СЕКВЕНИРОВАНИЕ ОТ A ДО N

Анастасия Жарикова

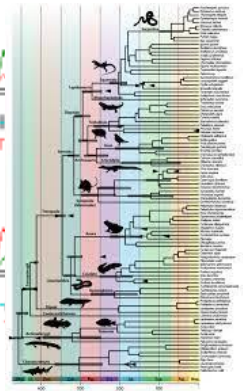
ФББ МГУ - 29 марта 2023

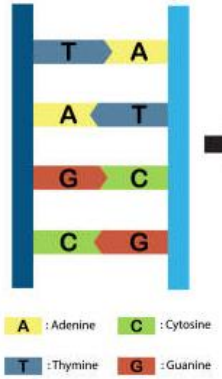
СЕКВЕНИРОВАНИЕ

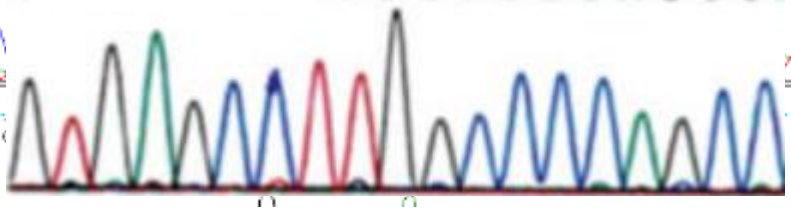
Криминалистика

Генетическое тестирование

Найден ген долголетия, ожирения, убийцы, гениальности

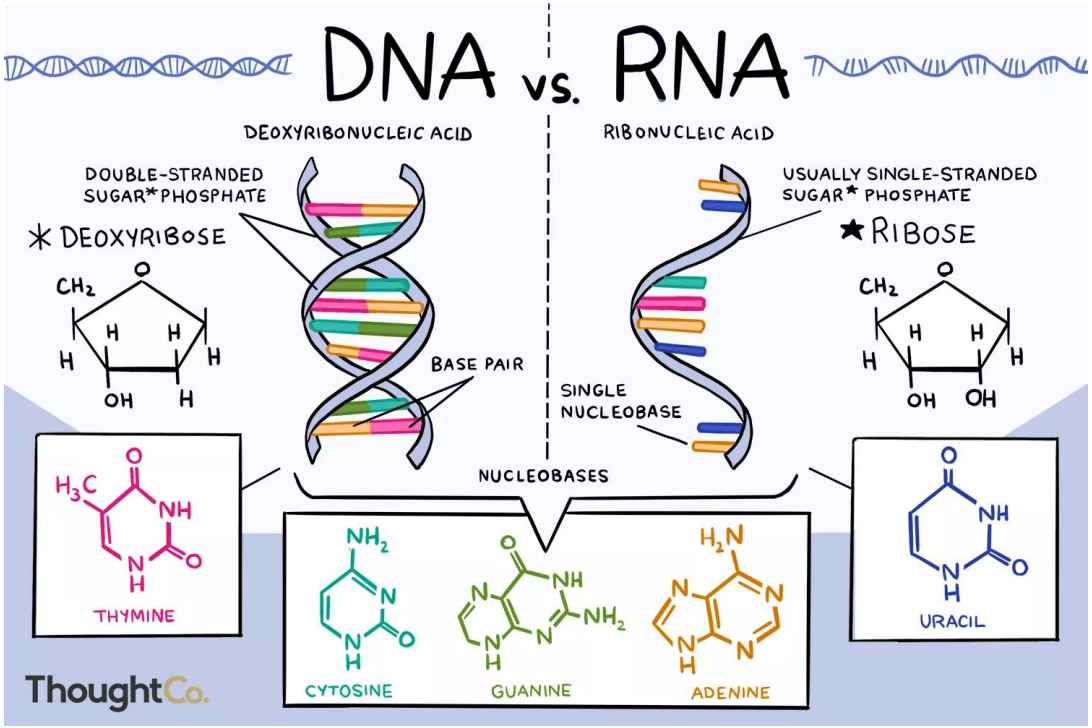






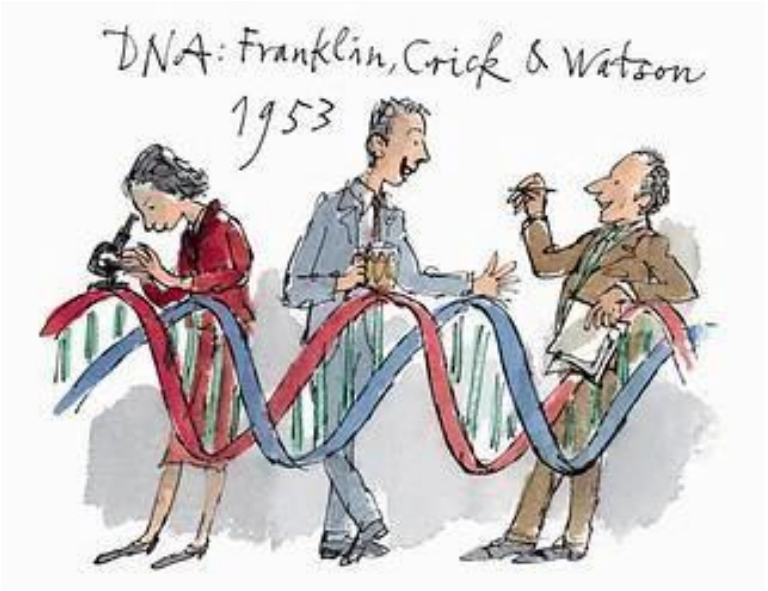
СЕКВЕНИРОВАНИЕ

- Установление последовательности нуклеиновых кислот (с белками тоже можно, но иначе)



ref

СТРУКТУРА ДВОЙНОЙ СПИРАЛИ ДНК



equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).

[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).

[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (3) (1950).

[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik.* (Stockholm), **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three inter-twined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate–sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate di-ester groups joining β-D-deoxy-ribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the *z*-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical *z*-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on inter-atomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

ЧТО КОНКРЕТНО

- ДНК: секвенировать весь геном

Откуда взять геном?

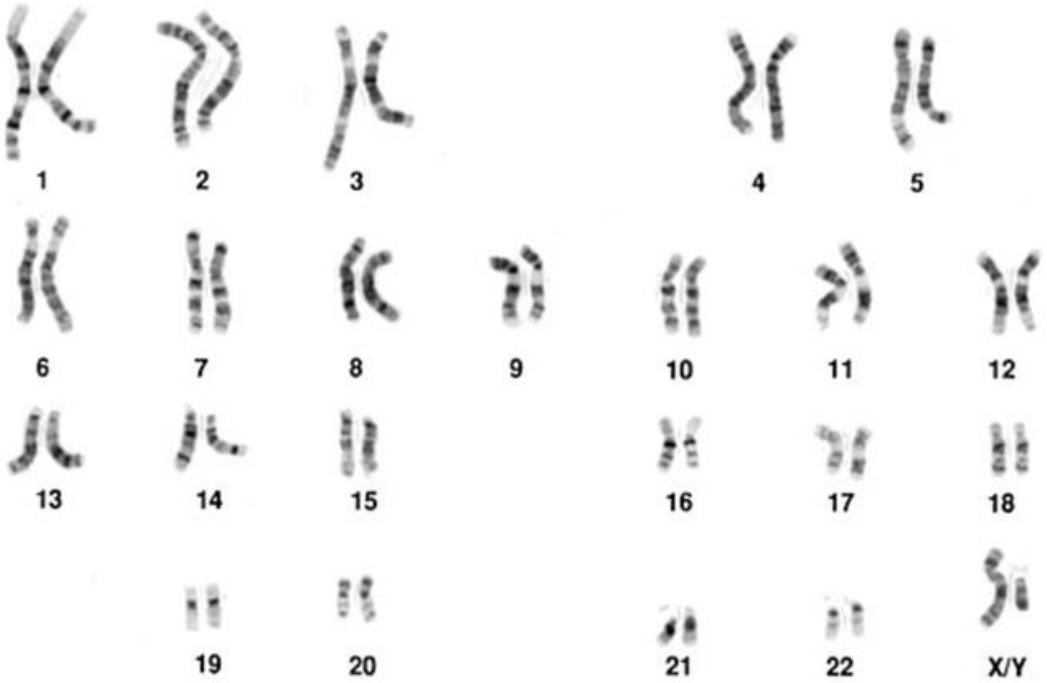
Нужно выделить ДНК

ОТКУДА ВЫДЕЛИТЬ ДНК?

- Кровь
- Слюна
- Кожа
- Волосы
- Лист
- Много клеток
- Вся муха
- Кусок органа
- Биопсия опухоли
- …

У ЧЕЛОВЕКА 23 ПАРЫ ХРОМОСОМ

- Это много или мало?



ref

ЧИСЛО ХРОМОСОМ У РАЗНЫХ ВИДОВ

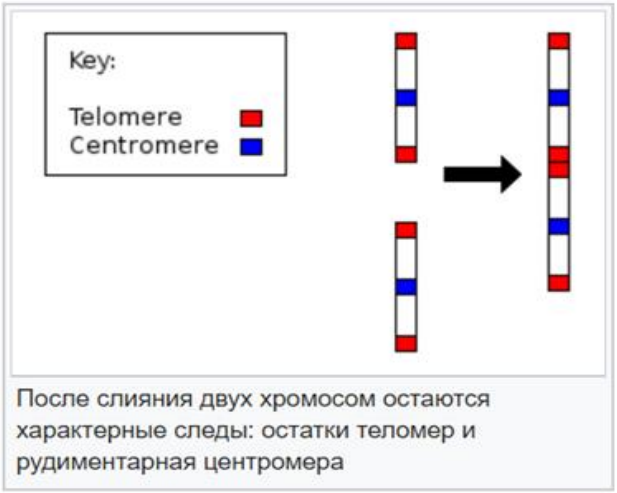
Гиббоны - 44

Макака - 42

Капуцин - 54

48

46



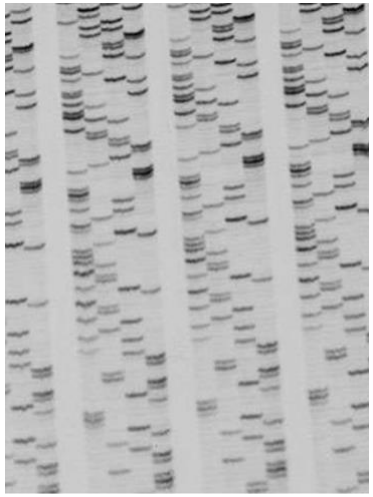
После слияния двух хромосом остаются характерные следы: остатки теломер и рудиментарная центромера

ЧИСЛО ХРОМОСОМ У РАЗНЫХ ВИДОВ

Муравей (Myrmecia pilosula) – 2
Плодовая мушка – 8
Арабидопсис – 10
Голубь – 16
Кошка – 38
Лиса – 34
Мышь – 40
Собака – 78
Утка – 80
Сазан – 104
Корова – 120
Рак (Cambarus clarkii) – 200
Хвощ– 216
Краб – 254
Бабочка – 380

ref

ТЕХНОЛОГИИ СЕКВЕНИРОВАНИЯ «ТОГДА» И «СЕЙЧАС»



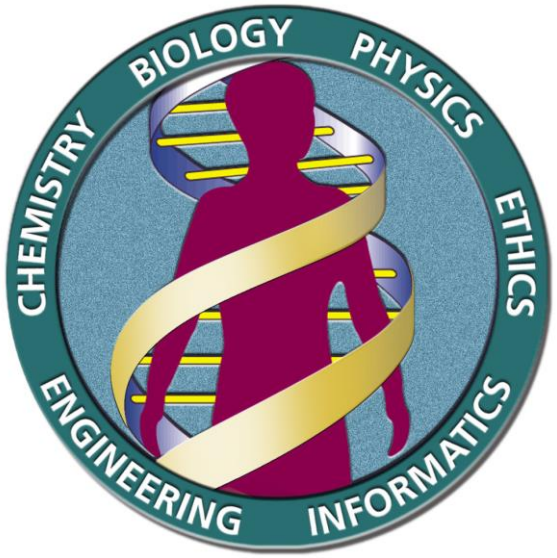
1 поколение

2 поколение

3 поколение

ПРОЕКТ «ГЕНОМ ЧЕЛОВЕКА»

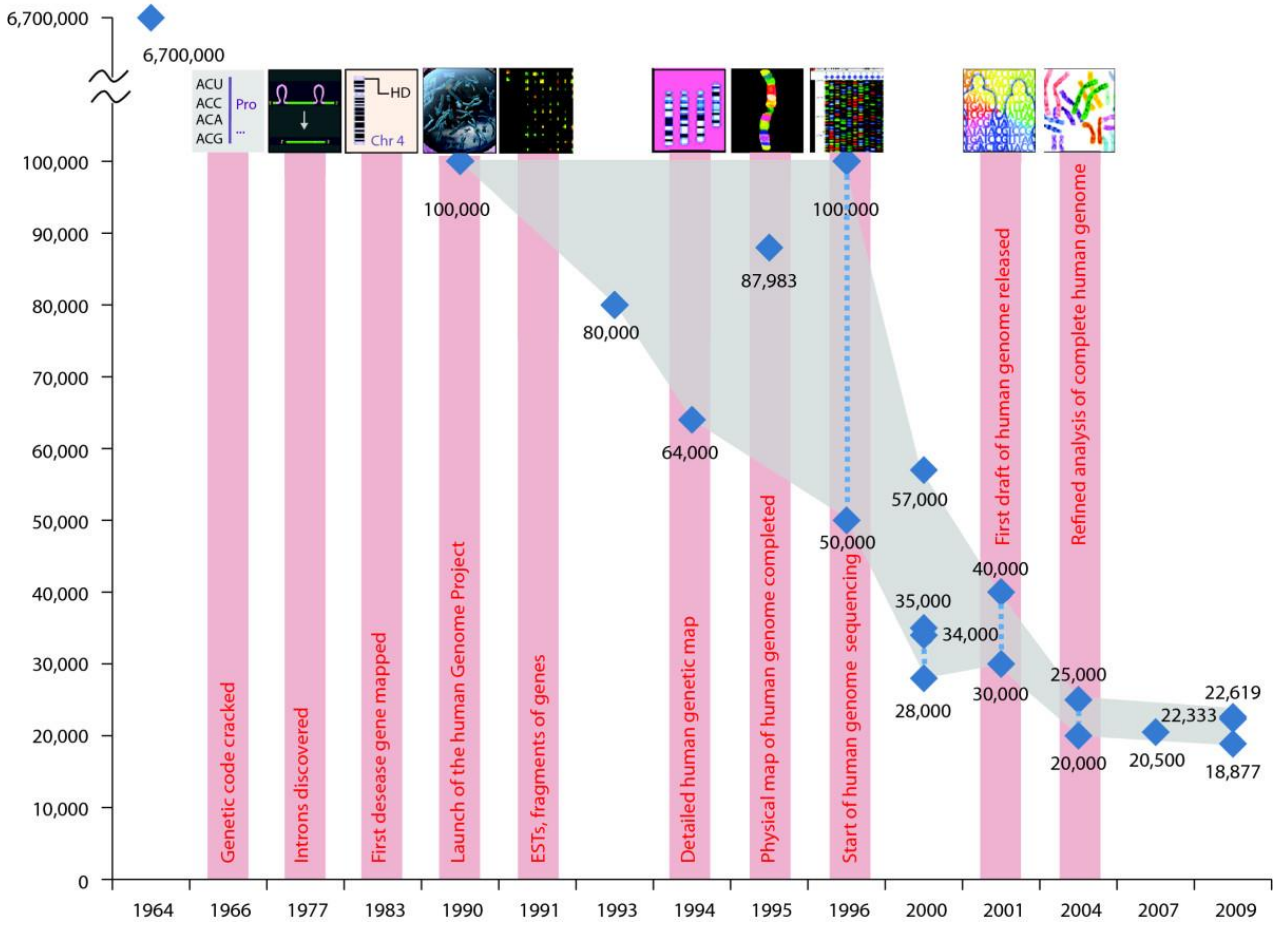
- Начали в 1990 году
- Приняло участие 18 стран
- 2001 – «черновое окончание проекта»
- ~ 10 лет и 3 млрд $



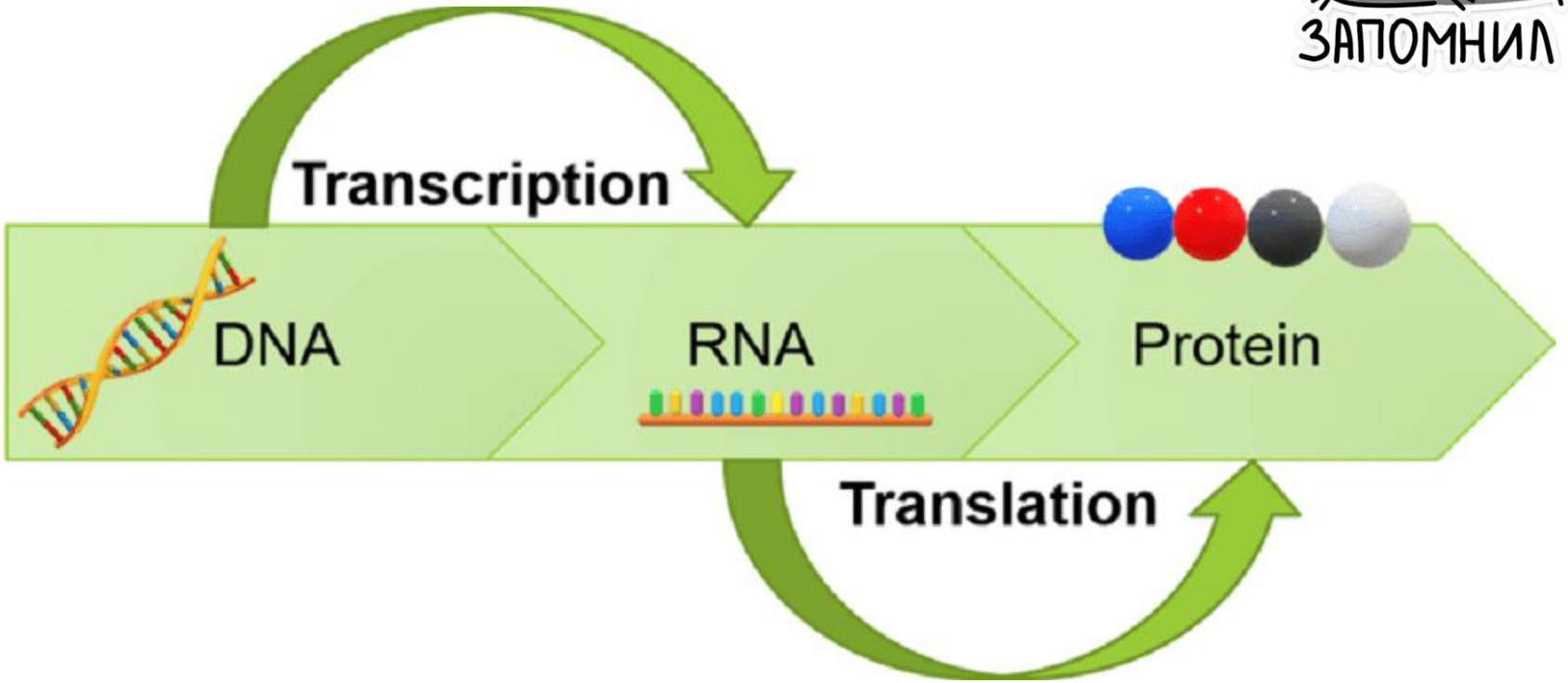
КОНКУРЕНЦИЯ

- Проект Крейга Вентера
- **Celera Corporation**

ОЦЕНКА КОЛИЧЕСТВА ГЕНОВ У ЧЕЛОВЕКА

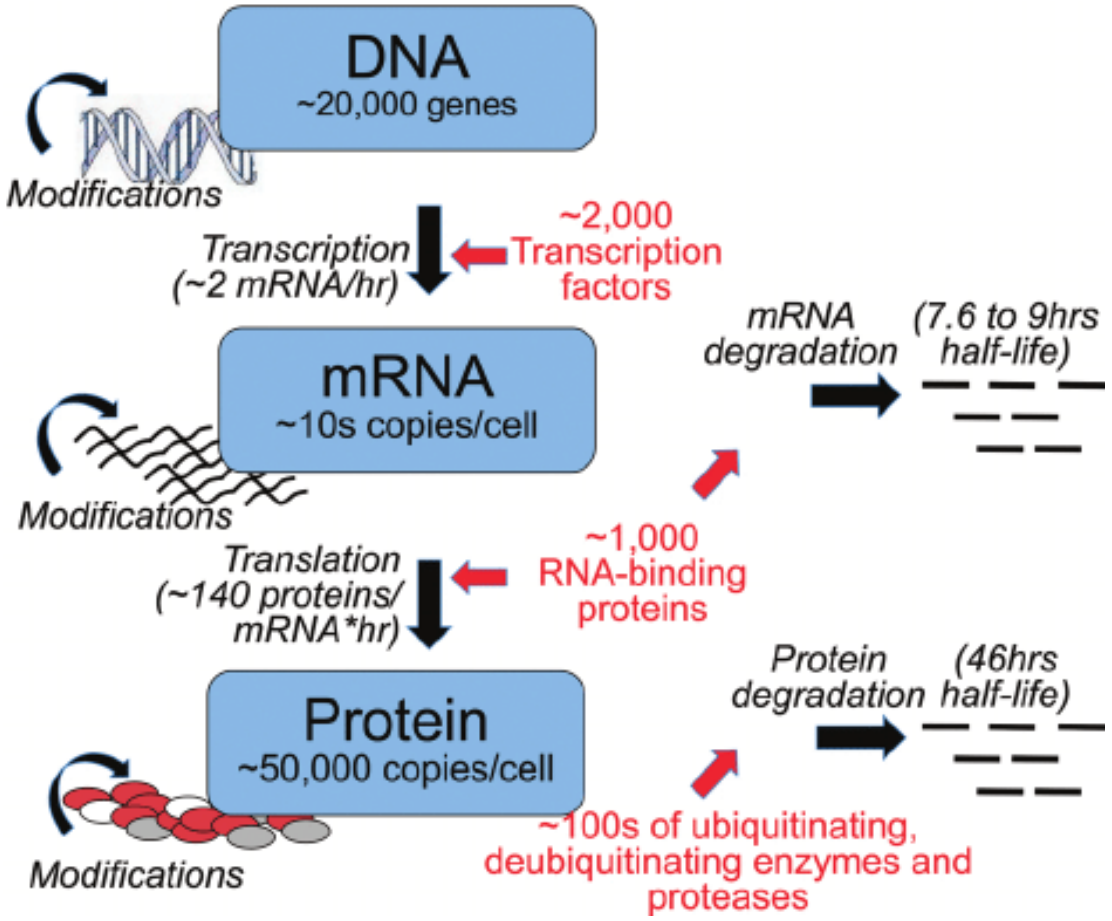


ЦЕНТРАЛЬНАЯ ДОГМА МОЛЕКУЛЯРНОЙ БИОЛОГИИ



ref

ЦЕНТРАЛЬНАЯ ДОГМА МОЛЕКУЛЯРНОЙ БИОЛОГИИ

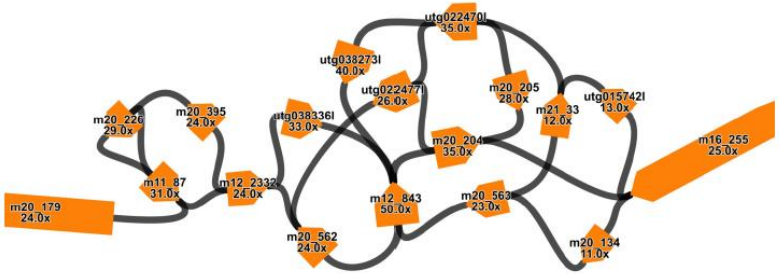


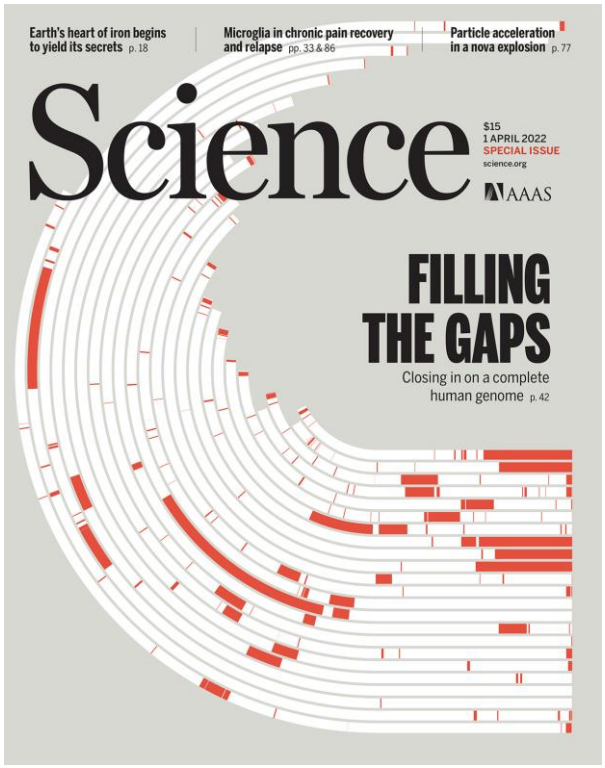
ref

БЕЛОК-КОДИРУЮЩИЕ ГЕНЫ

Картофель – 39 000
Человек ~ 20 000
Черви – 14 000
Мухи – 12 000
Грибы – 6 000
Бактерии – 2 000 – 4 000
Микоплазмы - 500
Вирус гриппа – 12

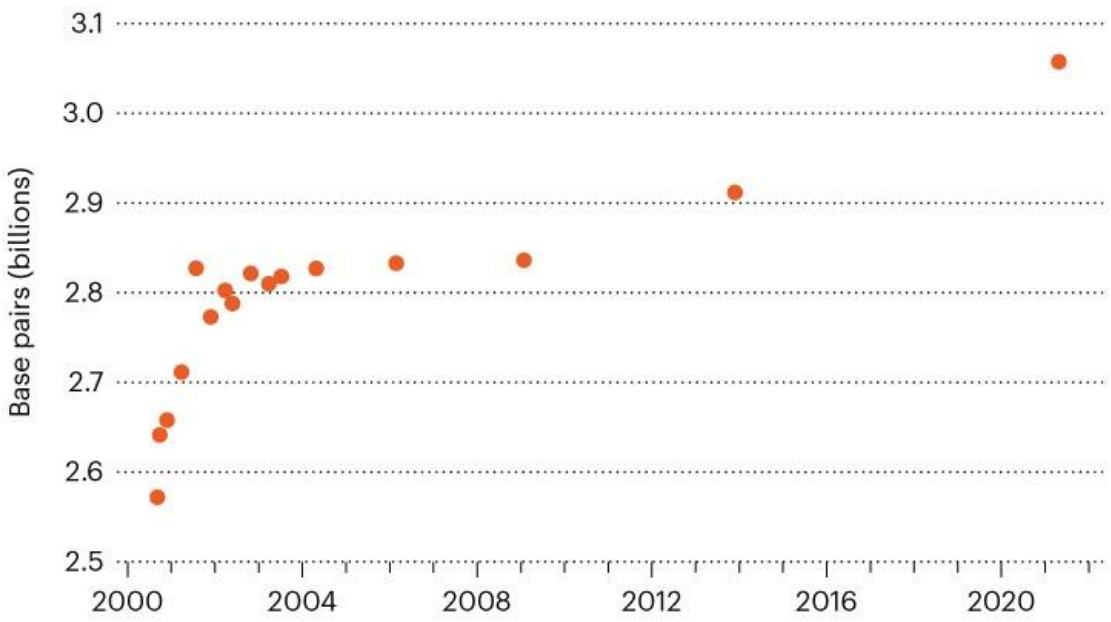
T2T





COMPLETING THE HUMAN GENOME

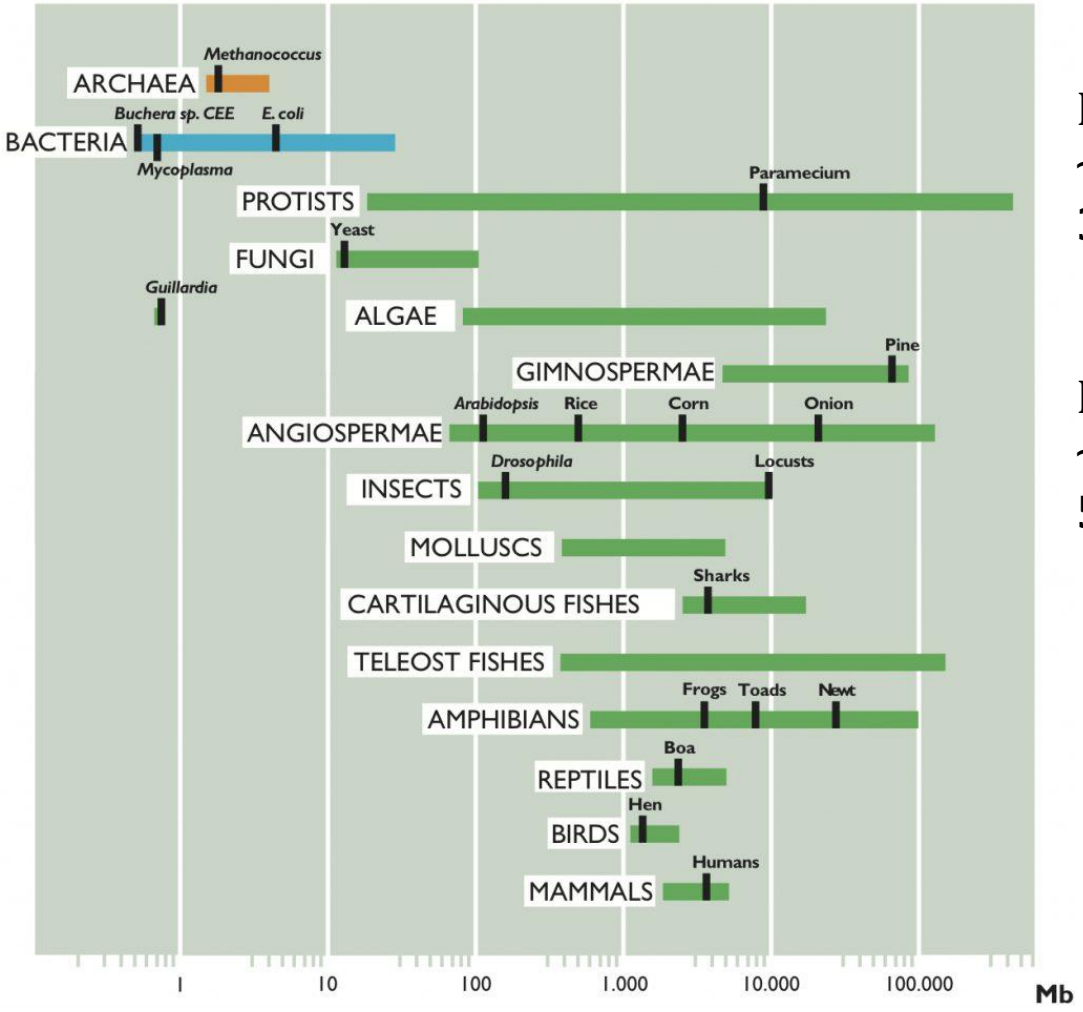
Researchers have been filling in incompletely sequenced parts of the human reference genome for 20 years, and have now almost finished it, with 3.05 billion DNA base pairs.



0.3% of sequence might still have errors. Includes X but not Y chromosome. Count excludes mitochondrial DNA.

©nature

РАЗМЕРЫ ГЕНОМОВ

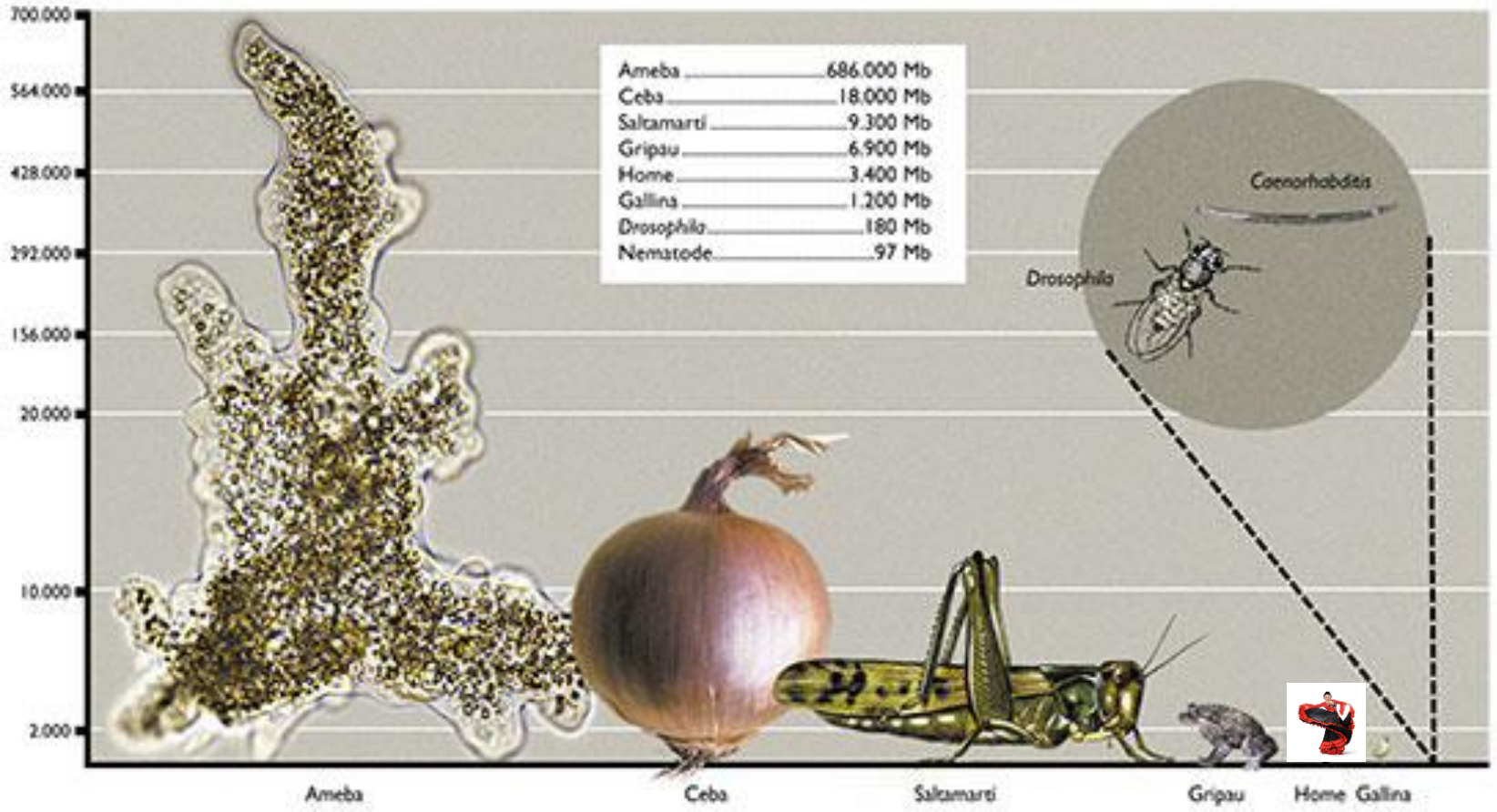


Размер генома человека:
~ 3 млрд пар нуклеотидов =
3 гигабазы

Размер генома кишечной палочки:
~ 5 млн пар нуклеотидов =
5 Мб

ref

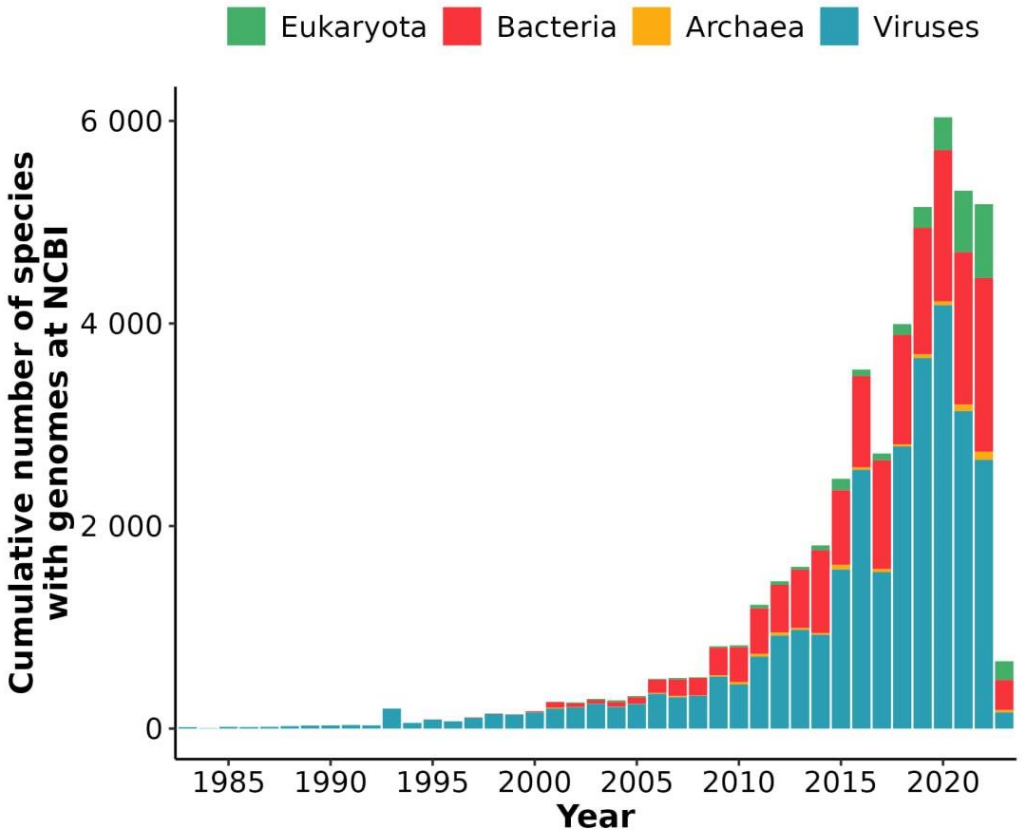
РАЗМЕРЫ ГЕНОМОВ



ref

НАКОПЛЕНИЕ ИНФОРМАЦИИ

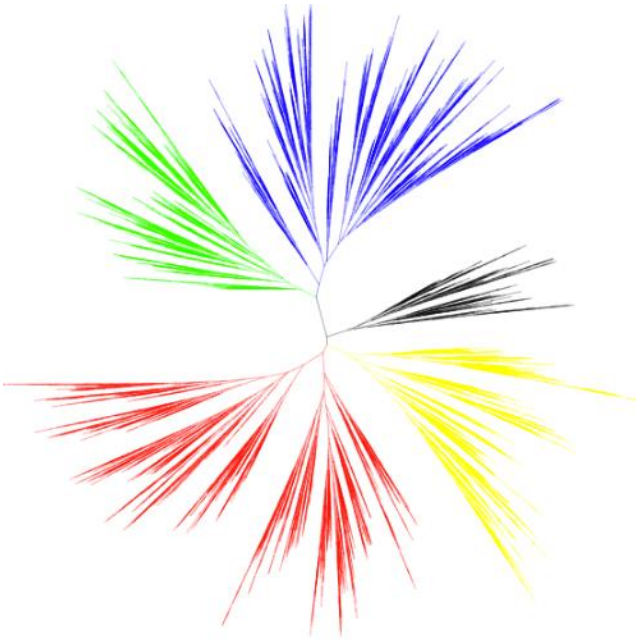
Хромосомы + полные геномы = 46903



by Anna Valyaeva – NCBI data – 29.03.2023

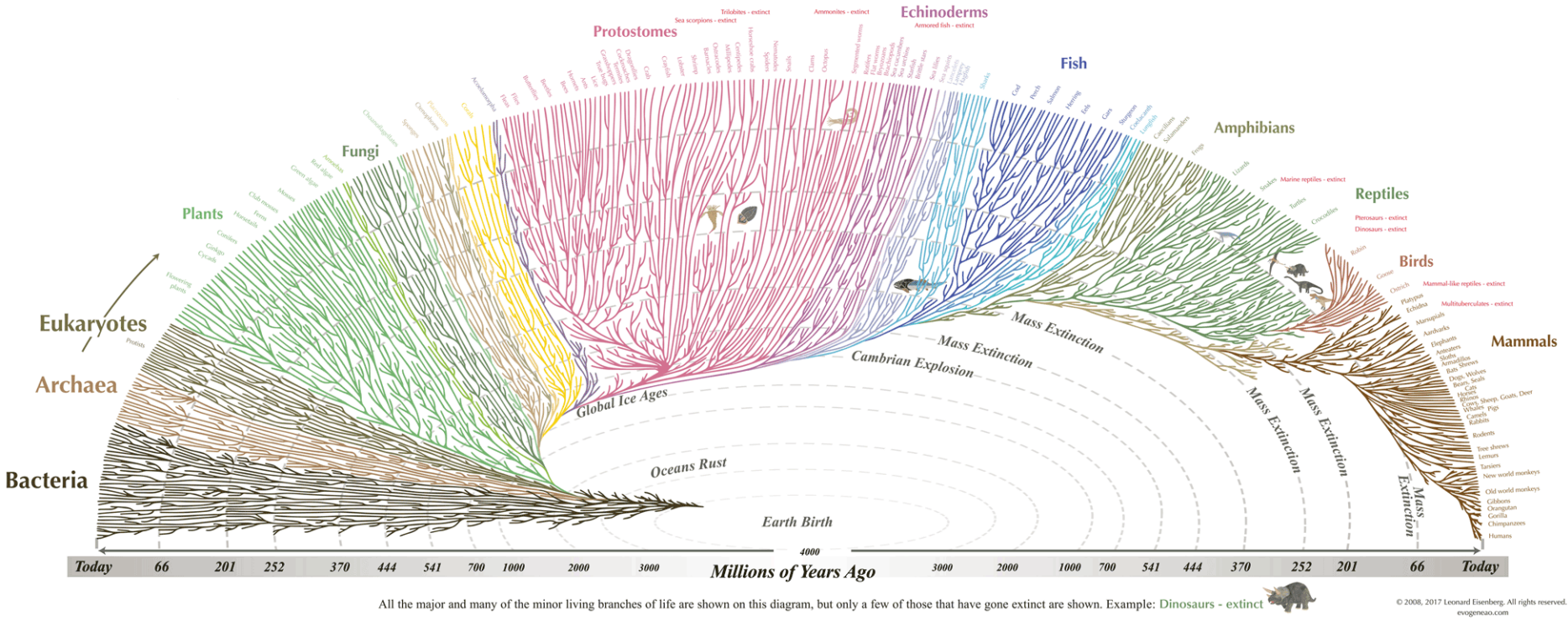
МОЛЕКУЛЯРНАЯ ФИЛОГЕНЕТИКА

- Установление родственных связей между организмами
- Исследование происхождения штаммов вирусов и бактерий
- Исследование злокачественных опухолей



Тут очень красиво!

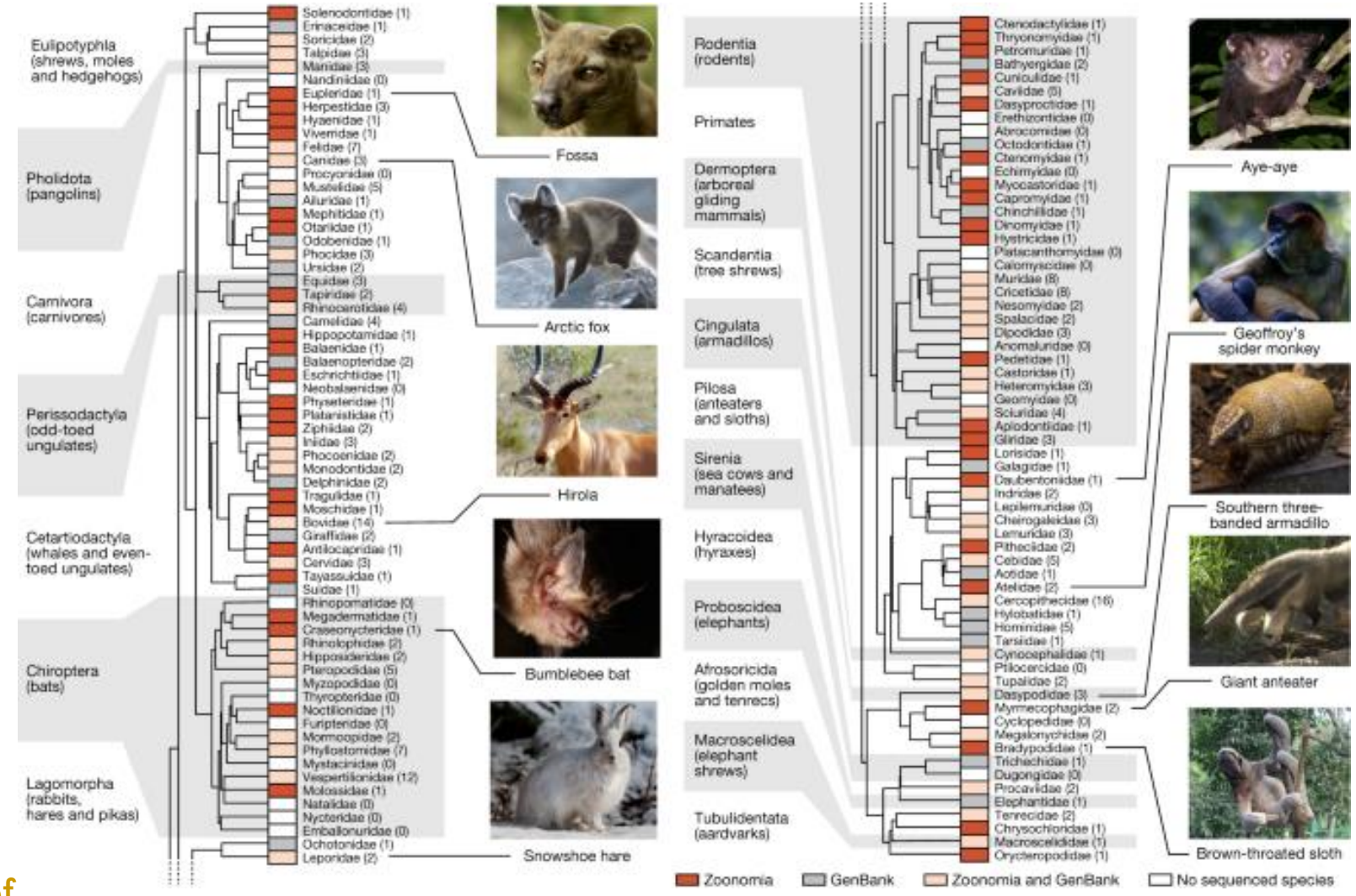
TREE OF LIFE



All the major and many of the minor living branches of life are shown on this diagram, but only a few of those that have gone extinct are shown. Example: Dinosaurs - extinct

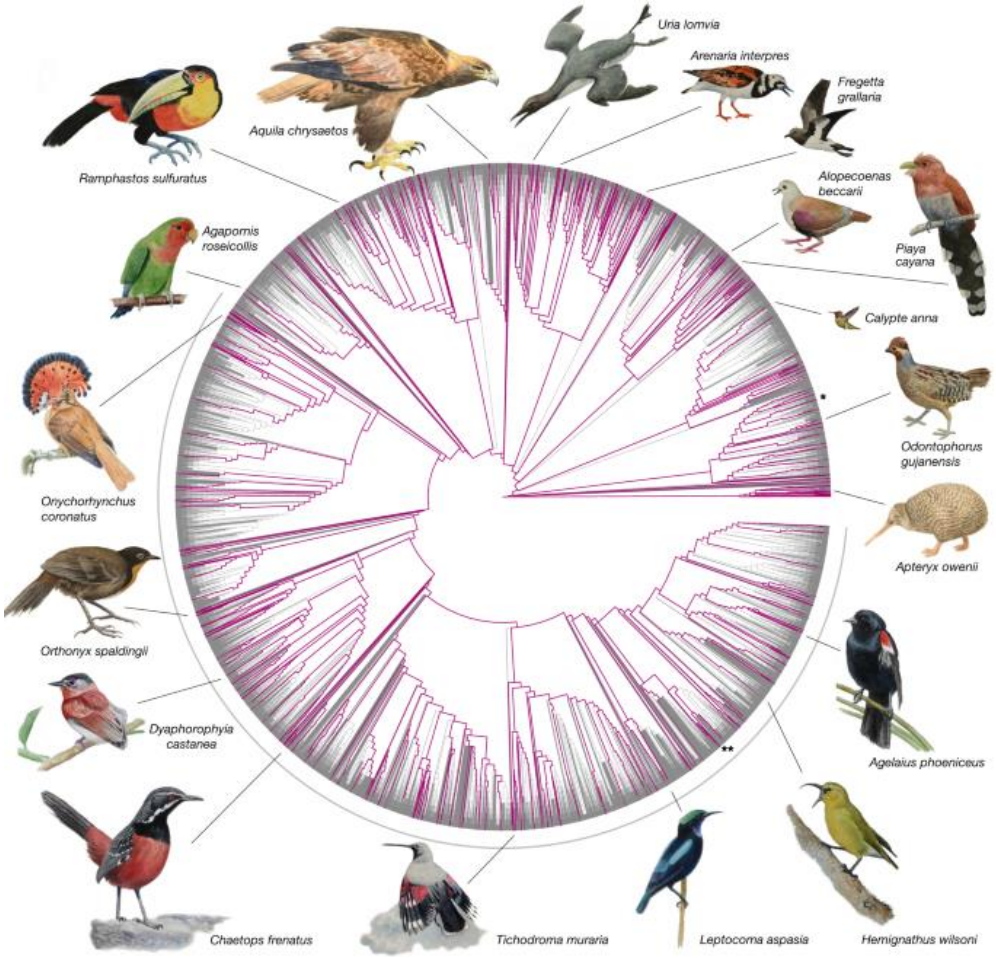
Tree of life

ZOONOMIA PROJECT



ref

10,135 BIRD SPECIES



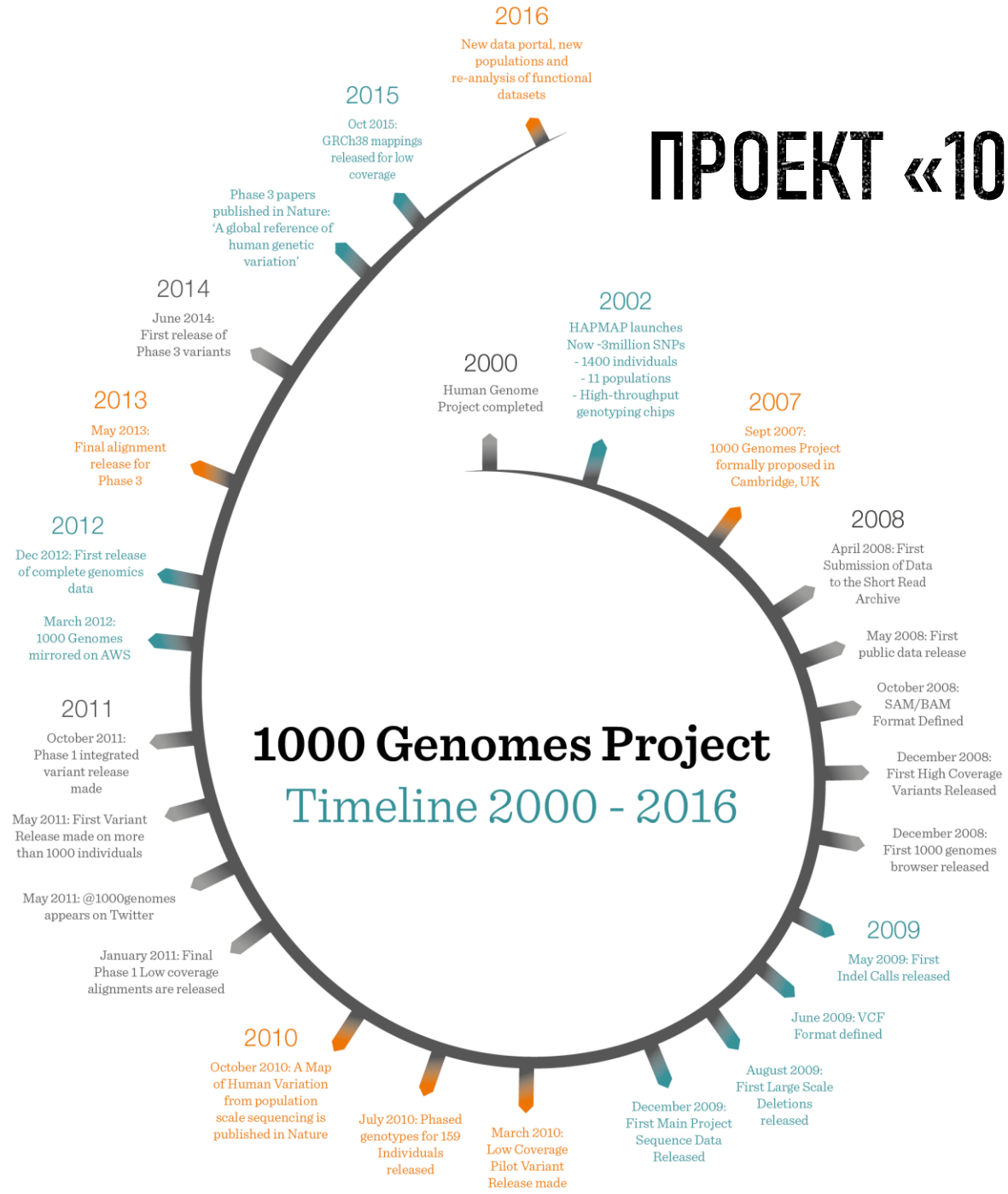
ref

МЕТАГЕНОМИКА

- Метагеномика изучает наборы генов всех микроорганизмов в образце
- Применение:
  - Изменение микрофлоры человека (конкретных органов и систем) коррелирует с изменениями в медицинских показателях
  - Идентификация видов
  - Изучение почвенных микробных сообществ важно в сельском хозяйстве
  - В химической и фармацевтической отрасли

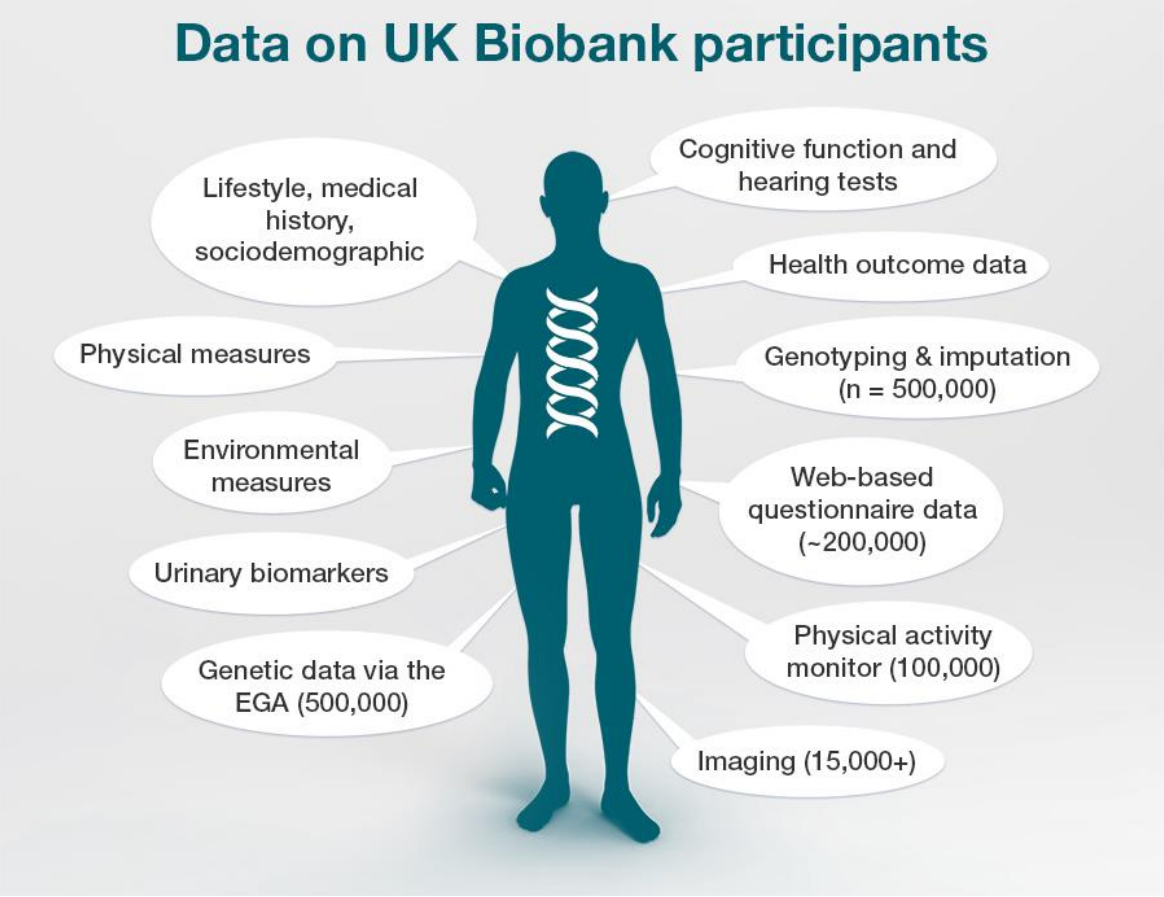
ref

ПРОЕКТ «1000 ГЕНОМОВ»



ref

UK BIOBANK



ref

БИОБАНК ФГБУ «НМИЦ ТПМ» МИНЗДРАВА РОССИИ

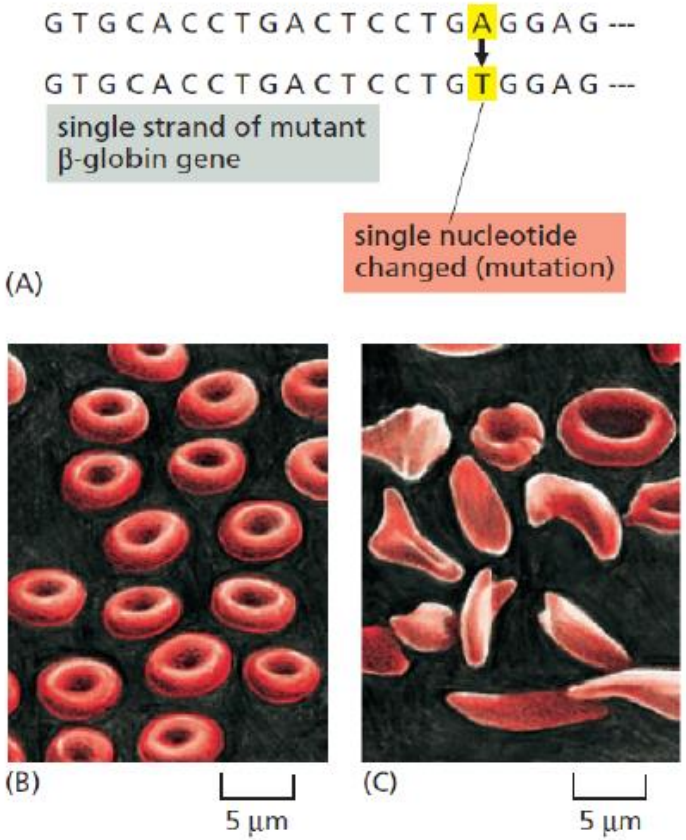
- коллекции биоматериала от пациентов с сердечно-сосудистыми заболеваниями (ИБС, аритмии, кардиомиопатии и др.), патологией желудочно-кишечного тракта, ожирением и другими ХНИЗ
- Коллекция биоматериала от репрезентативной выборки населения российских регионов (41 регион, 79516 участников)
- Общее количество образцов сыворотки, плазмы и цельной крови для исследовательских целей на начало 2023 года — >773 тыс.

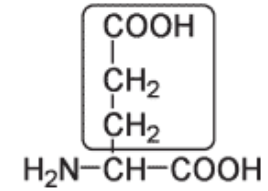
КЛИНИЧЕСКИЕ ВОПРОСЫ

- 1000 genomes, gnomAD: частоты вариантов в популяциях
- GWAS: поиск полиморфизмов, ассоциированных с болезнями:
  - моногенные (муковисцидоз, ген CFTR)
  - полигенные (ишемическая болезнь сердца, шизофрения, ...)
- Фармакогенетика и индивидуальные особенности
  - варфарин
  - исследование генов из системы свертывания крови
- PSCK9 и холестерин

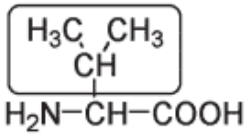
СЕРПОВИДНО-КЛЕТОЧНАЯ АНЕМИЯ

- Замена всего одного (!!!!) нуклеотида





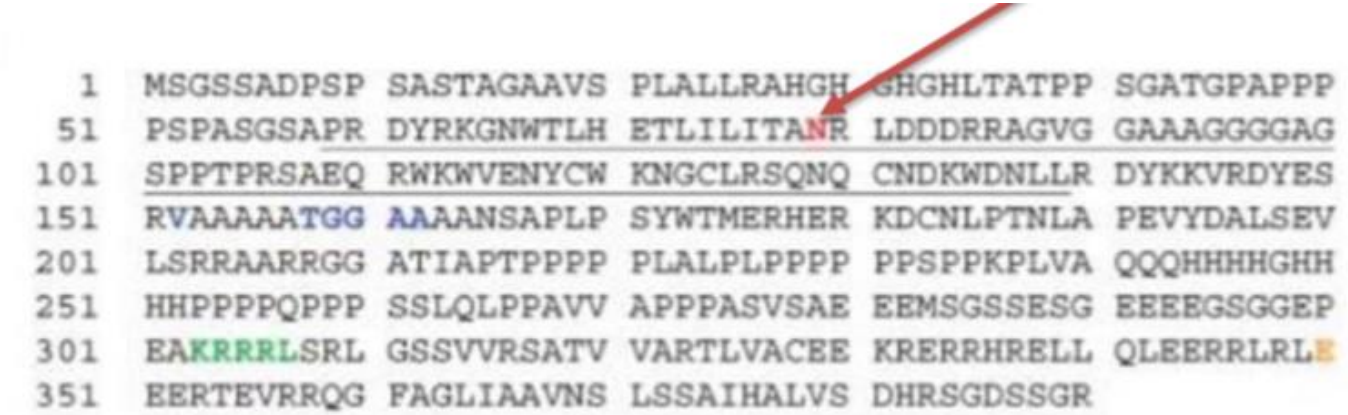
Глутаминовая кислота



Валин

Alberts - *Essential Cell Biology*

ДОМЕСТИКАЦИЯ РИСА



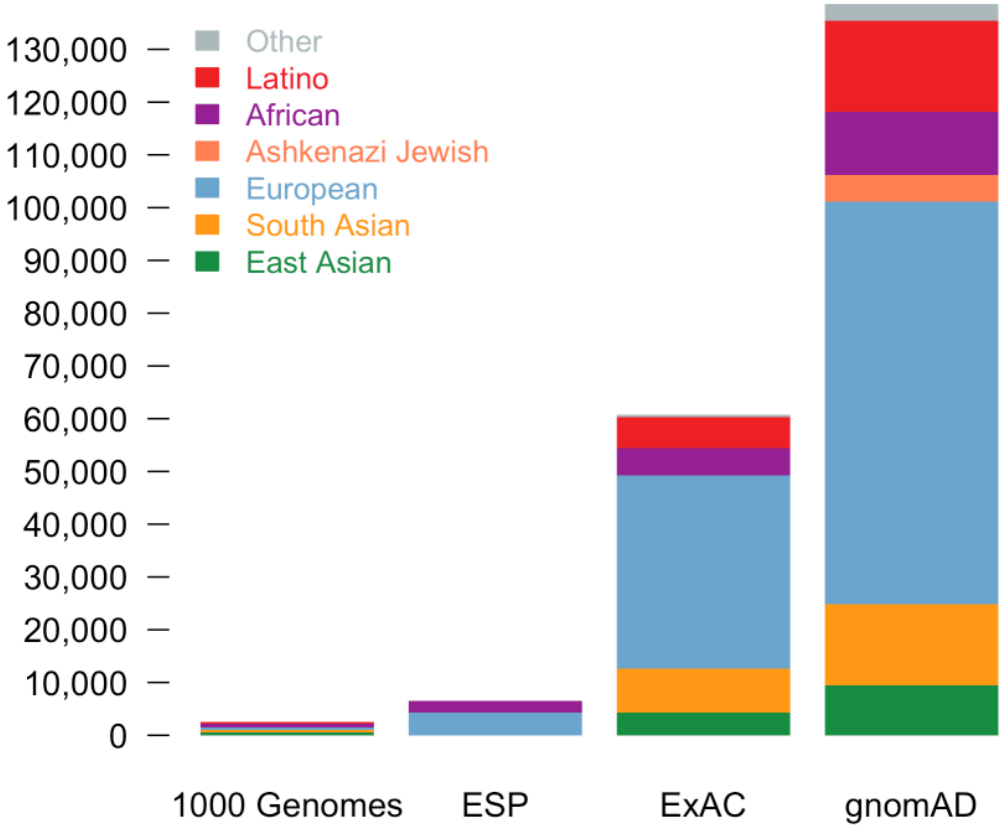
```
  1  MSGSSADPSP SASTAGAAVS PLALLRAHGH GHGHLTATPP SGATGPAPPP
 51  PSPASGSAPR DYRKGNWTLH ETLILITANR LDDDRRAGVG GAAAGGGGAG
101  SPPTPRSAEQ RWKWVENYCW KNGCLRSQNQ CNDKWDNLLR DYKKVRDYES
151  RVAAAAATGG AAAANSAPLP SYWTMERHER KDCNLPTNLA PEVYDALSEV
201  LSRRAARRGG ATIAPTPPPP PLALPLPPPP PPSPPKPLVA QQQHHHHGHH
251  HHPPPPQPPP SSLQLPPAVV APPPASVSAE EEMSGSSESG EEEEGSGGEP
301  EAKRRRLSRL GSSVVRSATV VARTLVACEE KRERRHRELL QLEERRLRLE
351  EERTEVRRQG FAGLIAAVNS LSSAIHALVS DHRSGDSSGR
```

sh4

Li et al., Science, 2006

Дикий рис	Культурный рис
AAG	AAT
Лизин	Аспарагин

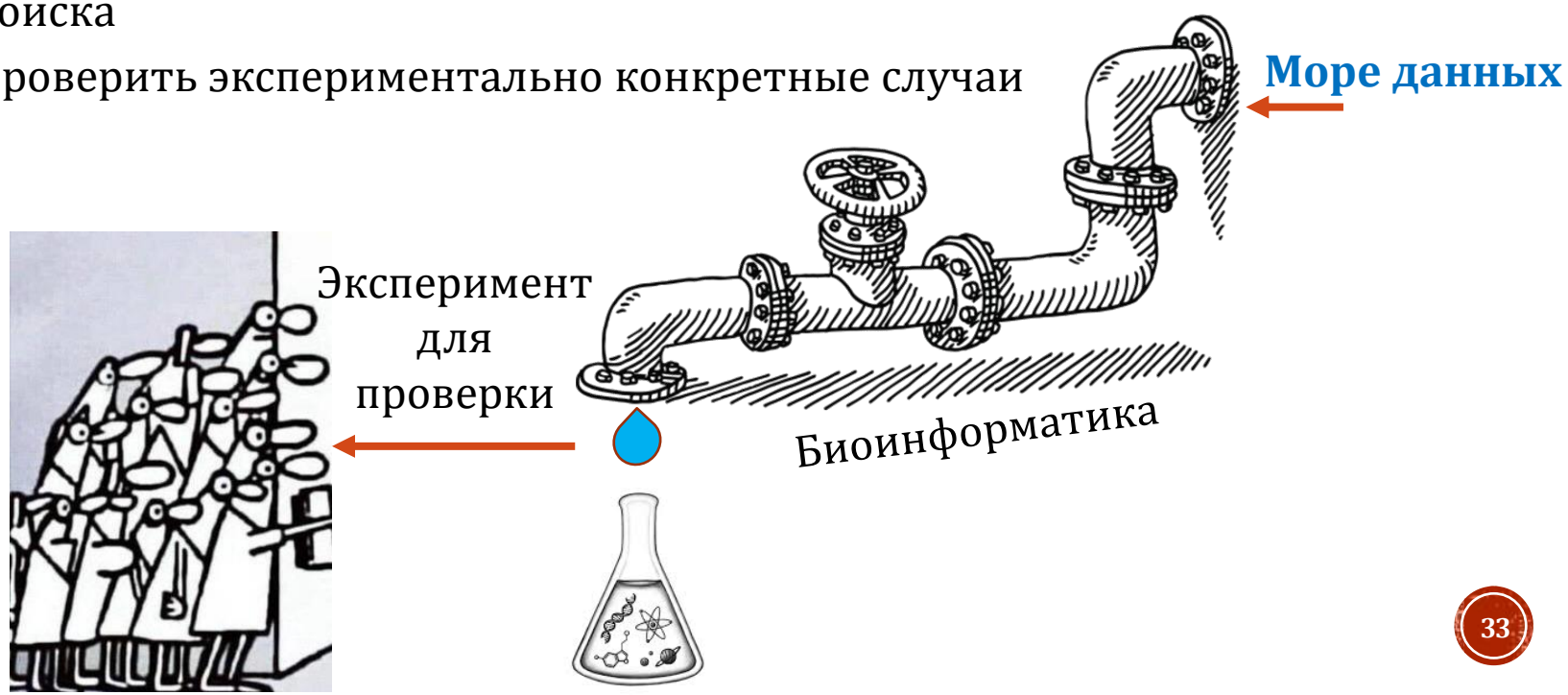
CLINVAR


130,000
120,000
110,000
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Other
Latino
African
Ashkenazi Jewish
European
South Asian
East Asian
1000 Genomes
ESP
ExAC
gnomAD

ref

ЗАЧЕМ БИОИНФОРМАТИКА?

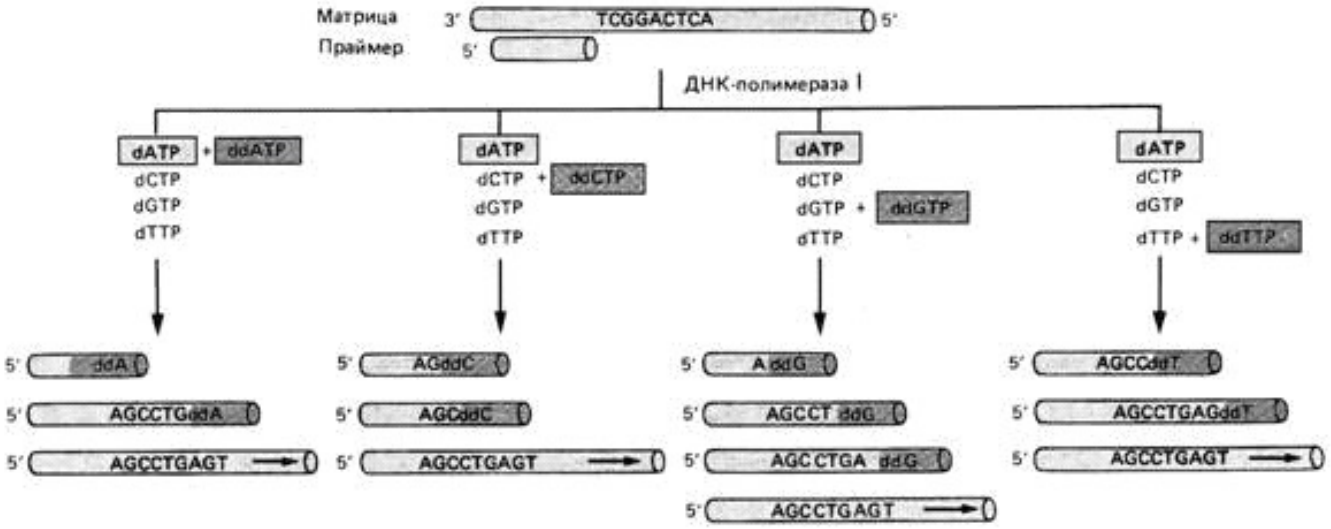
- Накоплено ОЧЕНЬ много данных
- Обработать, проанализировать и структурировать всю имеющуюся информацию вручную не представляется возможным
- Идея (одна из):
  - С помощью автоматических подходов минимизировать область поиска
  - Проверить экспериментально конкретные случаи

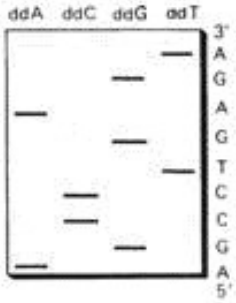


КАК СЕКВЕНИРОВАТЬ?

- 1 поколение: маленькими кусочками, долго, точно (Сэнгер)
- 2 поколение: маленькими кусочками, но сразу много (**NGS**)
- 3 поколение: всю молекулу целиком, быстро, много ошибок (одномолекулярное секвенирование)

МЕТОД ТЕРМИНАТОРОВ – 1977 ГОД





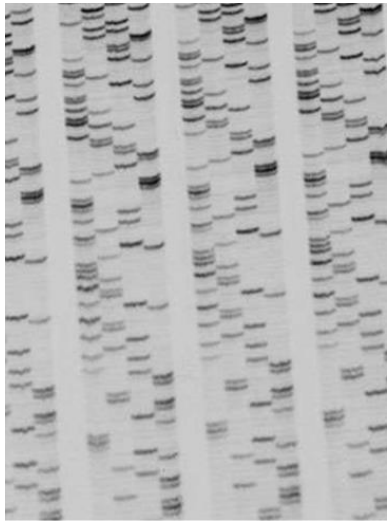
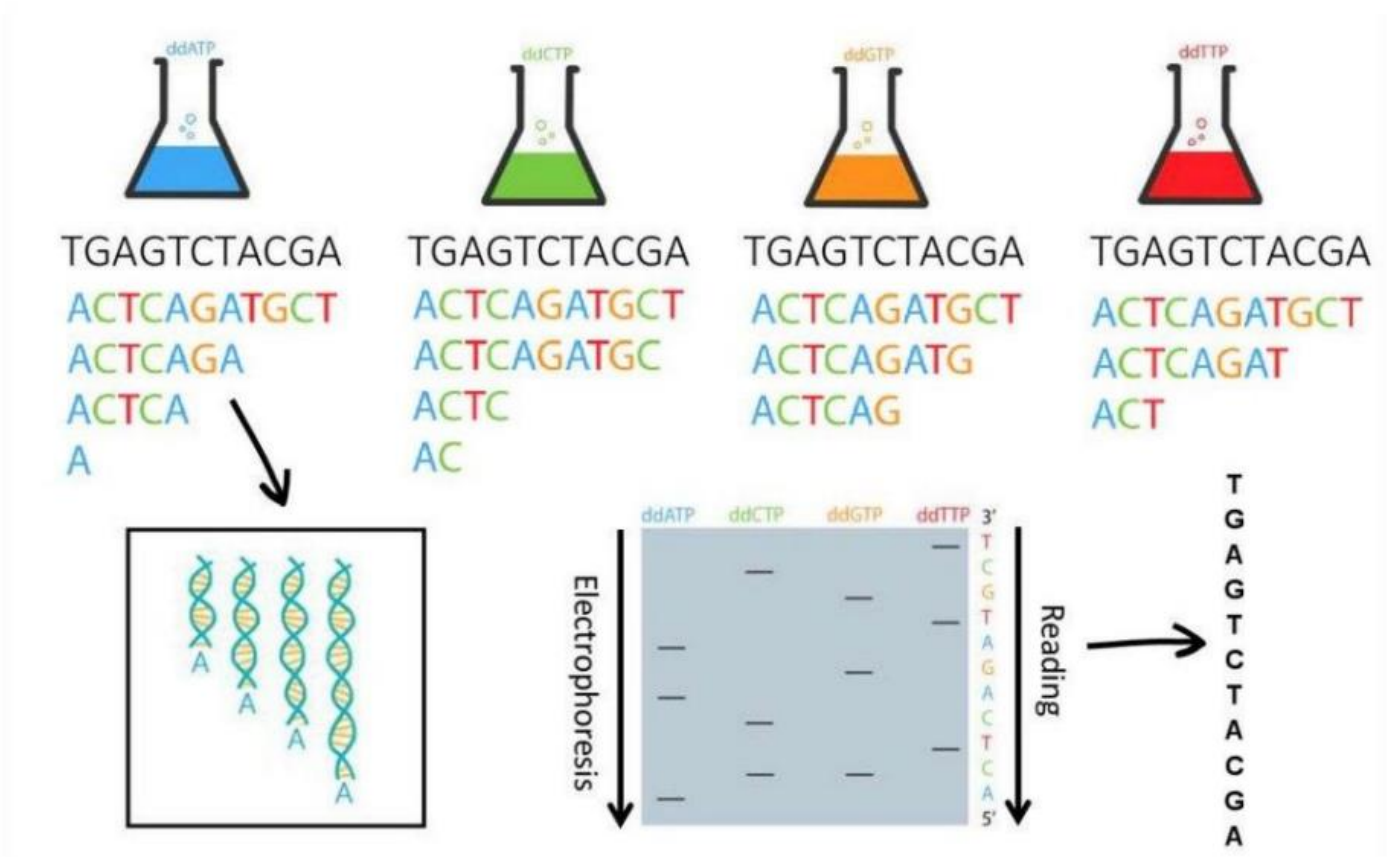


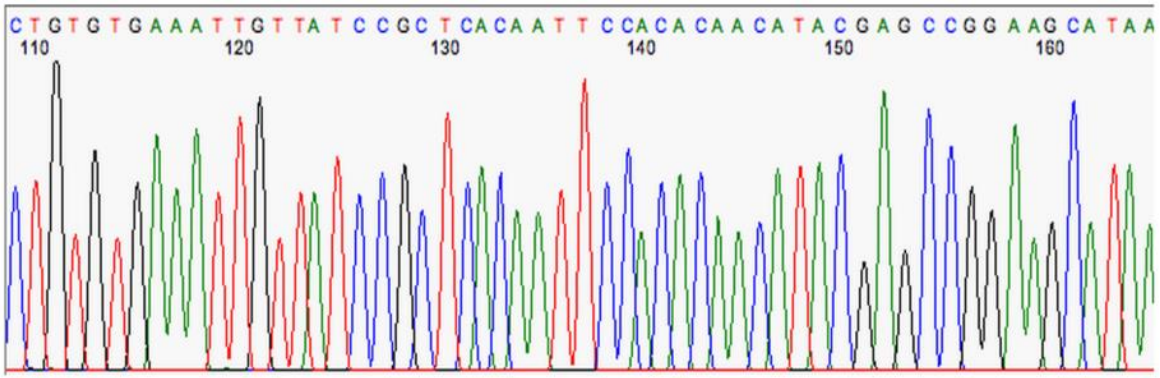
СХЕМА ПРОЦЕССА СЕКВЕНИРОВАНИЯ ПО СЭНГЕРУ



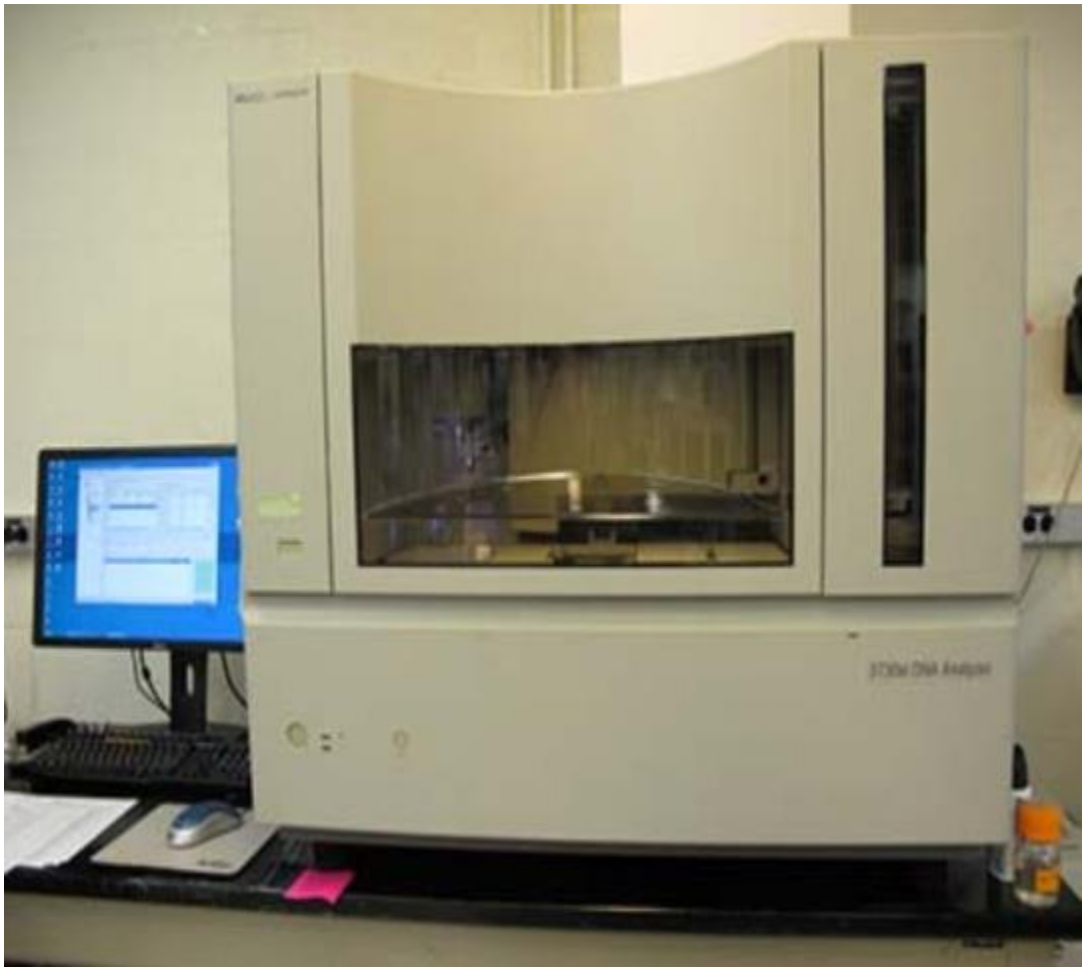
https://www.youtube.com/watch?v=FvHRio1yyhQ

СОВРЕМЕННАЯ МЕТОДИКА

- Сейчас вместо того, чтобы проводить форез в четырёх пробирках с разными ddNTP, проводят один форез на смеси, которая содержала все четыре типа ddNTP, каждый из которых окрашен по-своему
- Форез проводится в капиллярах (а не в геле) — так у него получается больше разрешение
- Максимальная длина прочтения — не больше 1000-1200 нуклеотидов
- Почему до сих пор применяется, если есть более производительные методы секвенирования?



СОВРЕМЕННЫЙ КАПИЛЛЯРНЫЙ СЕКВЕНАТОР

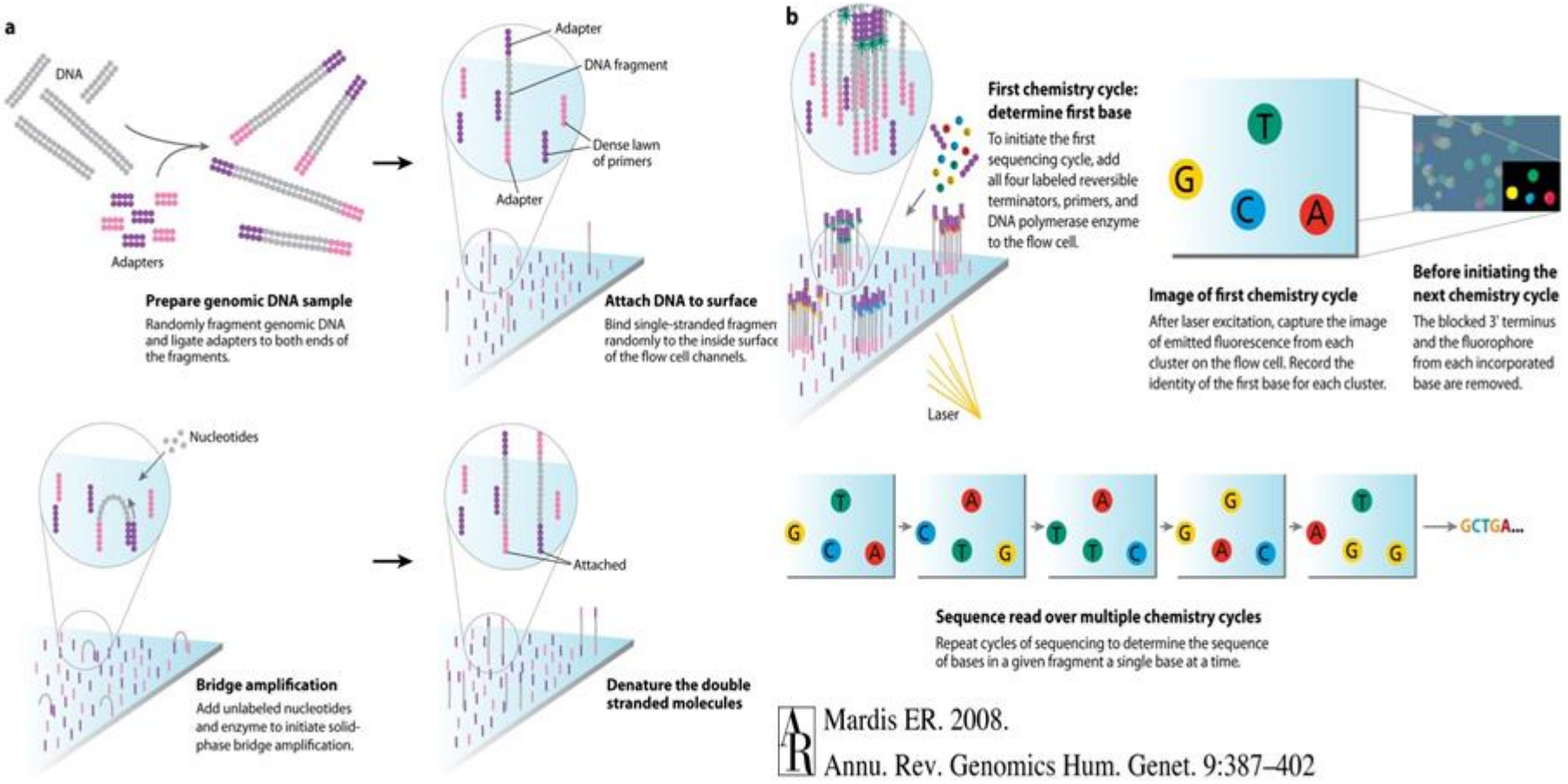


Много капилляров – секвенируем параллельно несколько фрагментов ДНК

NGS

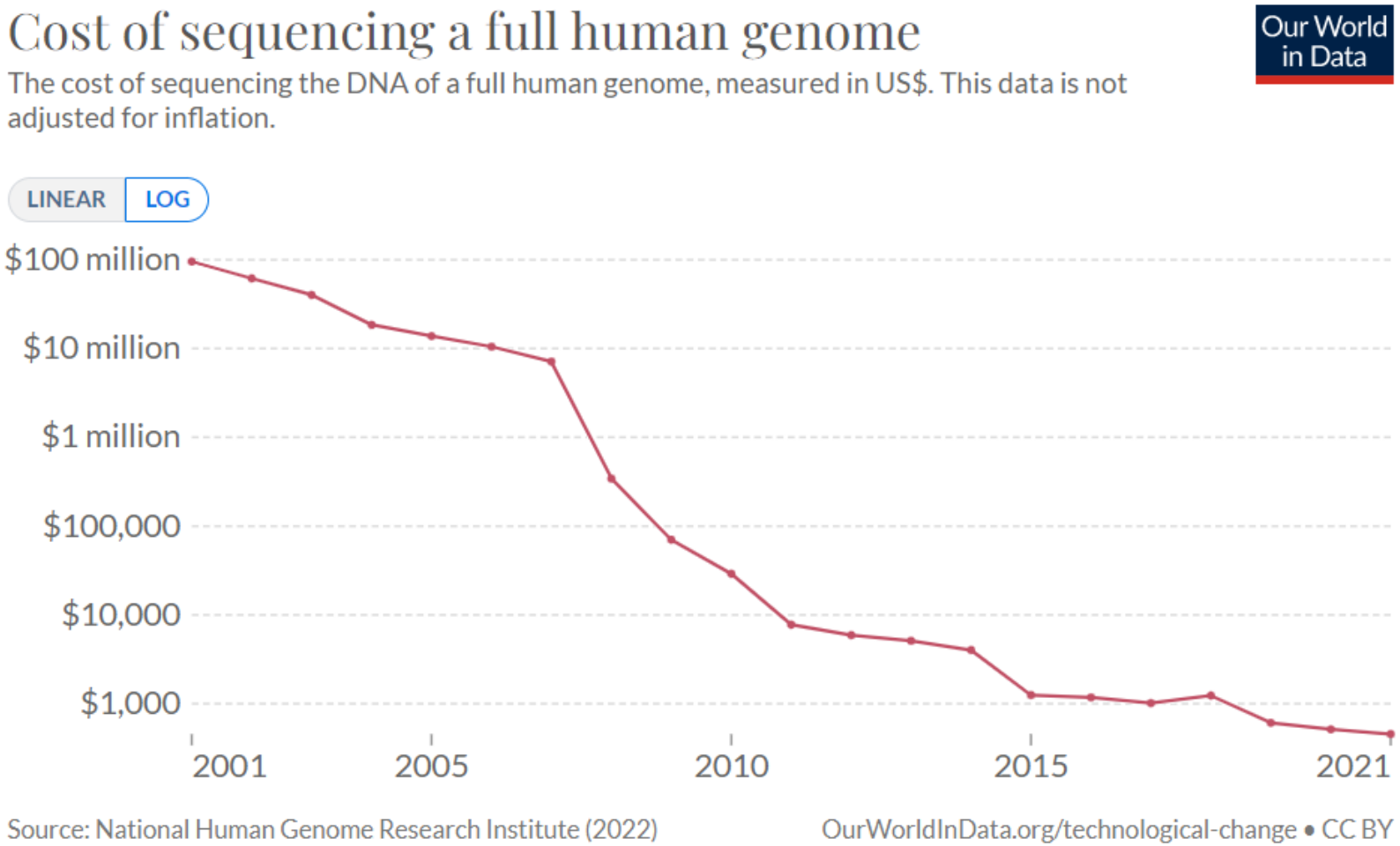
- **Next generation sequencing**
- “+” - одновременно идет сиквенс большого количества разных фрагментов
- “-” - прочтения длиной 75 - 150 нукл

NGS - ILLUMINA



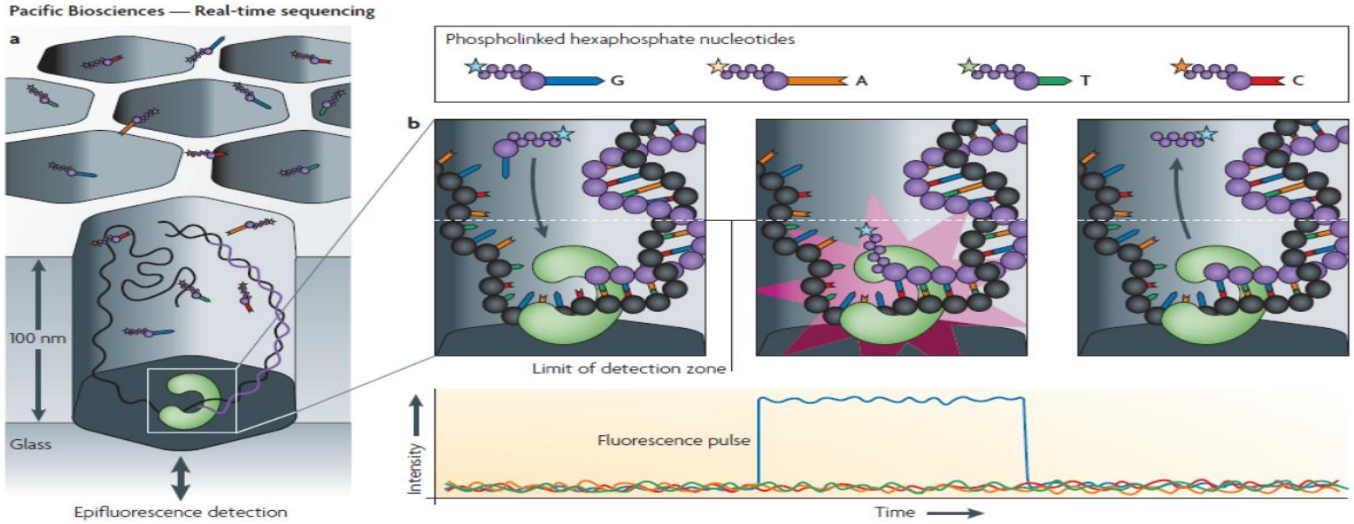
Mardis ER. 2008.
Annu. Rev. Genomics Hum. Genet. 9:387–402

СКОЛЬКО СТОИТ?



ОДНОМОЛЕКУЛЯРНОЕ СЕКВЕНИРОВАНИЕ

- ДНК-полимераза достраивает вторую цепь молекулы ДНК
- нуклеотиды, меченные различными флуоресцентными метками
- регистрация сигналов с помощью конфокальной микроскопии высокого разрешения



Pacific Biosciences

PACIFIC BIOSCIENCE

- “плюсы”
  - длина прочтений 20000-60000
  - без амплификации
  - быстро
- “минусы”
  - большой процент ошибок
  - цена

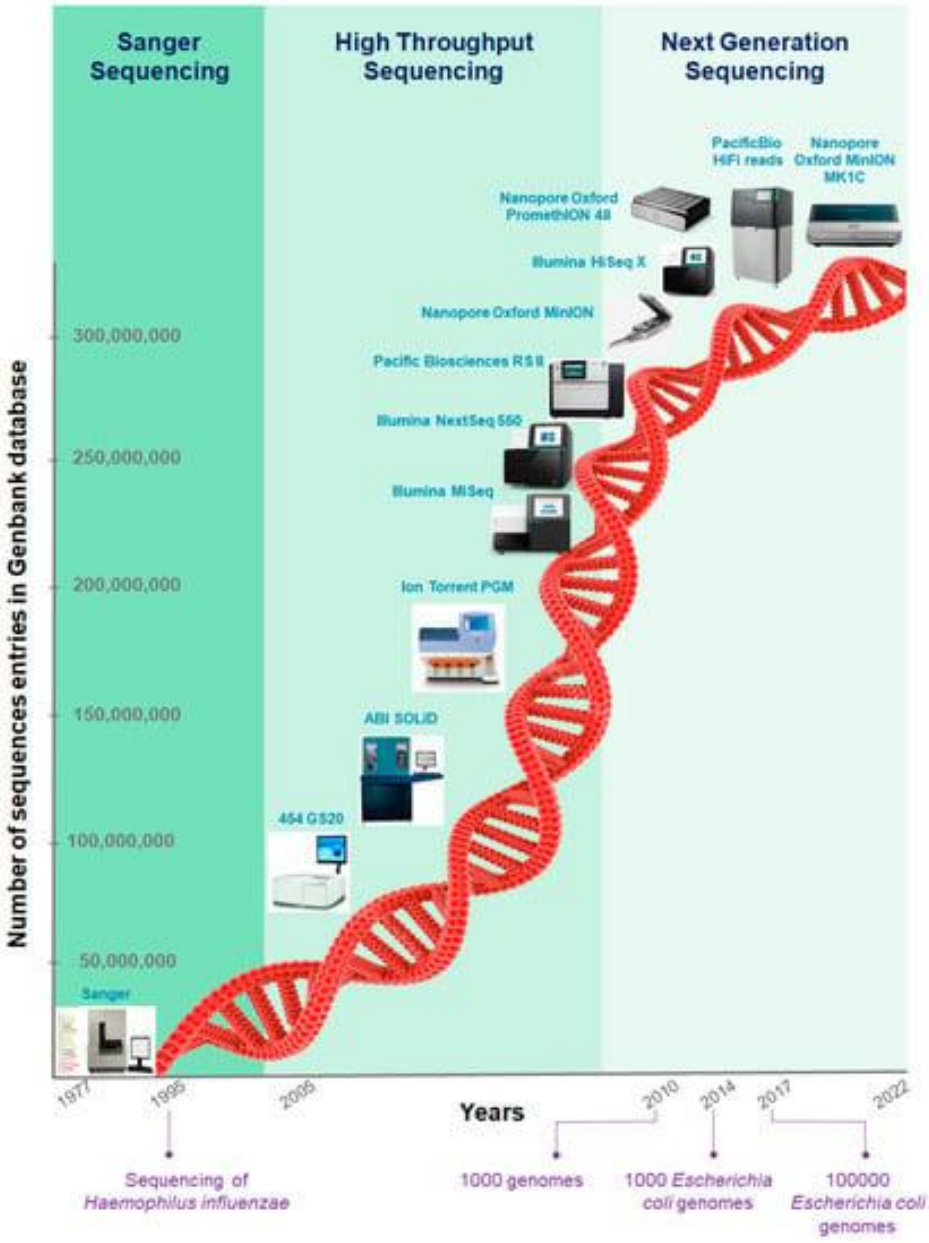
ОДНОМОЛЕКУЛЯРНОЕ СЕКВЕНИРОВАНИЕ

- Oxford Nanopore

- “плюсы”
  - длина прочтений 20000-60000
  - без амплификации
  - быстро
  - компактность и мобильность
- “минусы”
  - большой процент ошибок

ТЕХНОЛОГИИ СЕКВЕНИРОВАНИЯ

Sequencing Technology	Primary Errors	Error Rate
Capillary Sanger Sequencing	substitutions	0.1 %
454 FLX	Indel	1 %
PacBio SMRT	CG deletions	13 %
Oxford Nanopore	deletions	7 %
Ion Torrent	indels	1 %
Illumina	substitutions	0.1 %



ref

ЧТО ЖЕ ВЫБРАТЬ?

- Все зависит от задачи
- Комбинировать платформы
- Увеличивать покрытие

- T2T так и сделали!


СПАСИБУШКИ