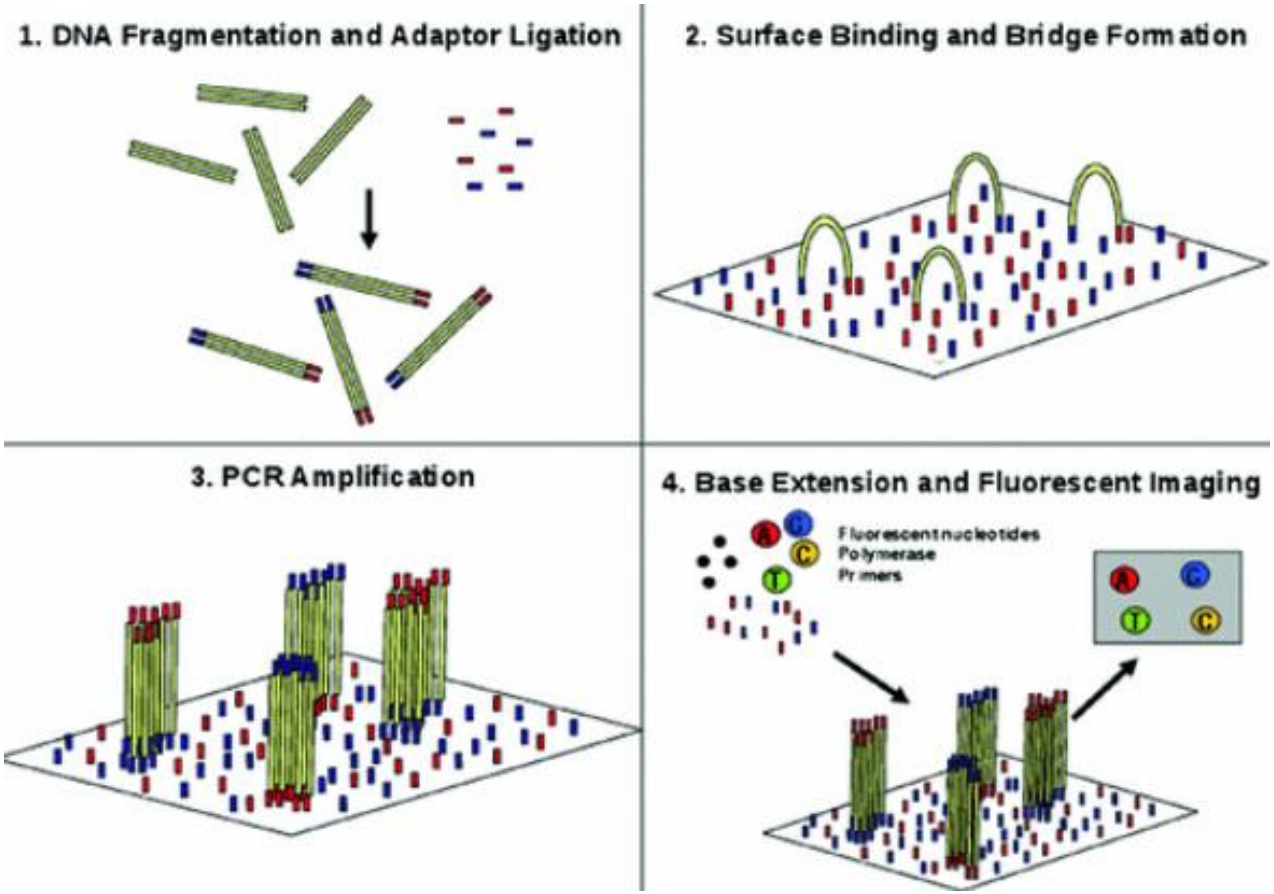
To print is more expensive than to sequence

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



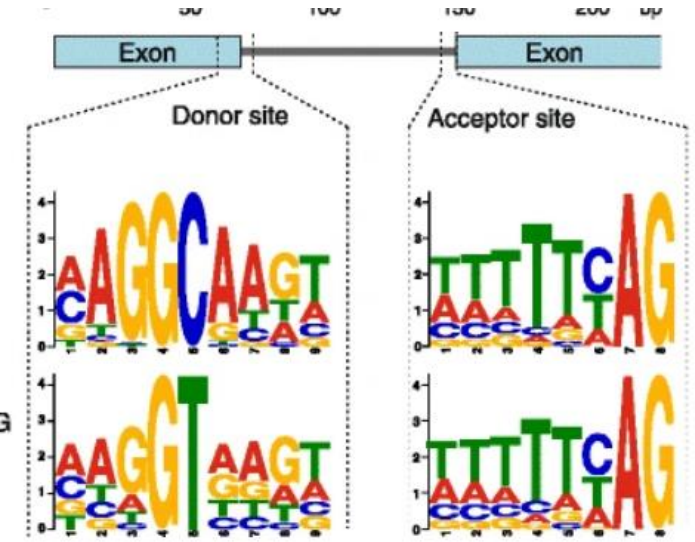
Massive parallel sequencing

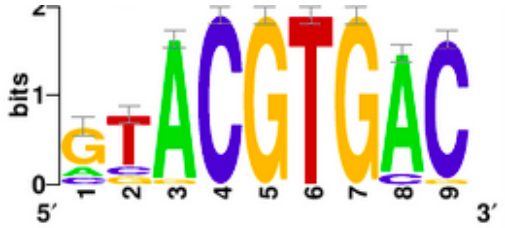
1 million to 43 billion short reads per instrument run

Слои геномной кодировки

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

DNA-seq, RNA-seq, ChIP-seq

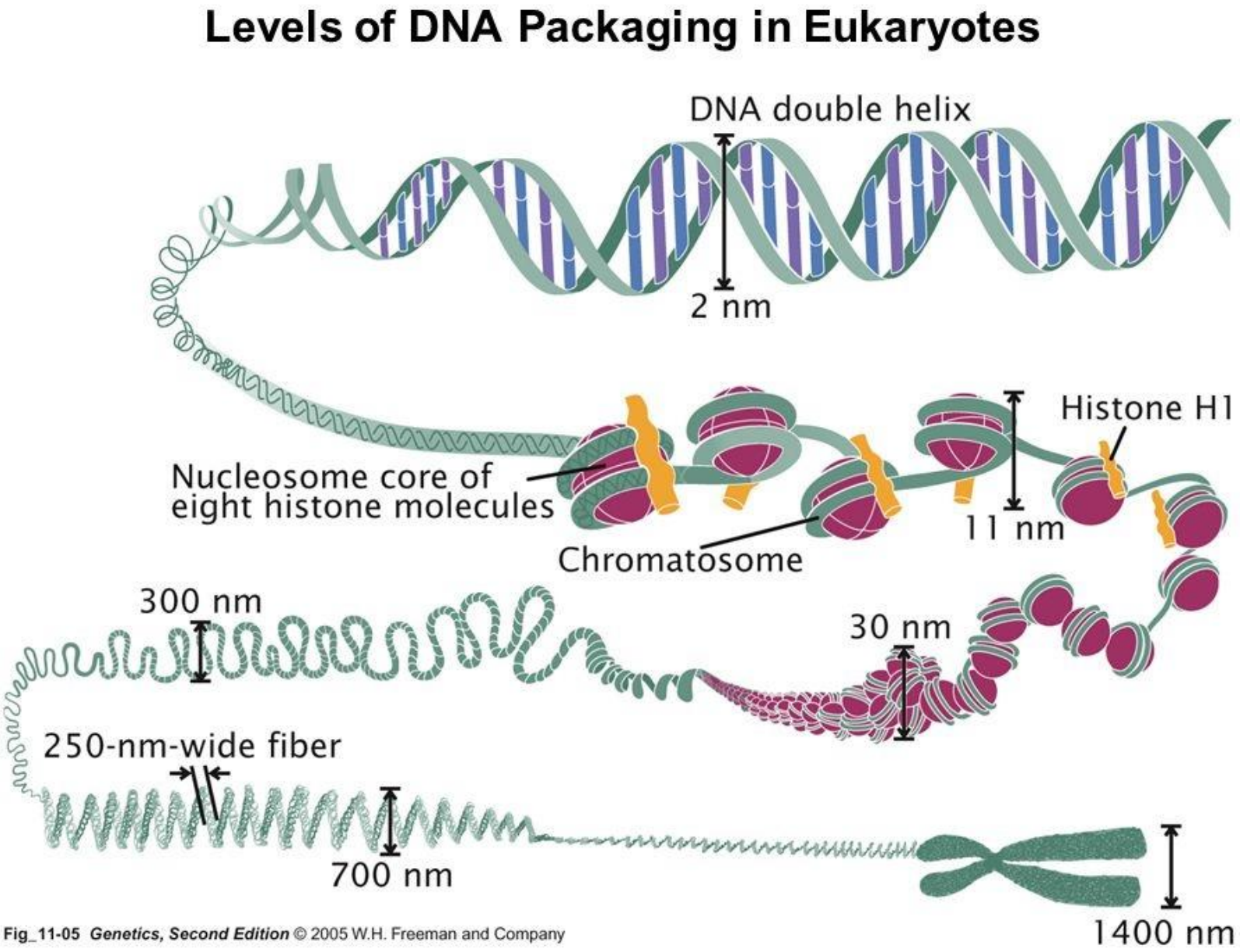




	U	C	A	G	
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA TER	UGA TER	A
	UUG Leu	UCG Ser	UAG TER	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile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

Hidrophobic - Imino
Hidrophobic - Aromatic
Hidrophobic - Aliphatic
Polar - Acid
Polar - Neutral
Polar - Basic

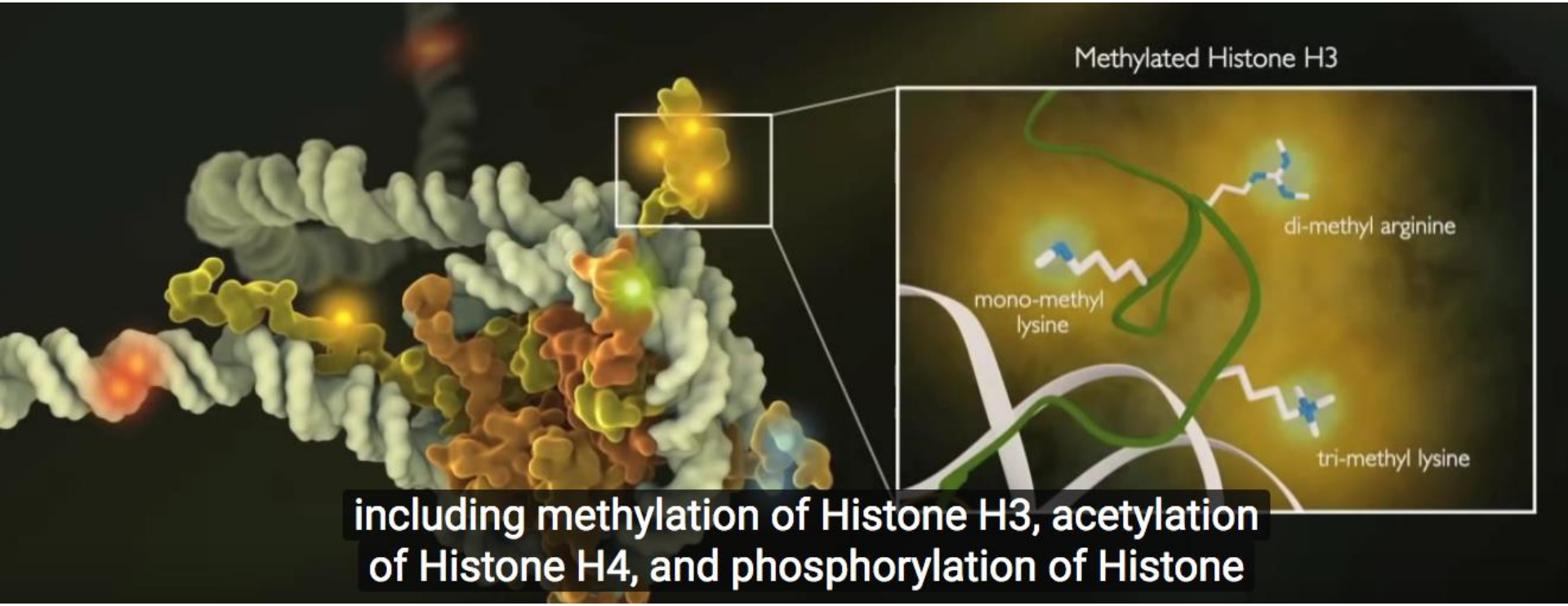
Эпигенетика



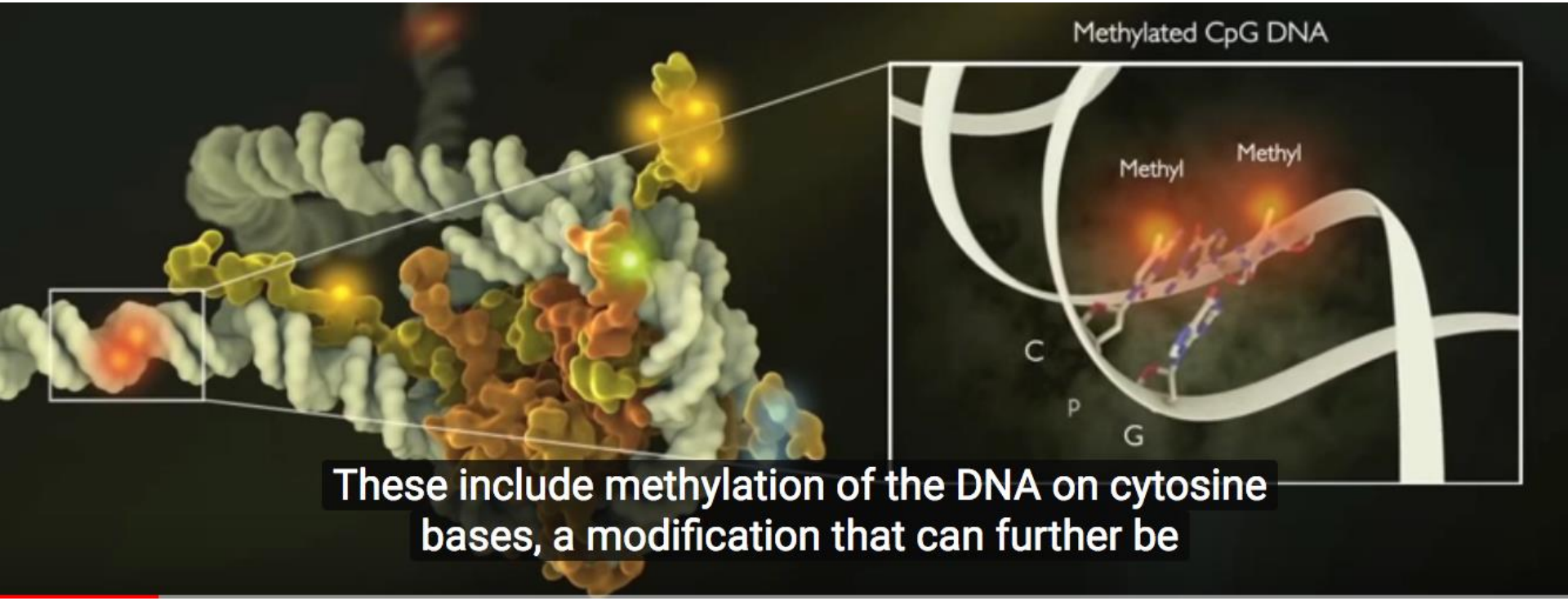
Слои геномной кодировки

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

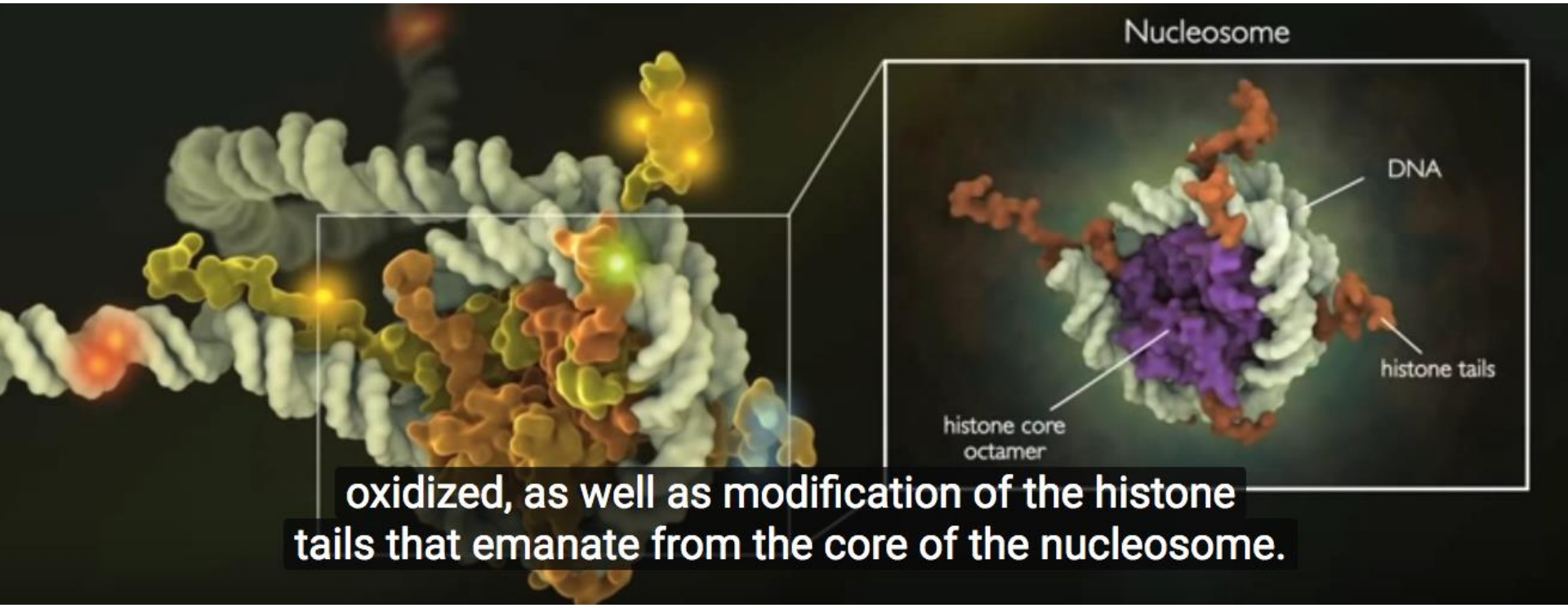
- DNA marks: Epigenomics
  - Methylations, histone modifications, chromatine packaging



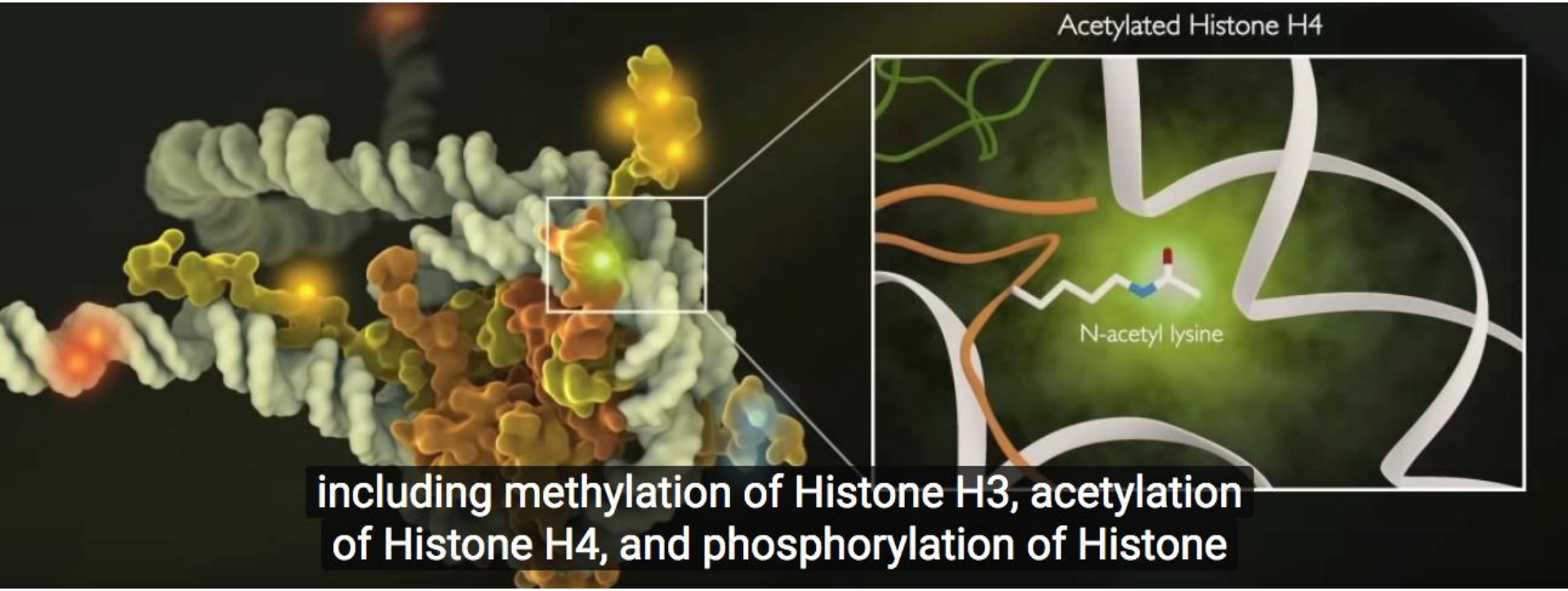
Второй слой геномной аннотации



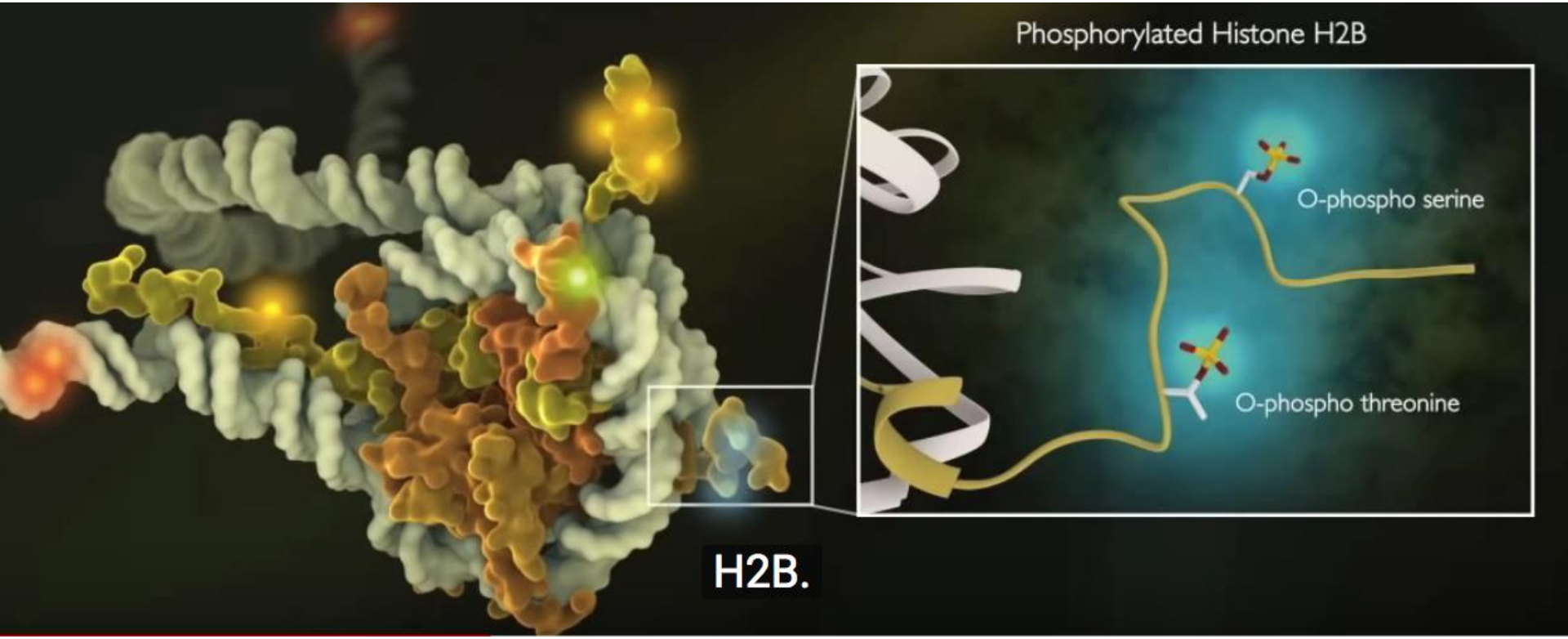
Второй слой геномной аннотации



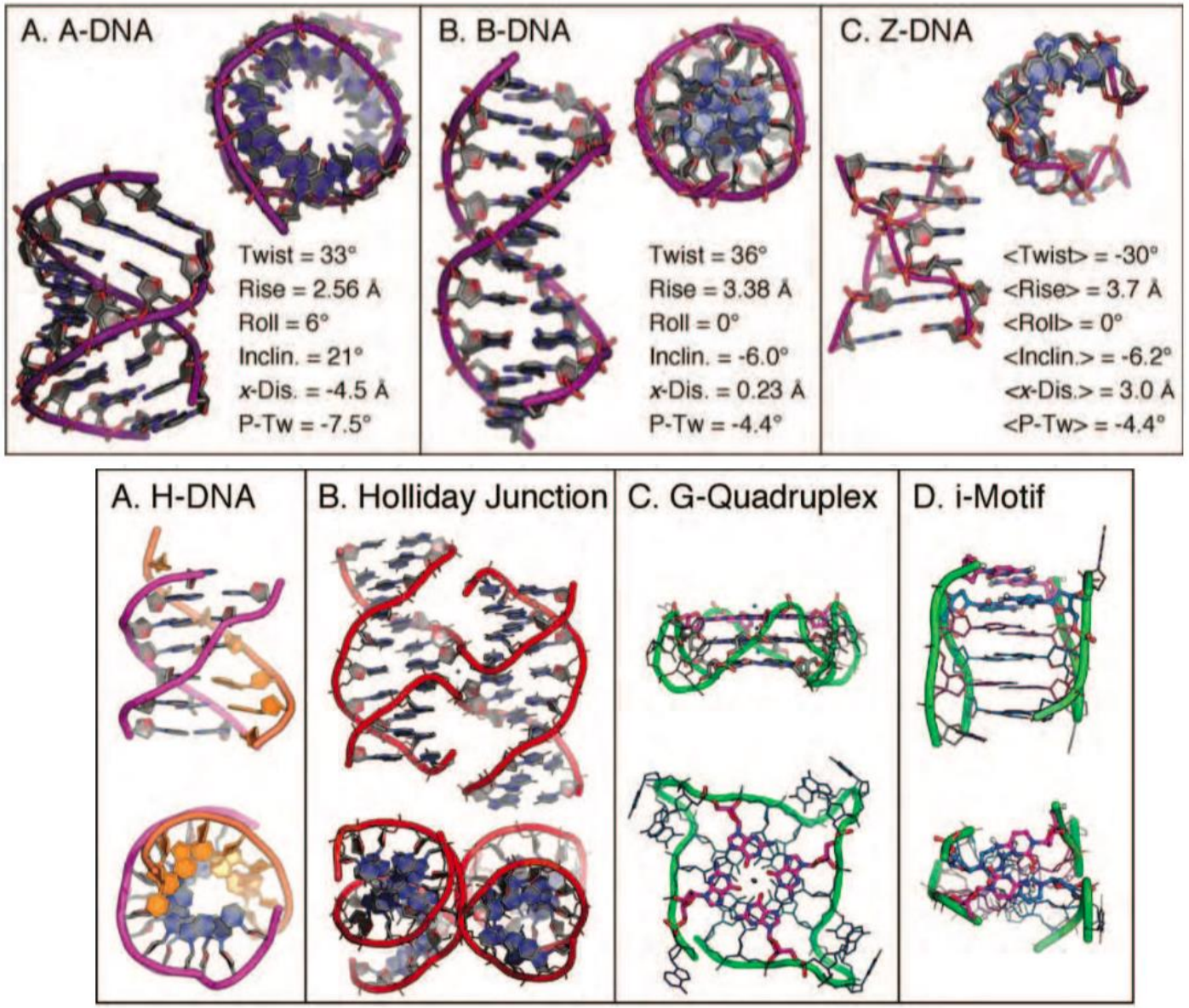
Второй слой геномной аннотации



Второй слой геномной аннотации



Третий слой – структуры ДНК

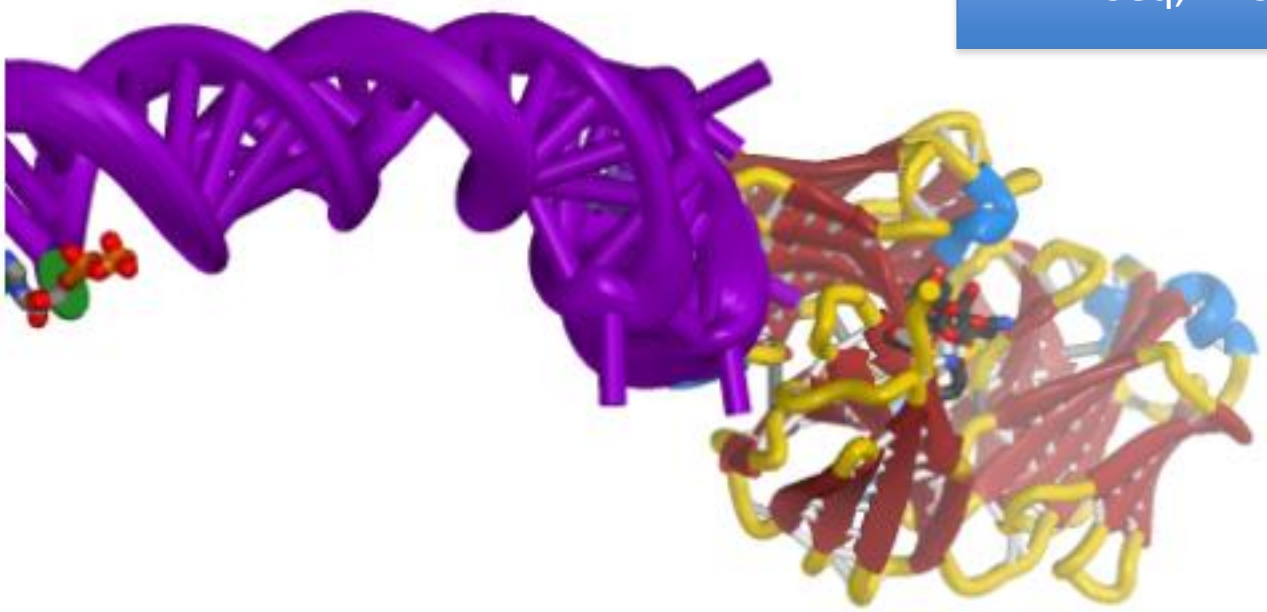


Book chapter “DNA Structure: Alphabet Soup for the Cellular Soul ” by P. Shing Ho and Megan Carter

Третий слой геномной кодировки – структуры ДНК

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

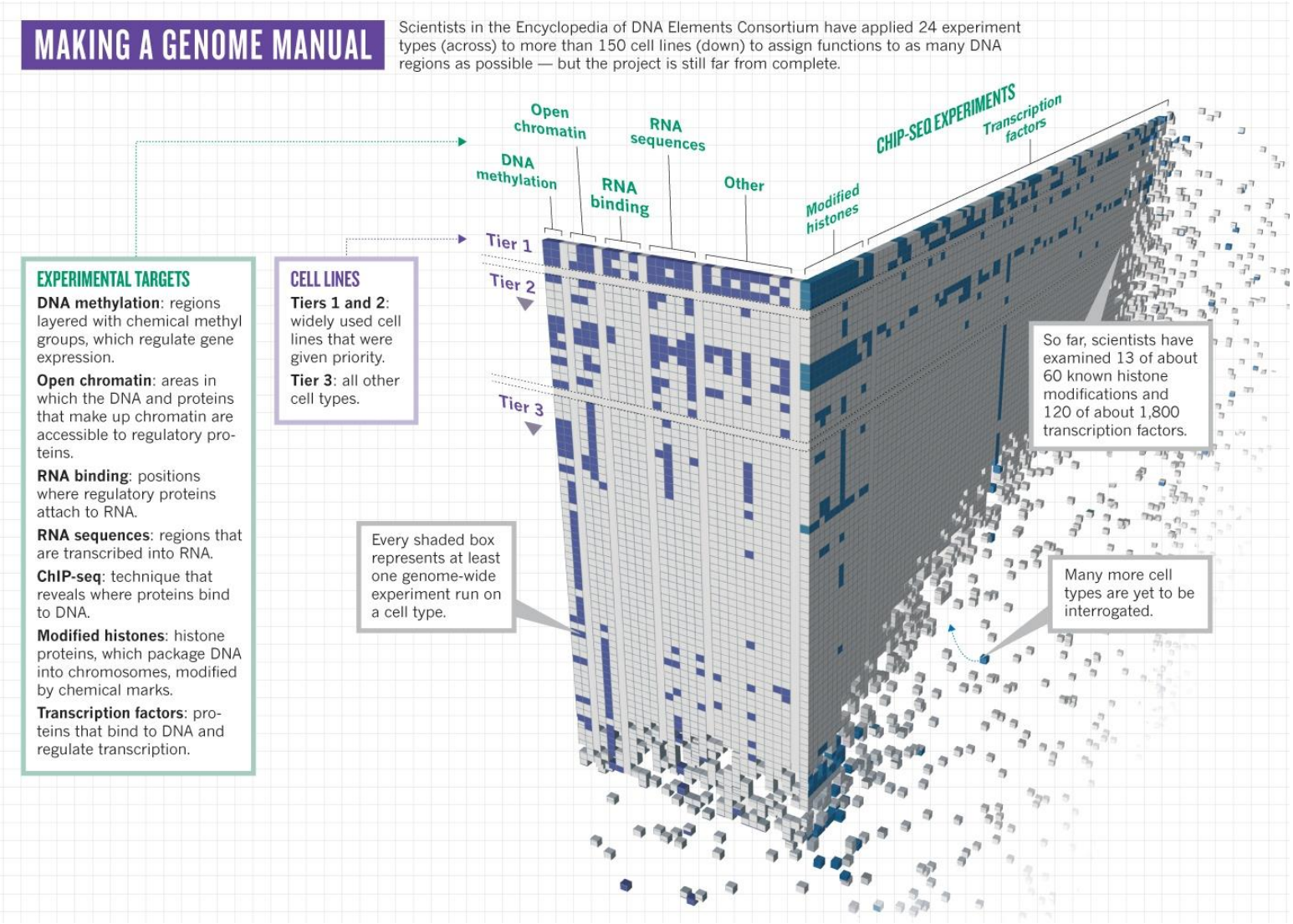
ChIP-seq, G4-seq, ssDNA-Seq, KAS-seq



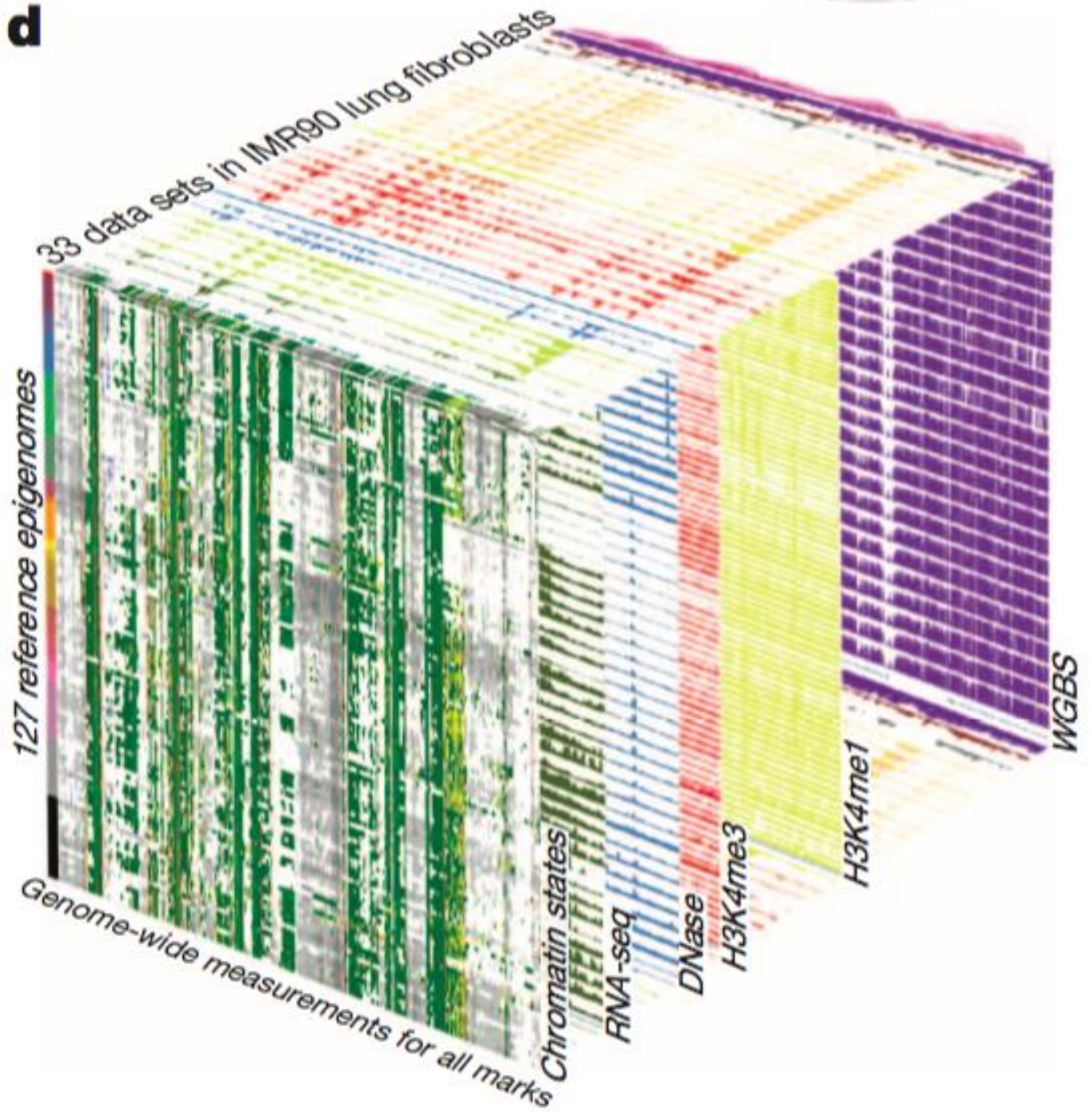
https://3dprint.nih.gov/discover/3DPX-010444

Накопление данных

ENCODE: Encyclopedia of DNA Elements

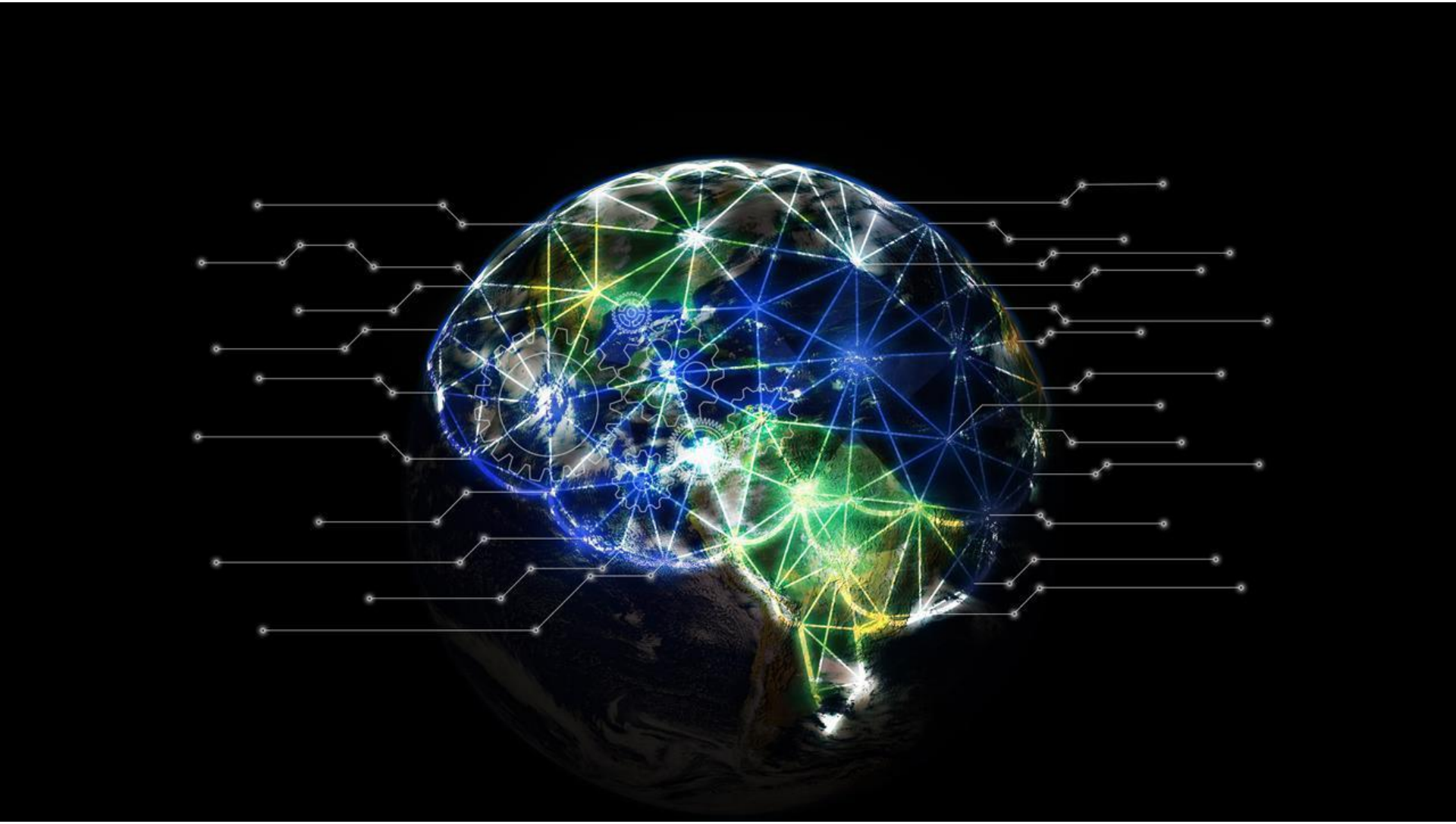


Relationship of figure panels highlights data set dimensions

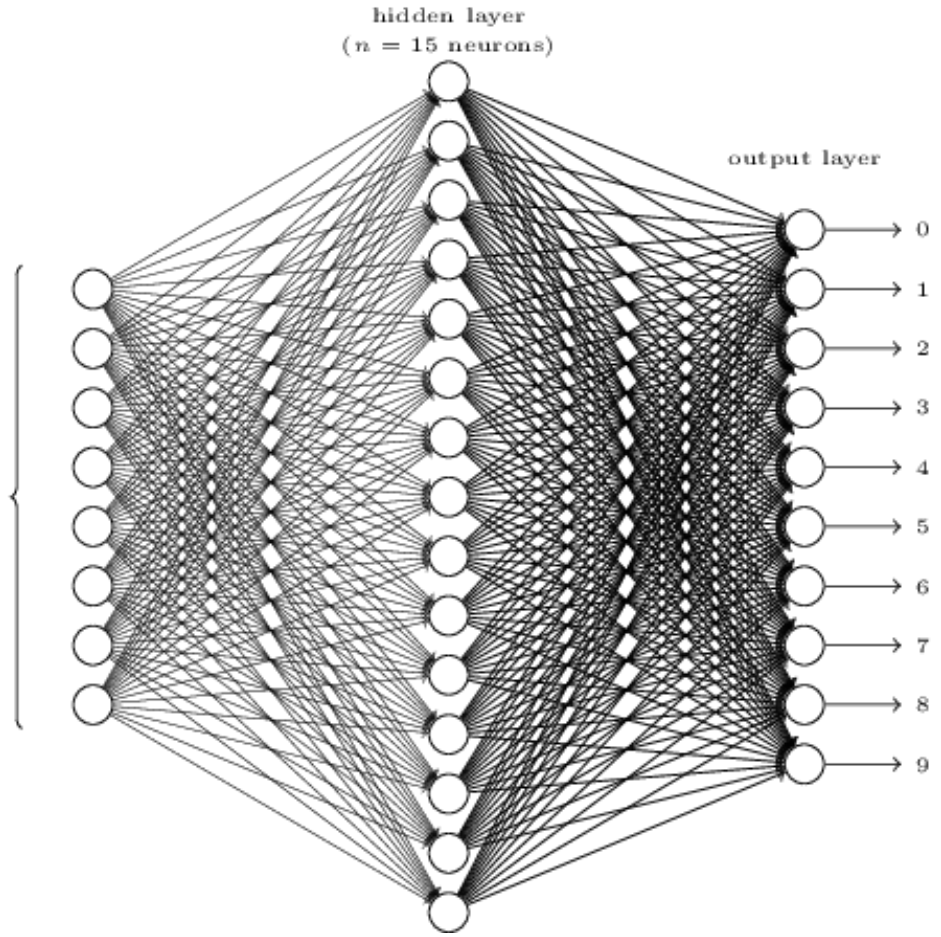


Что делать?

Машинное обучение

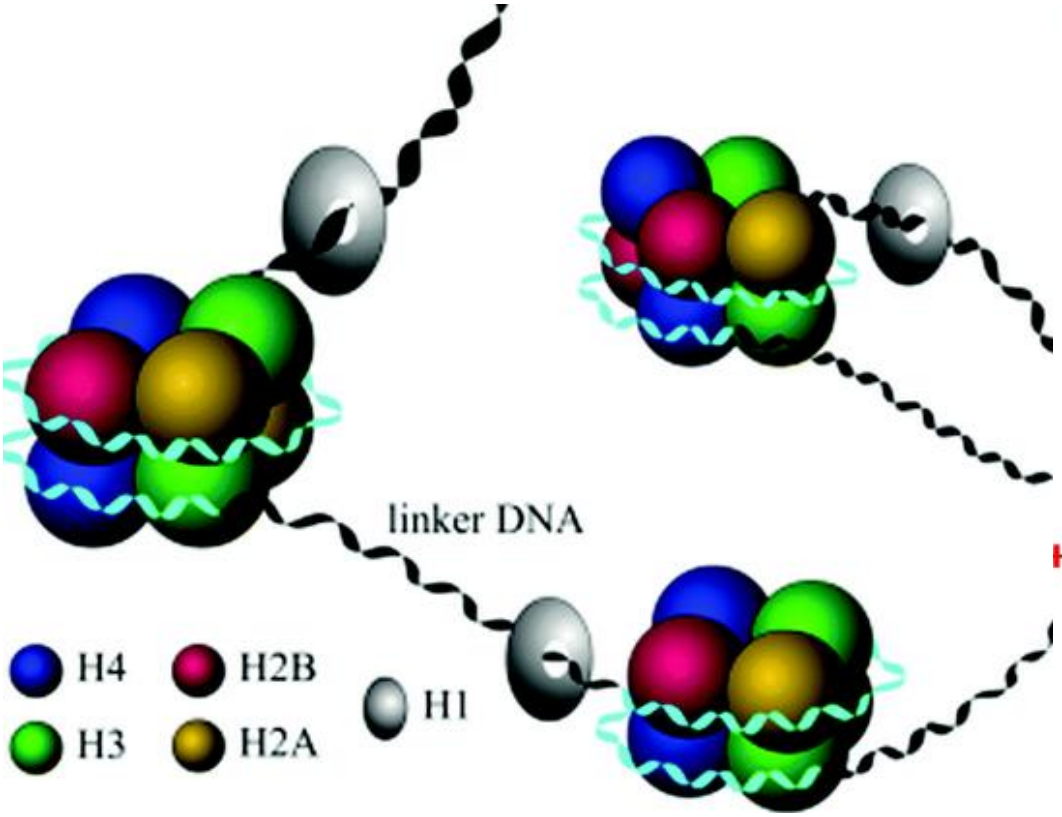


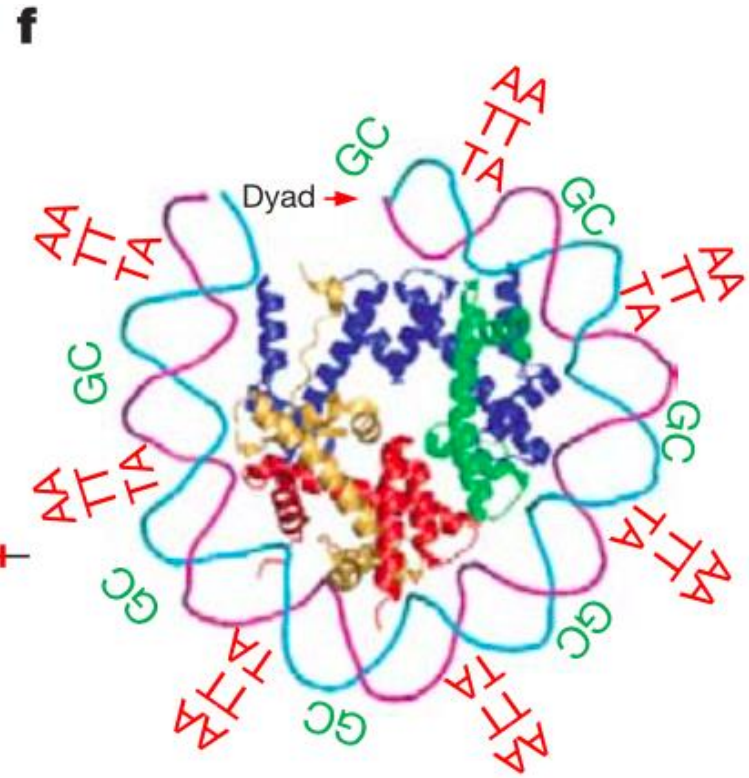
Нейронные сети



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]

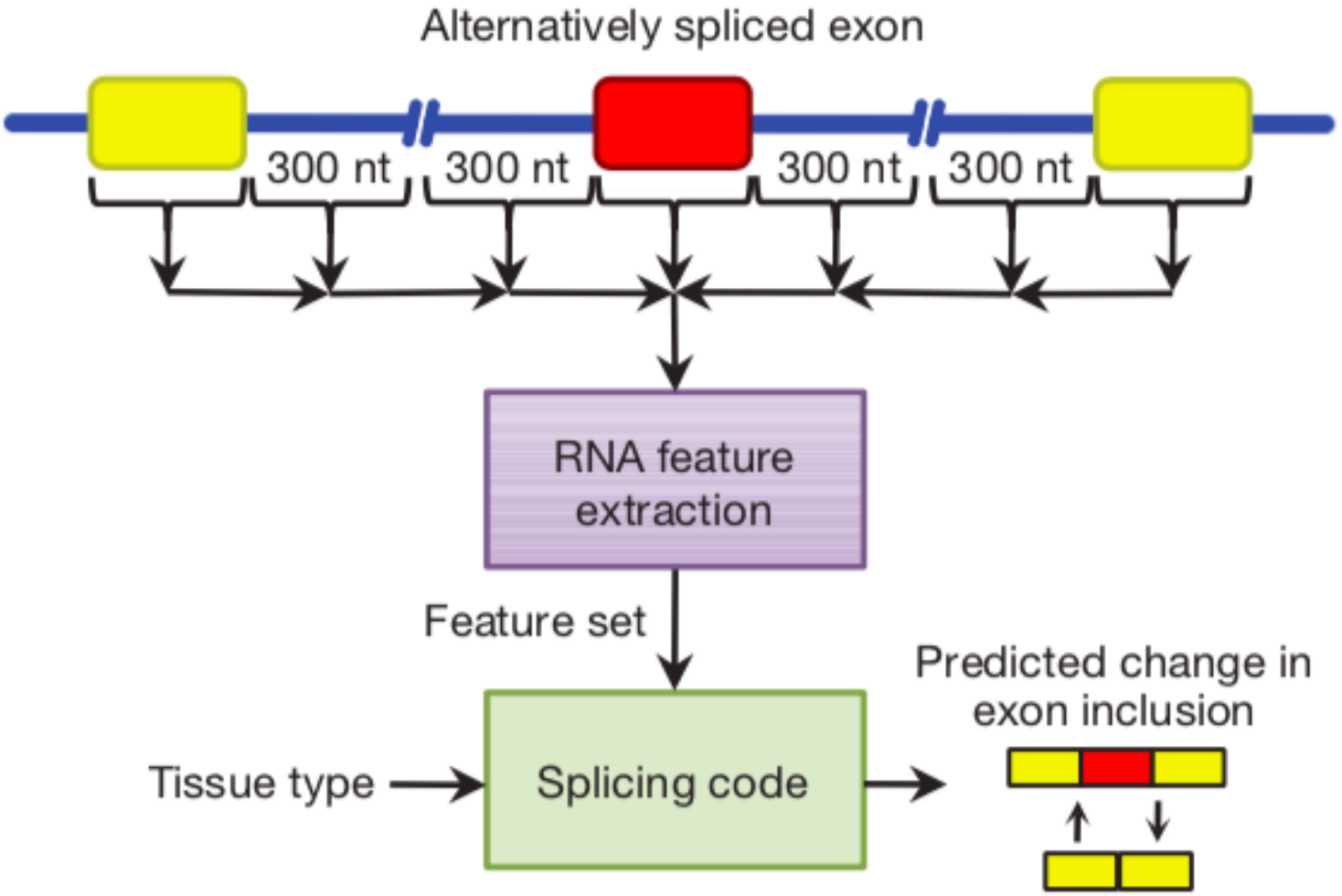




Nature 2006

Deciphering the splicing code

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]

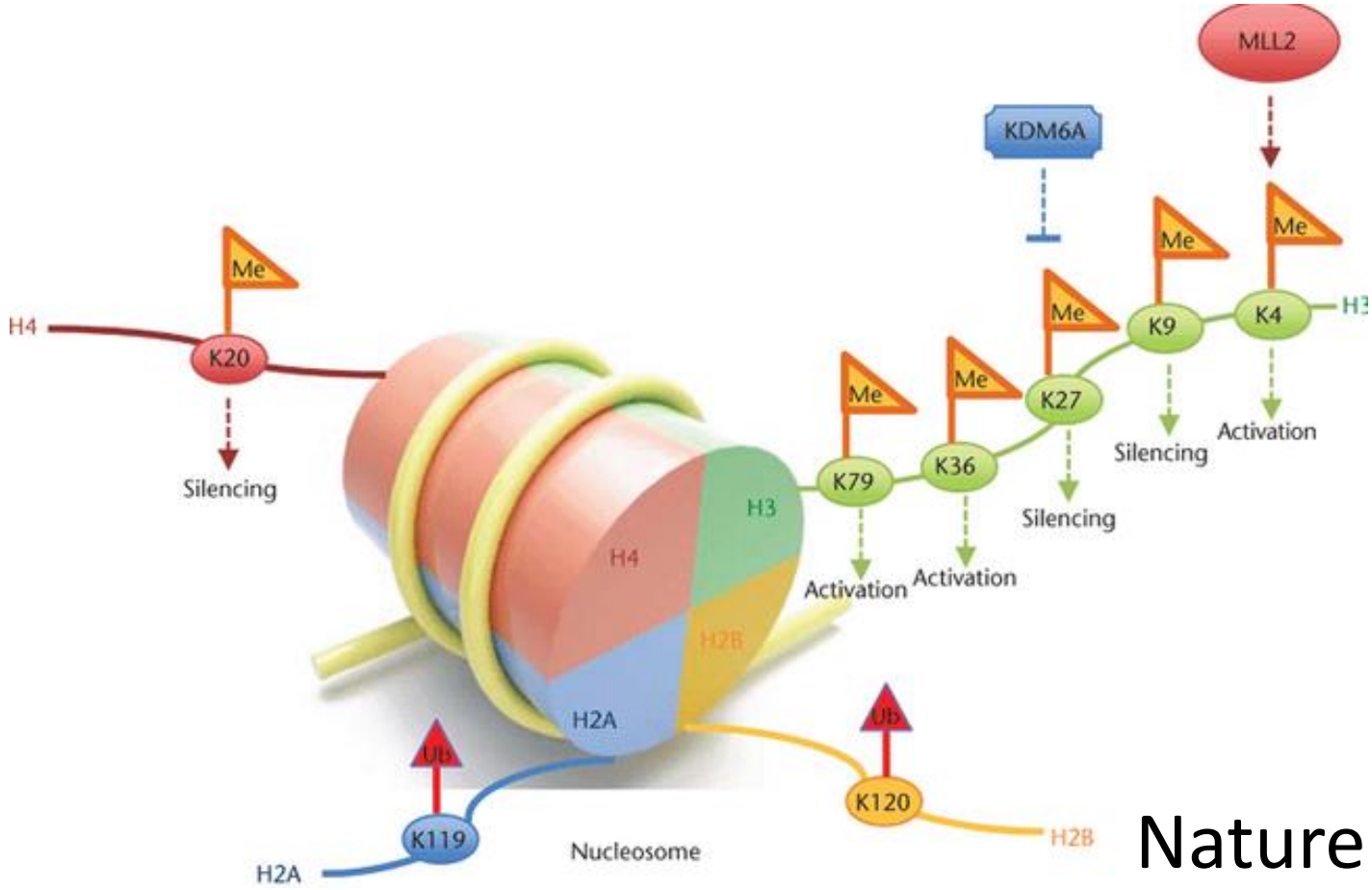


Nature 2010

Epigenetic code

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

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

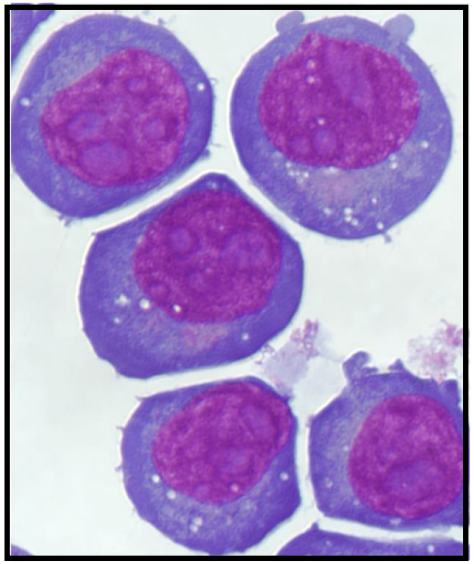


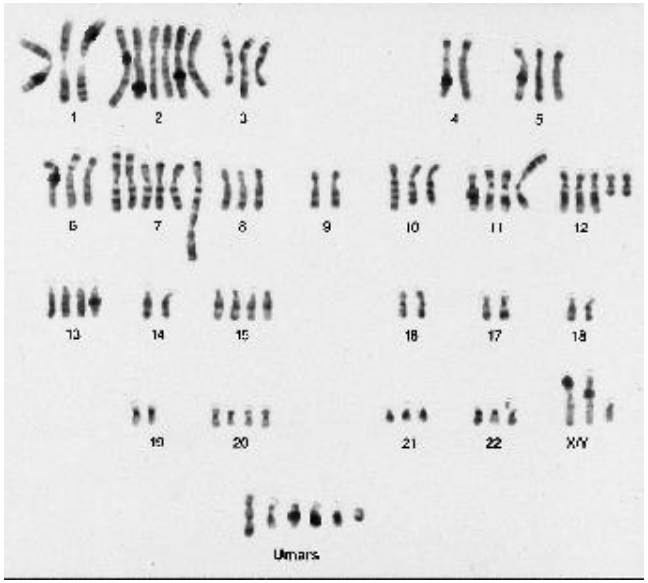
Nature 2010

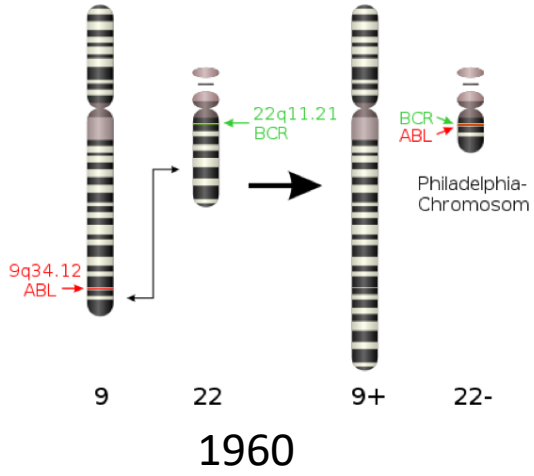
РАК

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 Карениной:

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

- **Challenge in Treating Cancer**

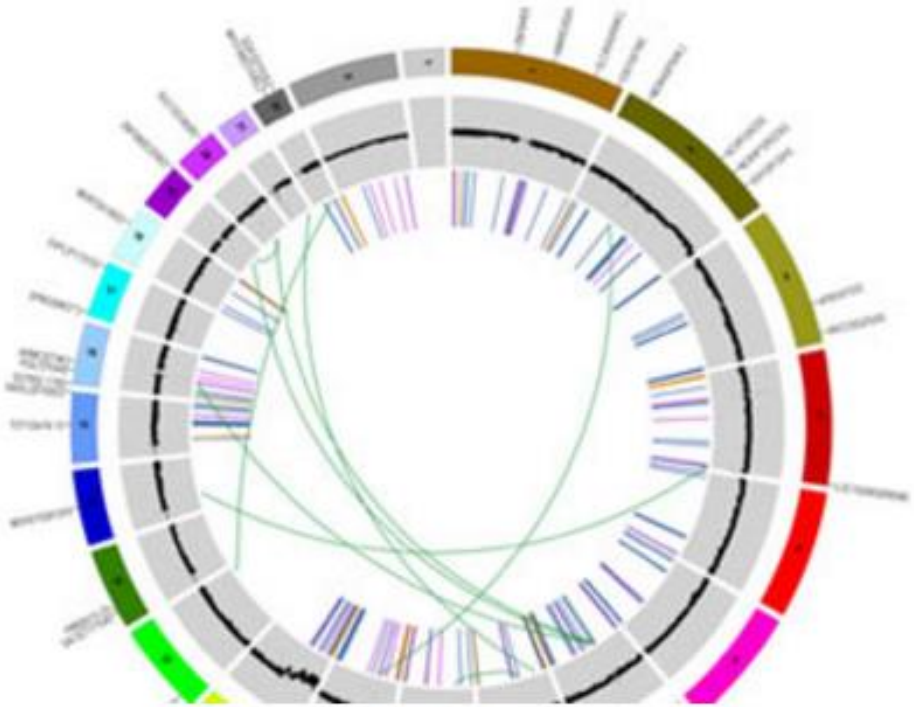
Every tumor is different
Every cancer patient is different

THE CANCER GENOME ATLAS

National Cancer Institute
National Human Genome Research Institute

Search

Home | About Cancer Genomics | Cancers Selected for Study | Research Highlights | Publicatio

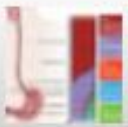


Program Overview

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

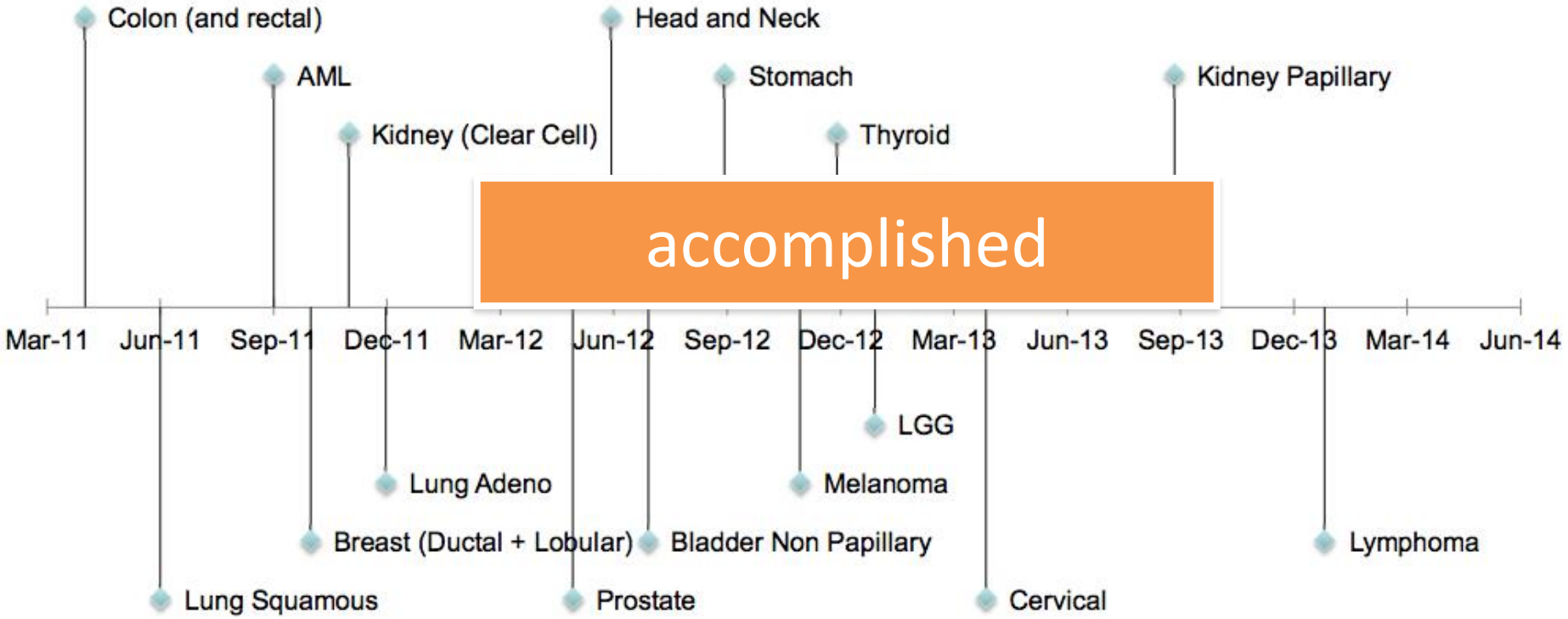
Cancers Selected for Study

About TCGA

Active Tumor Projects

THE CANCER GENOME ATLAS

Timeline to Completion of Comprehensive Analysis for Each Tumor Project



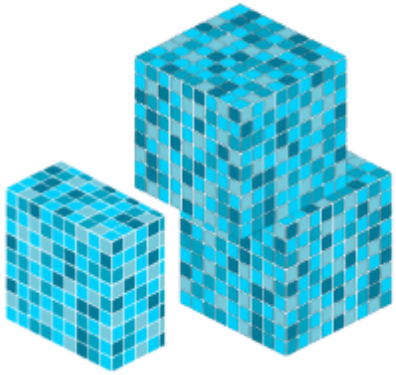
TCGA BY THE NUMBERS

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

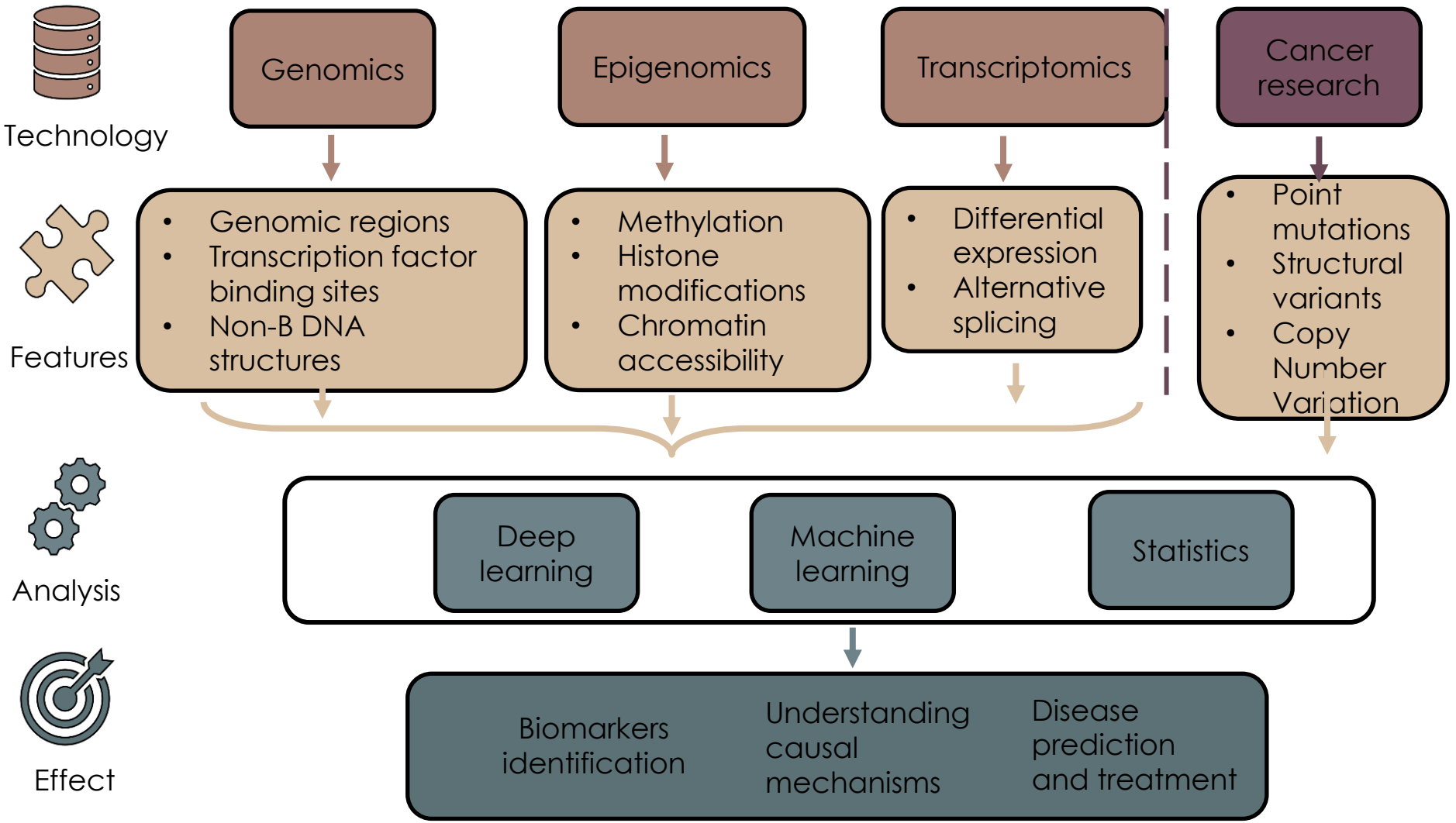
PATIENTS

...using

7

DIFFERENT DATA TYPES

Omics approach



RESEARCH ARTICLE Open Access

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

Kseniia 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.

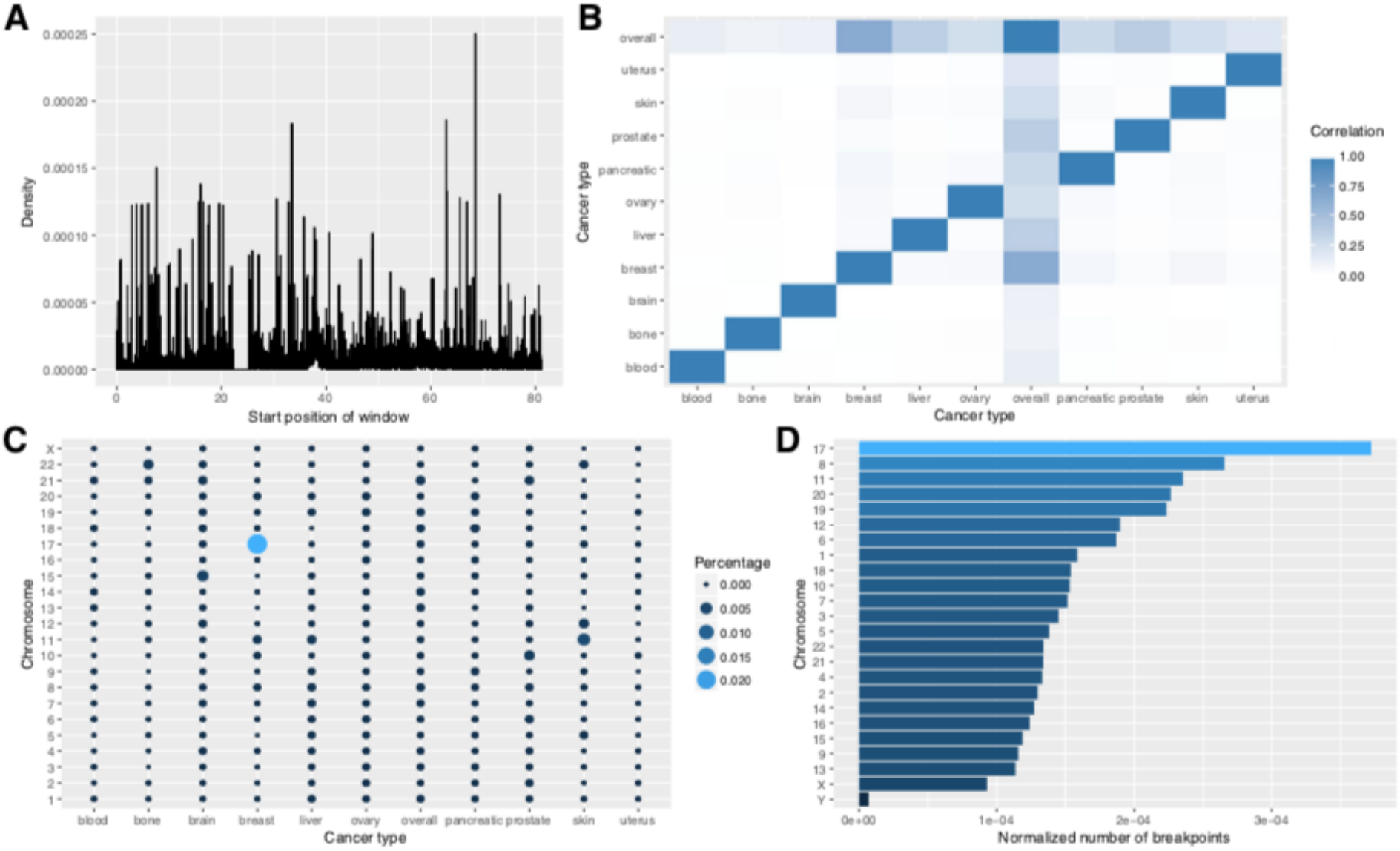


Fig. 1 a Number of breakpoint hotspots by chromosome and type of cancer for 0.1% labeling type. b Correlation between different cancer profiles. c Breakpoint density in general cancer profile across chromosome 17 (in Mb). d Number of breakpoints normalized by chromosome length

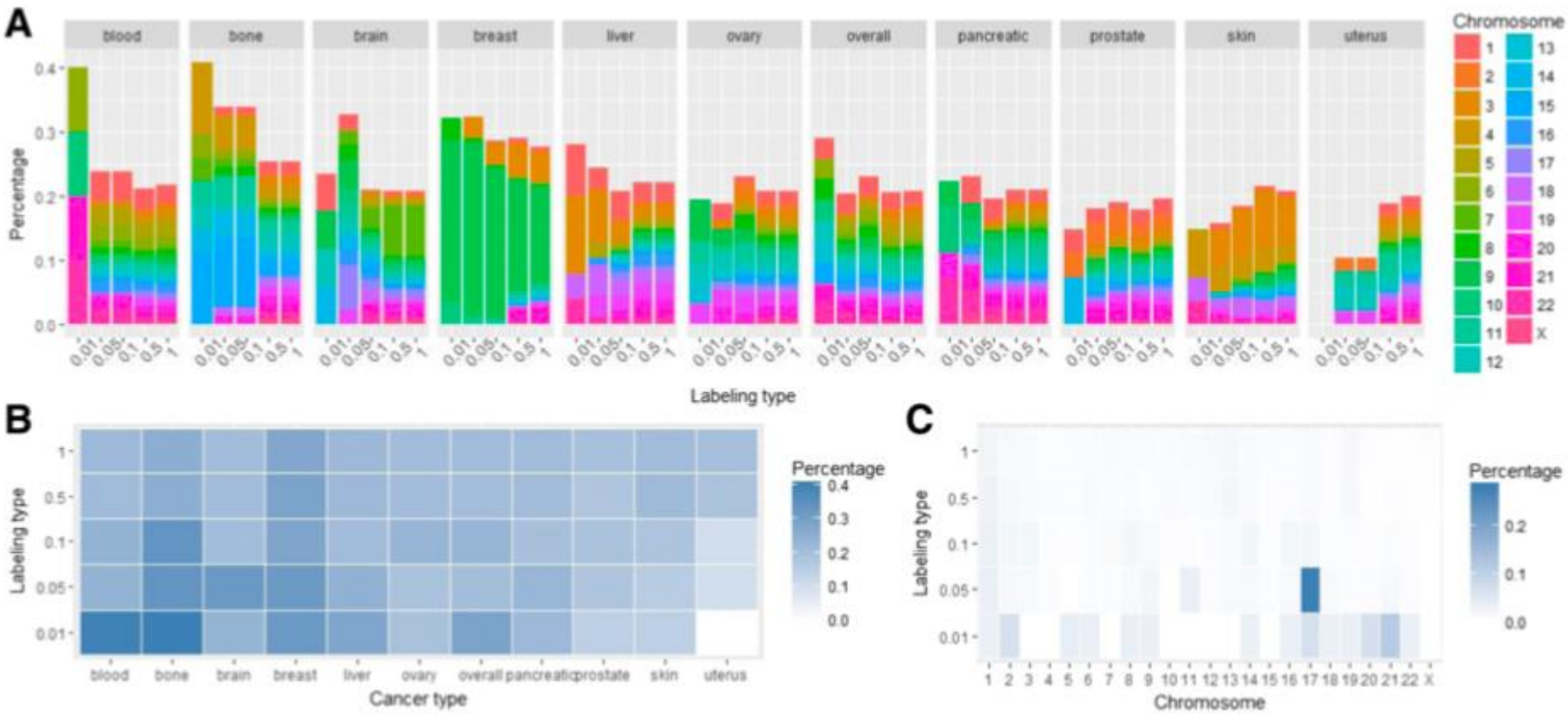
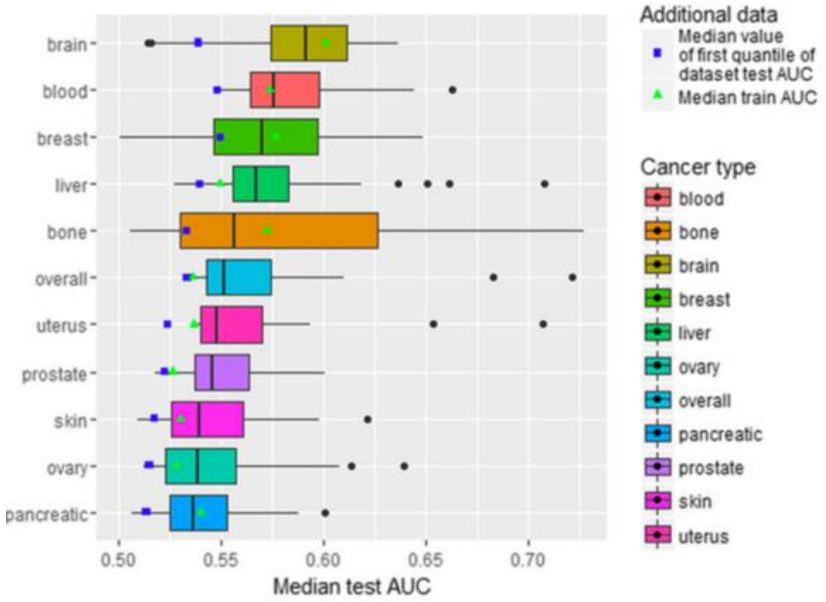
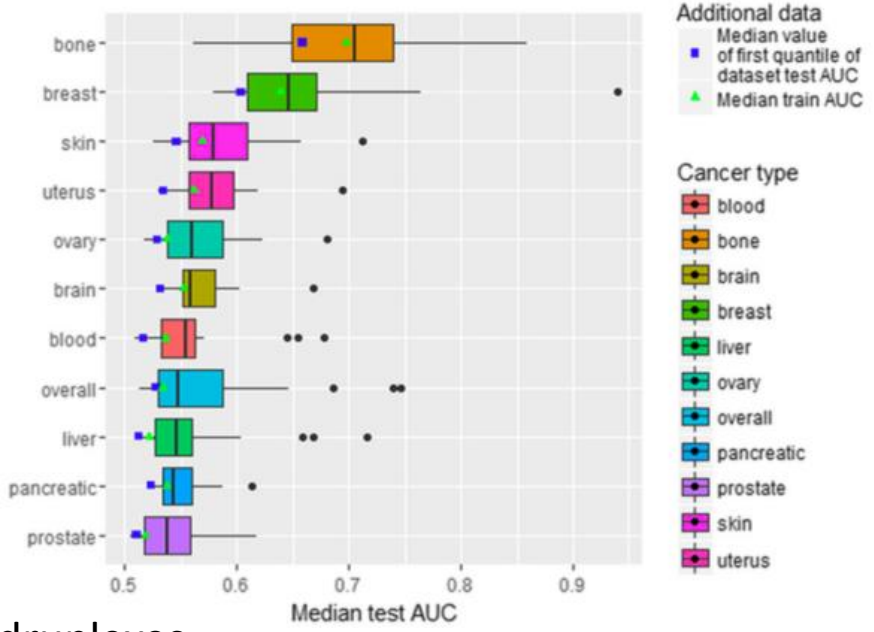


Fig. 2 a Percentage of breakpoint hotspots intersecting with whole gene regions in different cancer types for 5 different hotspot labeling types. b Heatmap of the percentage of the breakpoint hotspots' intersections with whole gene regions across different cancer types. c Heatmap of the percentages of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

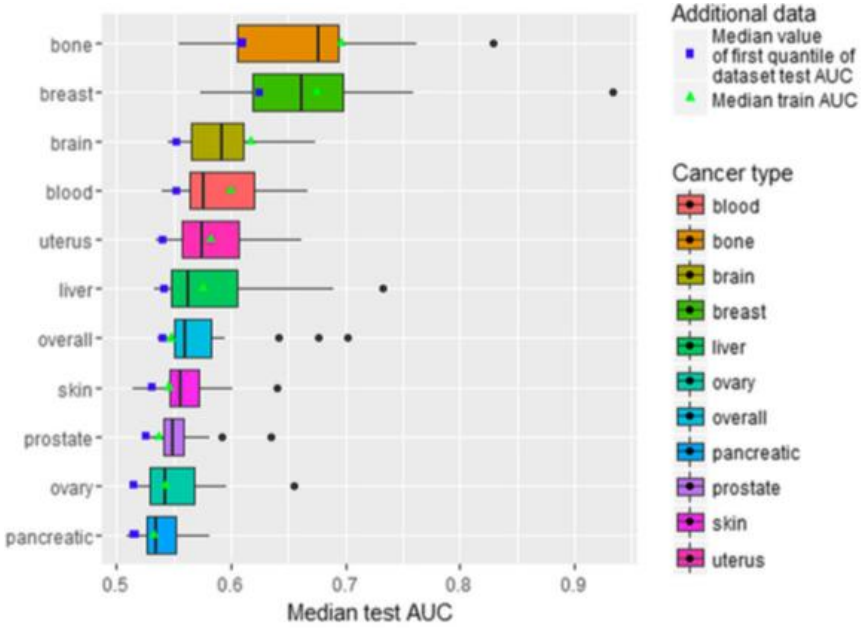
Stem-loops



Quadruplexes



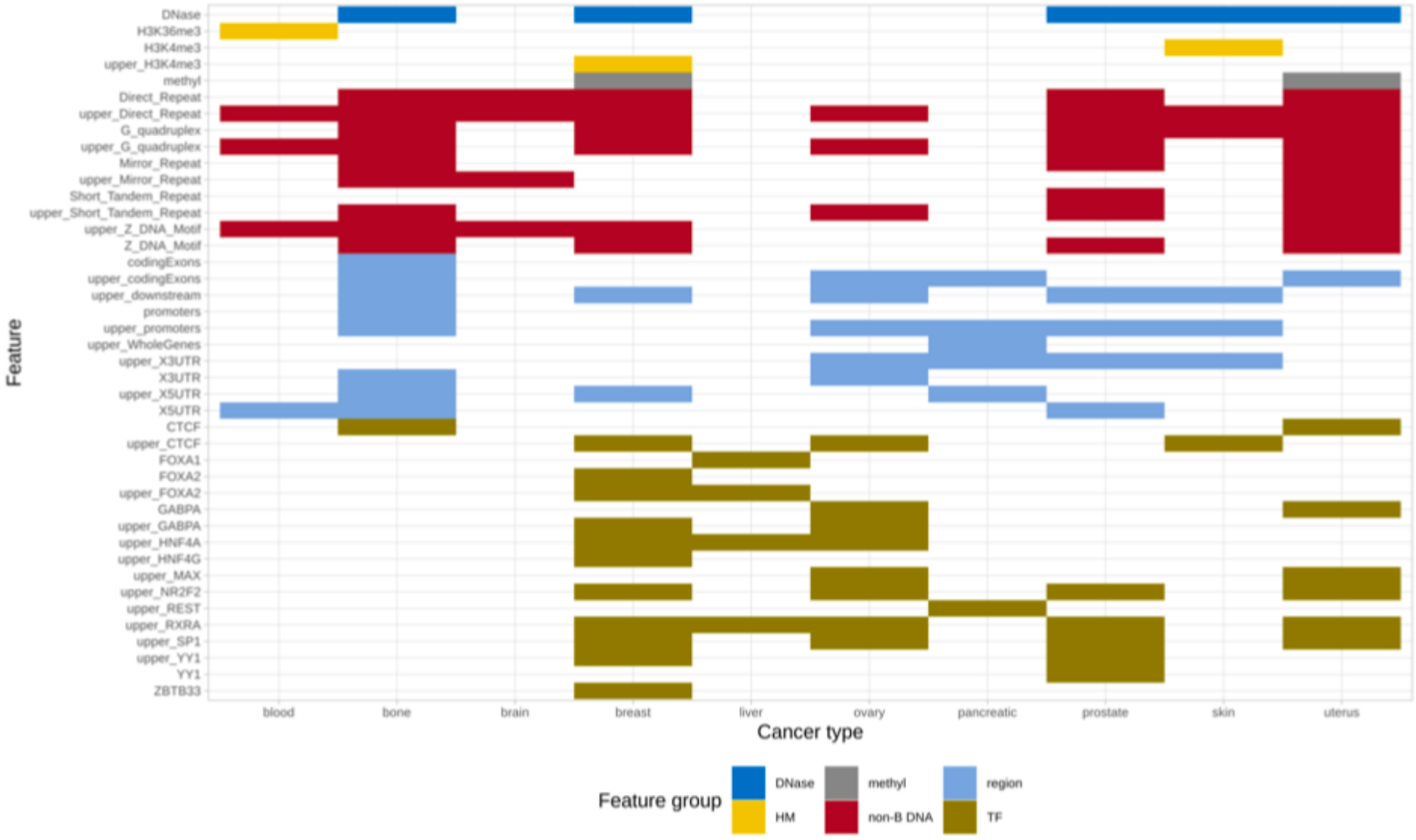
Stem-loops + quadruplexes

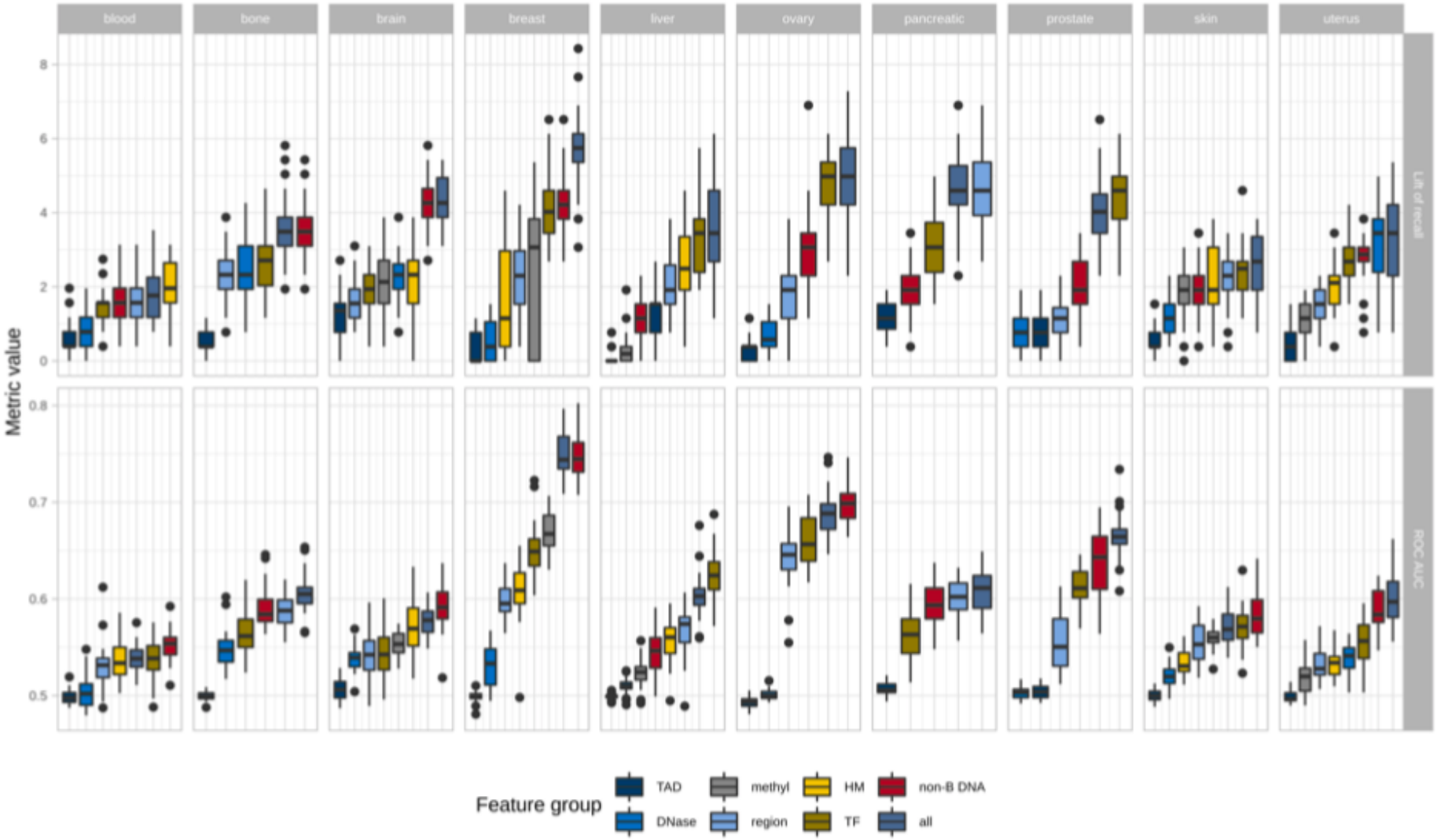


Comprehensive analysis of cancer breakpoint hotspots with machine learning models

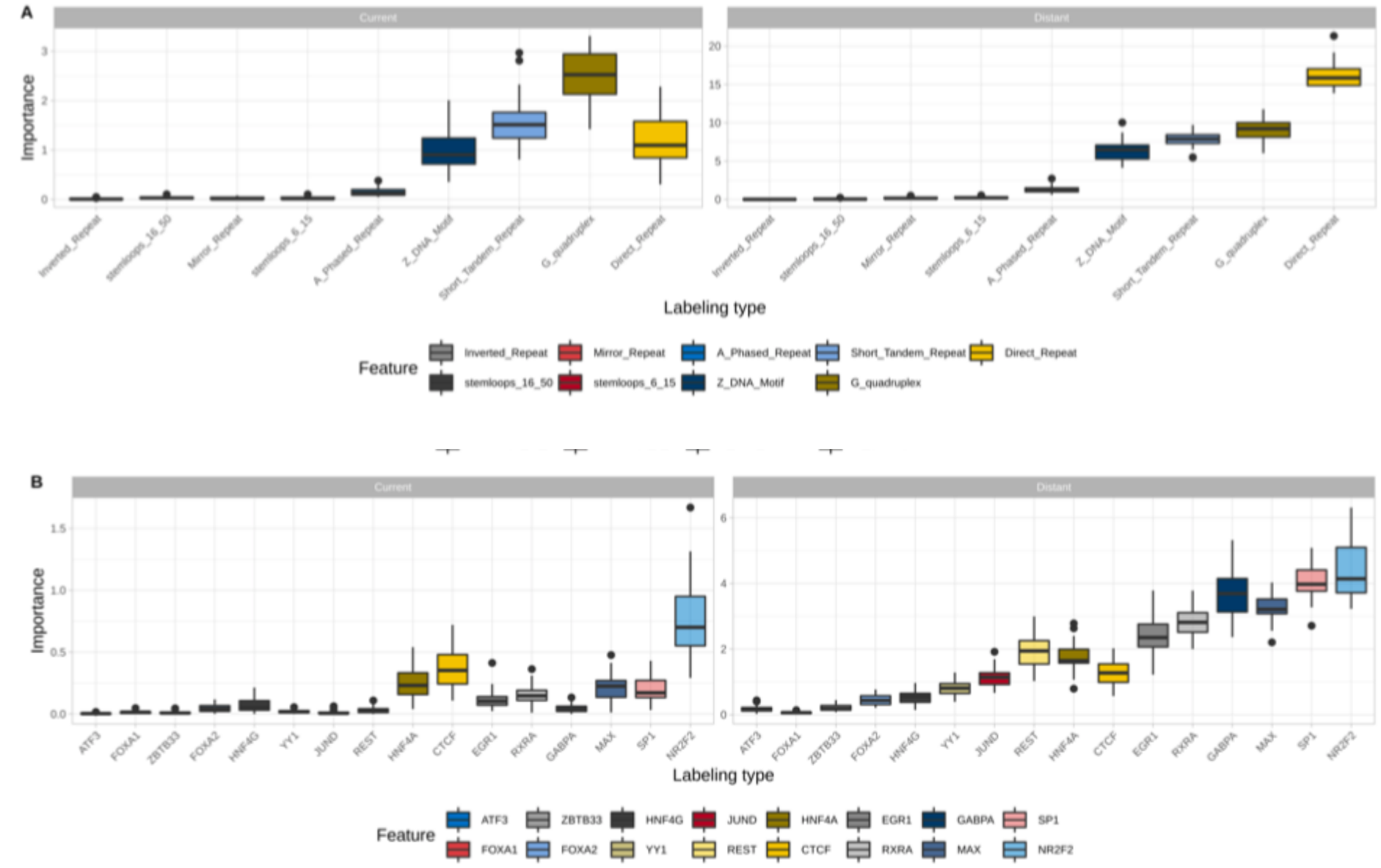
- non-B DNA structures
- genomic regions
- chromatin structure
- transcription factor binding sites
- epigenetic markers

Вклад разных групп





Contribution of different feature groups separately and all together.


A
Current
Distant
Importance
Inverted_Repeat
stemloops_16_50
Mirror_Repeat
stemloops_6_15
A_Phased_Repeat
Z_DNA_Motif
Short_Tandem_Repeat
G_quadruplex
Direct_Repeat
Labeling type
Feature
B
ATF3
FOXA1
ZBTB33
FOXA2
HNF4G
YY1
JUND
REST
HNF4A
CTCF
EGR1
RXRA
GABPA
MAX
SP1
NR2F2

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

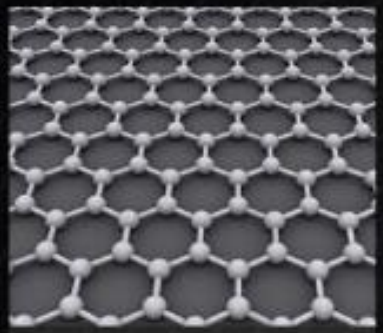
Технический директор Google

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

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

Creative AI is now capable of making complex art and music

Autonomous vehicles now dominate our roads

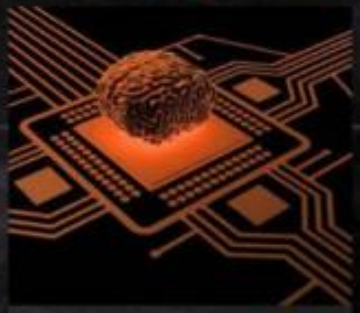
Humans begin to develop deep relationships with AI

Language translation machines are now routinely used in conversation

2029-2039

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



Computers are now capable of autonomously learning and creating new knowledge

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

VR eyeglasses and heaphones are replaced with computer implants

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

- 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

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.

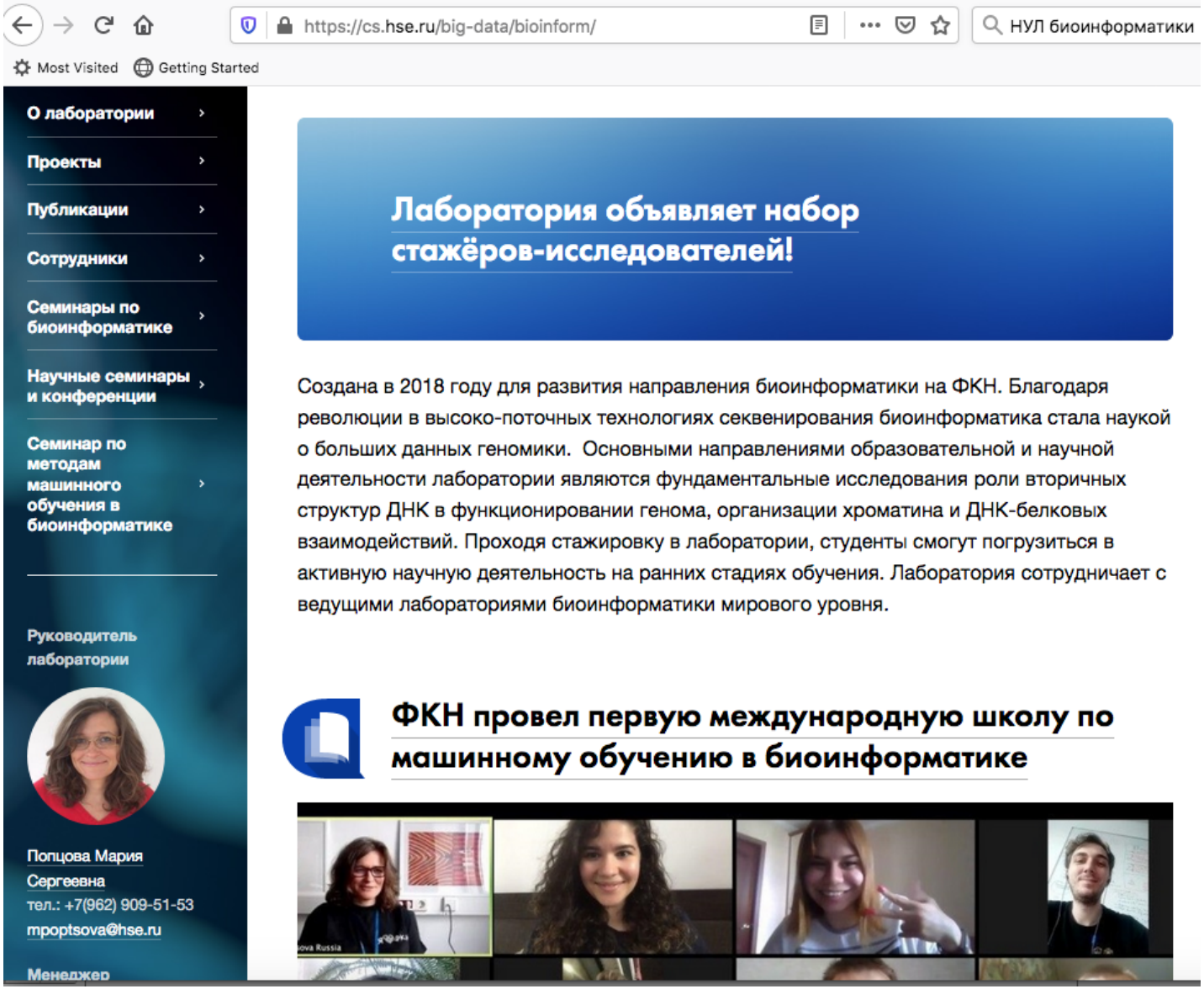
Программа нацелена на подготовку будущих лидеров биоинформатических исследований, обладающих компетенциями для разработки и применения на практике вычислительных методов с целью решения задач в различных областях биологии и медицины.

Задать вопрос о программе

Академический руководитель
Гельфанд Михаил Сергеевич

О программе
Абитуриентам
Студентам

Научно-учебная лаборатория биоинформатики



Спасибо за внимание