

В прошлый раз

Наивный алгоритм

Кнут – Моррис – Пратт

Рабин – Карп

Бойер – Мурр

Неточный поиск с заменами

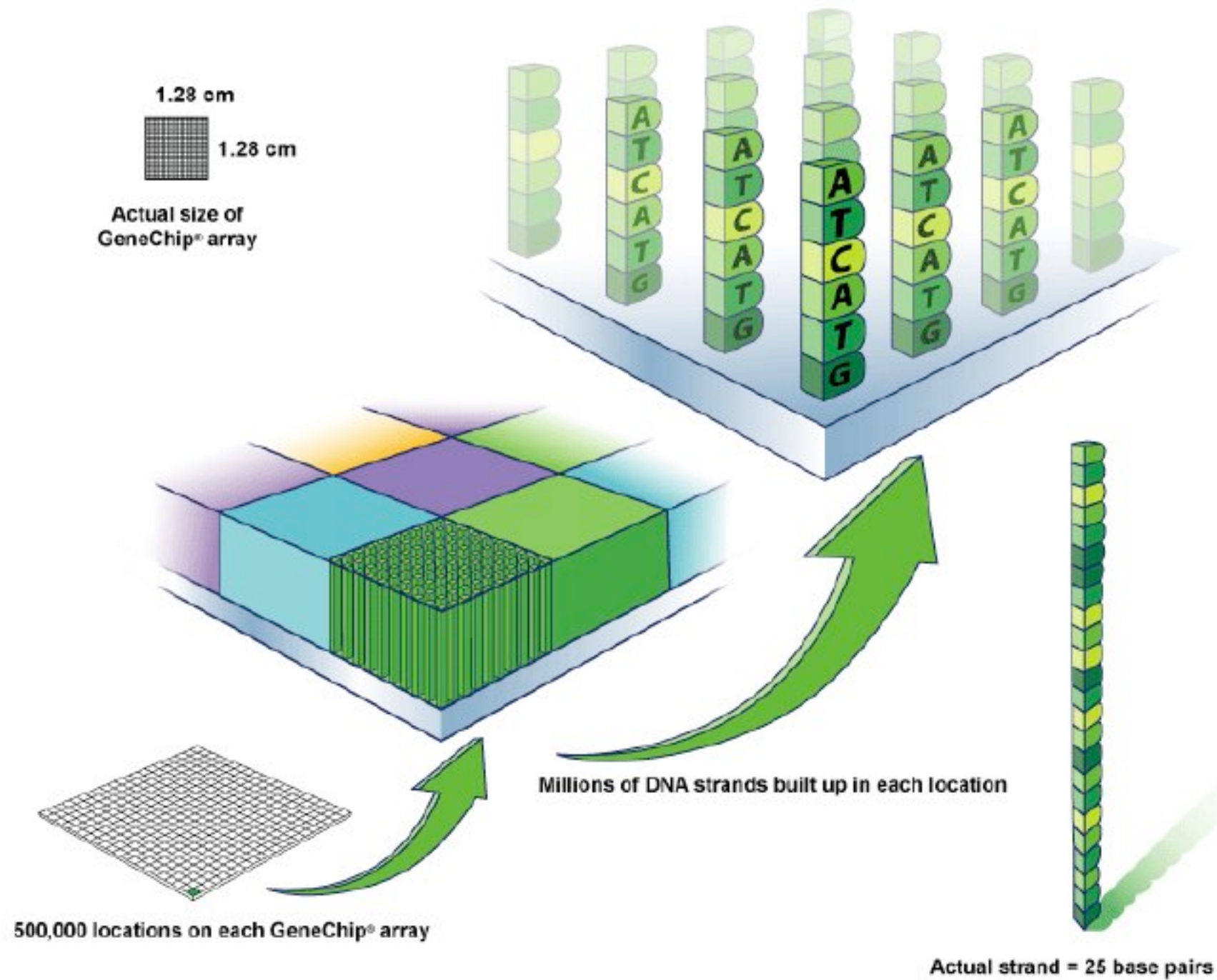
Сегодня

DNA Sequencing

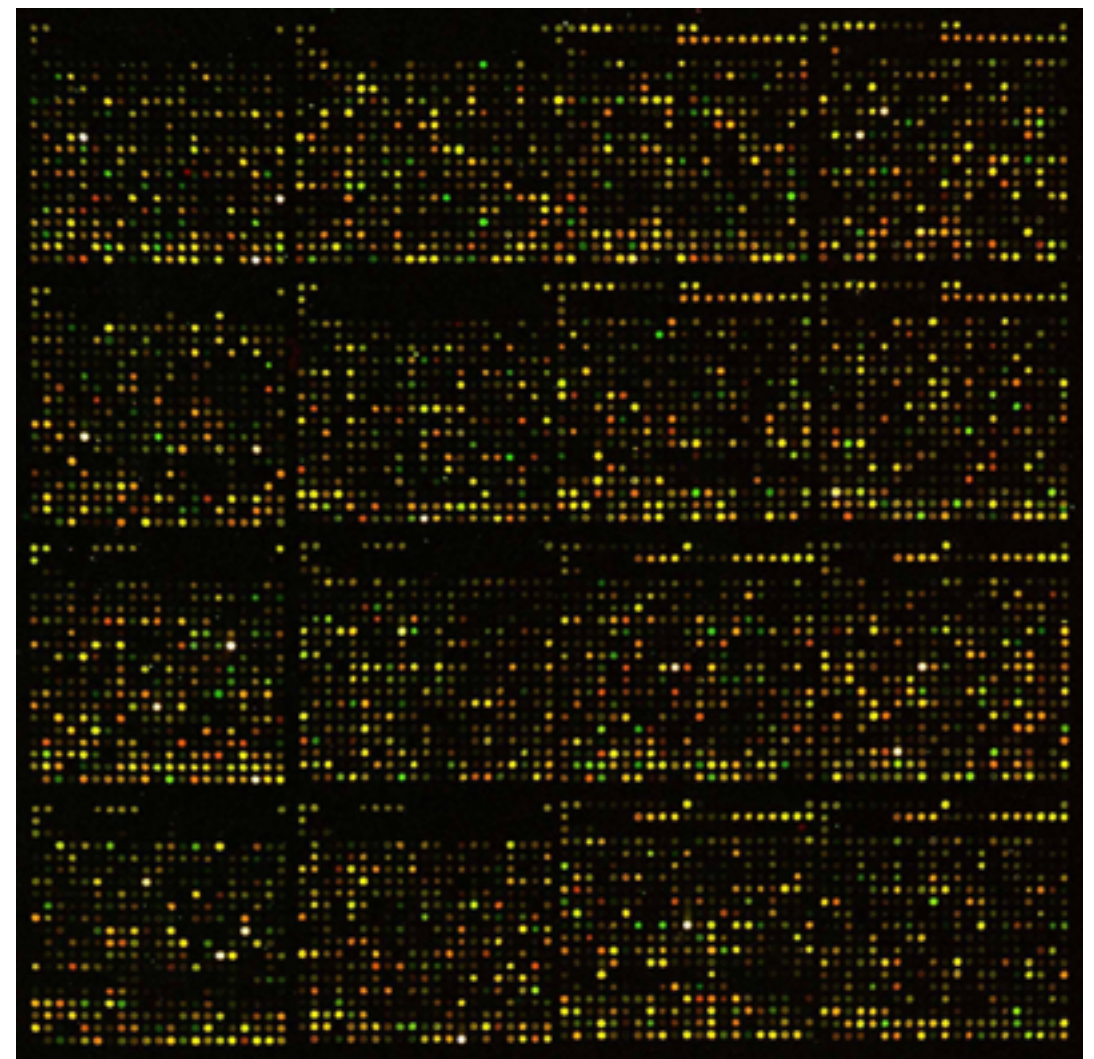
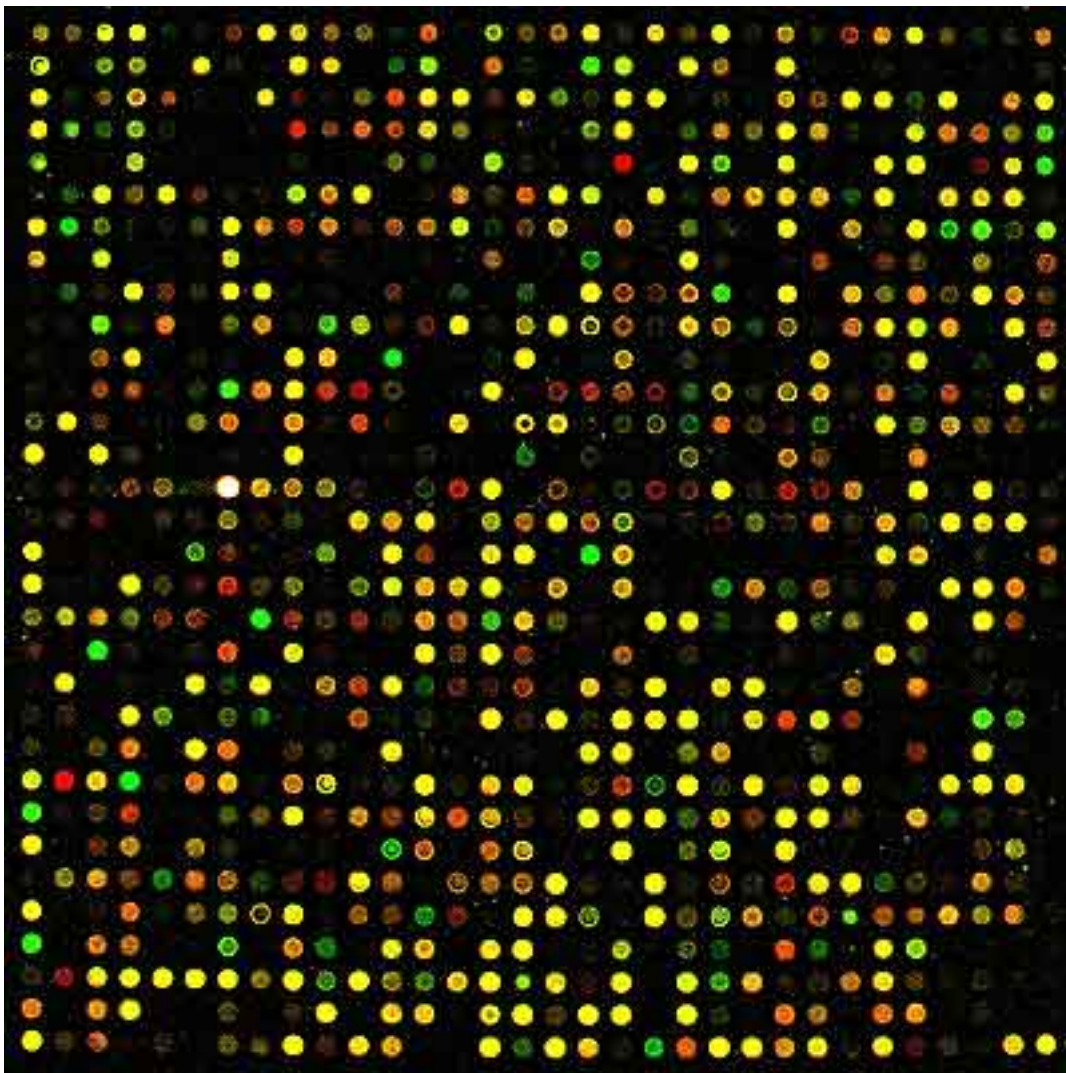
Бор

Ахо – Корасик

Микрочипы



Микрочипы

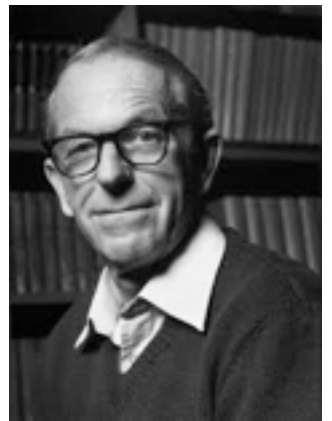


Краткая история

Конец **1970**-х: Уолтер Гилберт и Фредерик Сэнгер развивают независимые методы секвенирования.

1980: Они получают Нобелевскую премию по химии.

Их методы выявления последовательности слишком дороги для больших геномов.

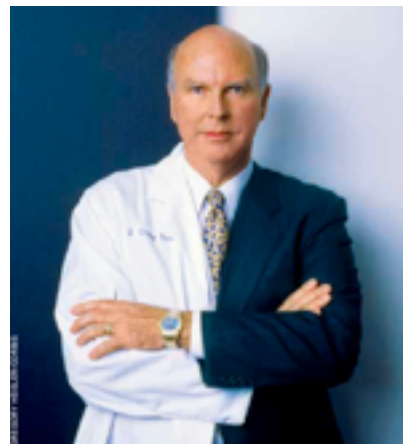


Краткая история

1990: Общественный проект «Человеческий геном», возглавляемый Фрэнсисом Коллинзом, задаётся целью расшифровать человеческий геном.

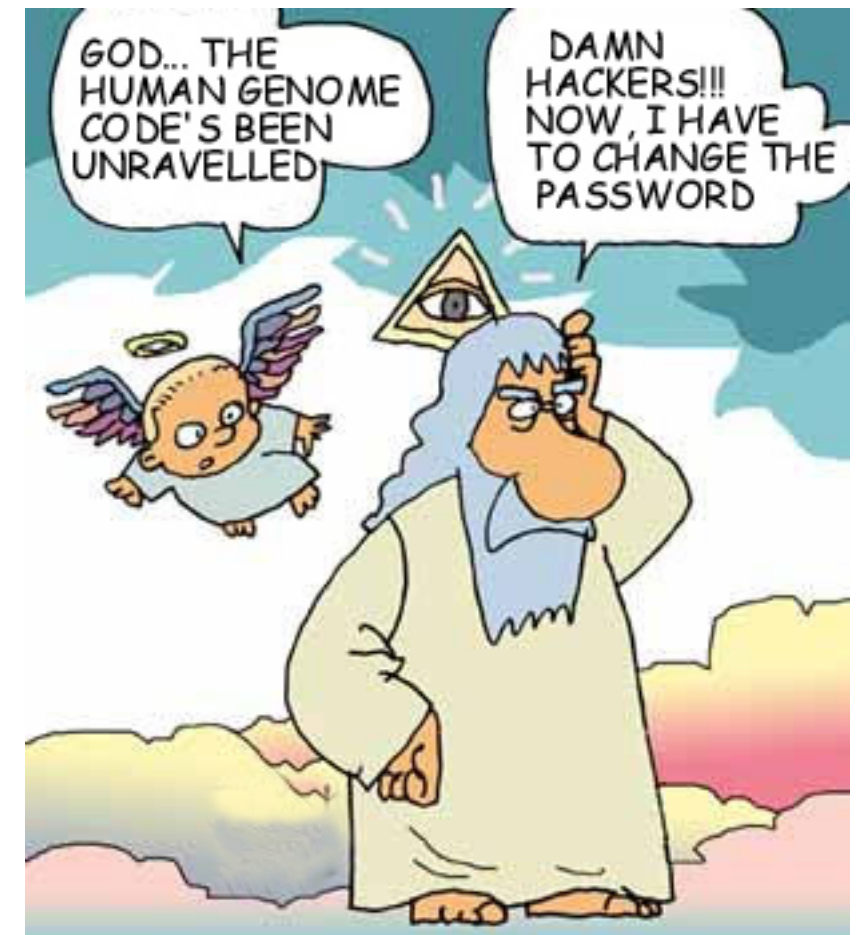


1997: Крейг Вентер создаёт частную компанию «Celera Genomics» с той же целью.



Краткая история

2000: Черновой вариант человеческого генома одновременно завершён (общественным) проектом «Человеческий геном» и (частной) компанией Celera Genomics.



Science

16 February 2001

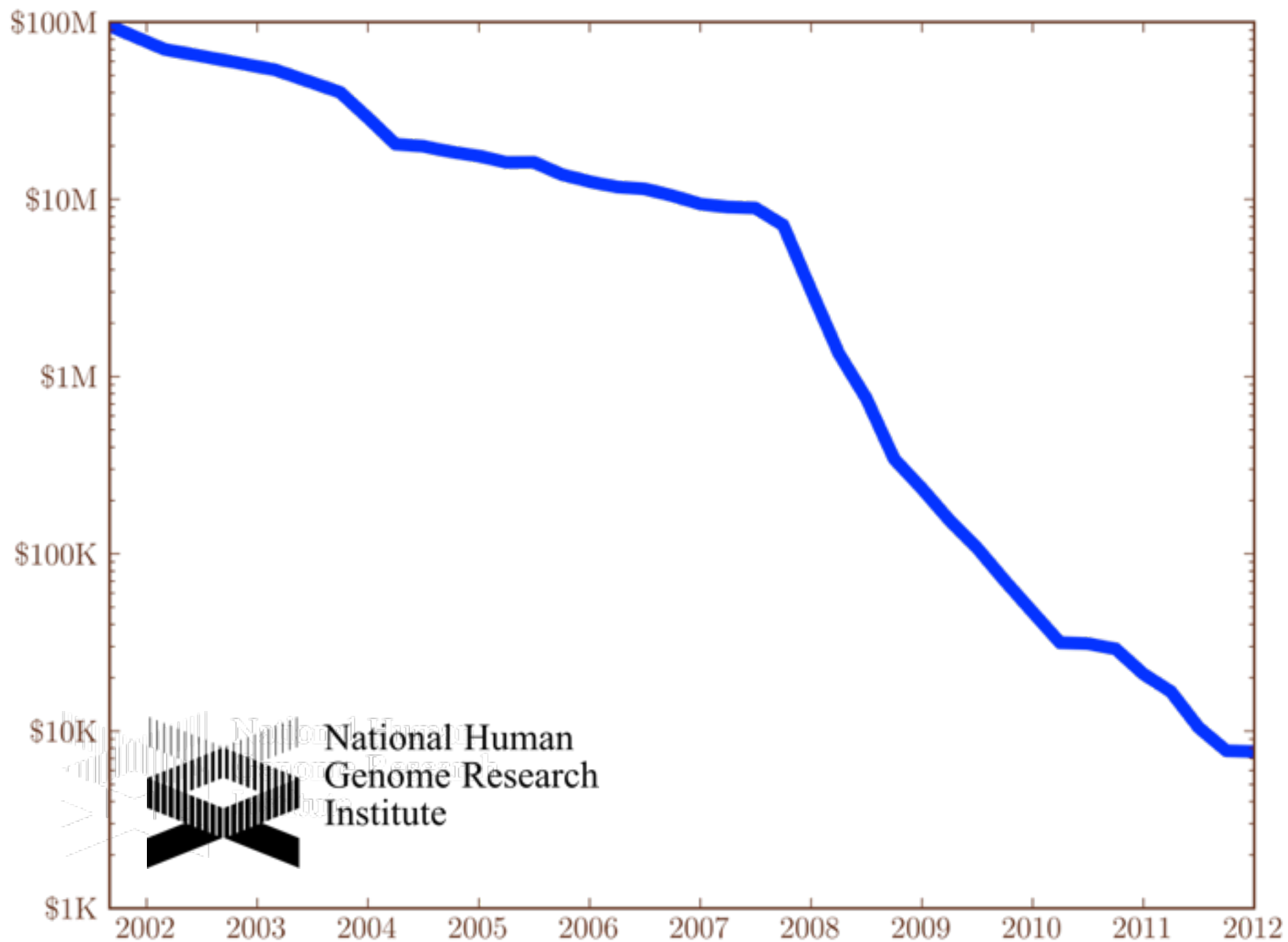
Vol. 291 No. 5507
Pages 1145-1434 \$9

THE HUMAN GENOME



AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

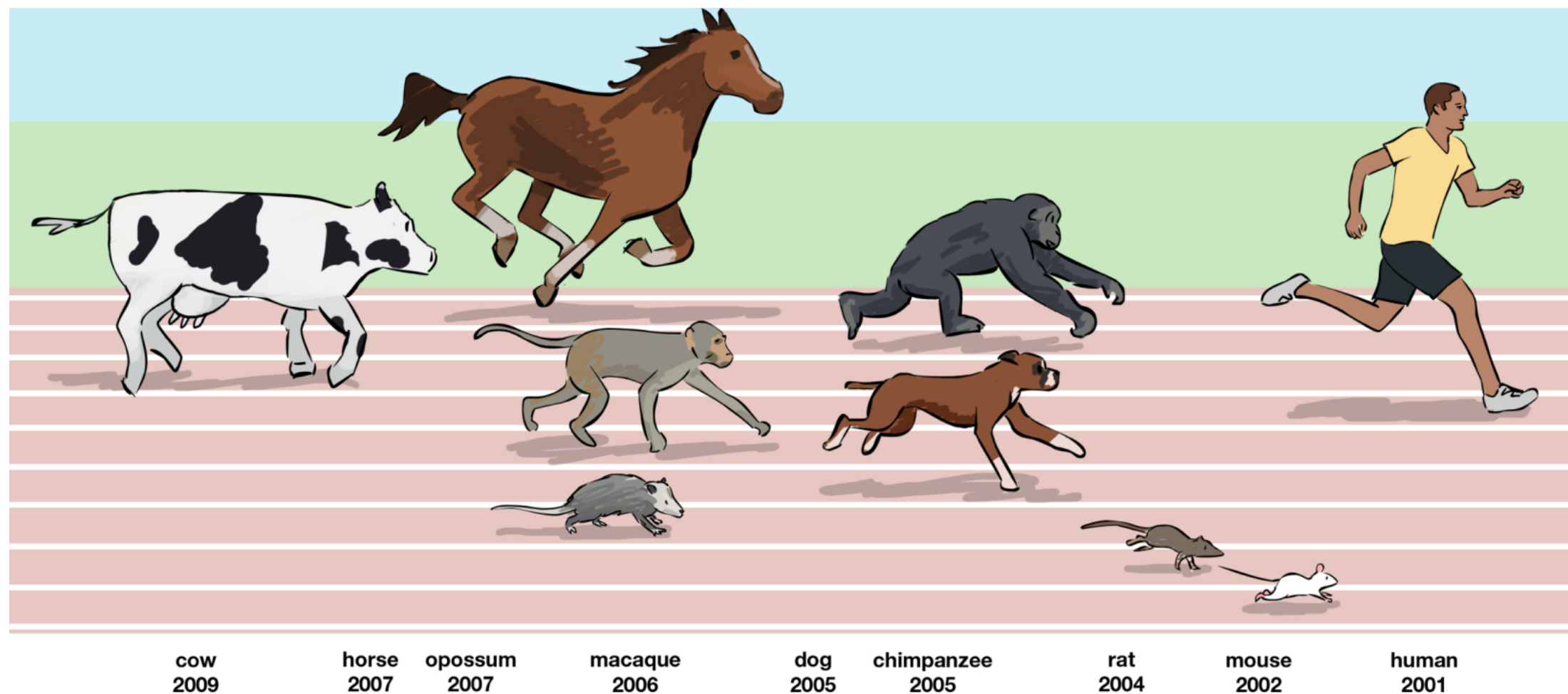
Эволюция технологий



ЭВОЛЮЦИЯ ТЕХНОЛОГИЙ

Год	Технология	Стоимость
2001	Sanger (ABI)	300 000 000 \$
2007	Sanger (ABI)	10 000 000 \$
2008	Roche (454)	2 000 000 \$
2008	Illumina	1 000 000 \$
2008	Illumina	250 000 \$
2009	Helicos	48 000 \$
2012	Illumina	5000 \$
2012	Complete Genomics	5000 \$

Эволюция технологий



Будущее

Секвенирование человеческого генома за 1000 долларов может стать реальностью уже в 2013-14 году.

Секвенирование индивидуального генома вскоре станет таким же рутинным делом, как рентгеновский снимок.



Эволюция технологий

Сегодня: 50 Gbp / день.

Нельзя прочитать ДНК целиком!

Только маленькими фрагментами:

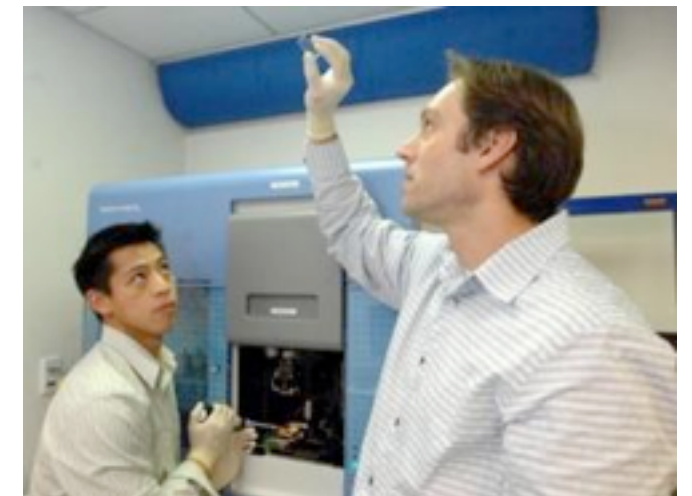
- 100 - 200 букв
- 1-2% ошибок чтения



Секвенирование

Чтение фрагментов (лабораторная):

Считать множество
фрагментов из многих копий
одного генома.

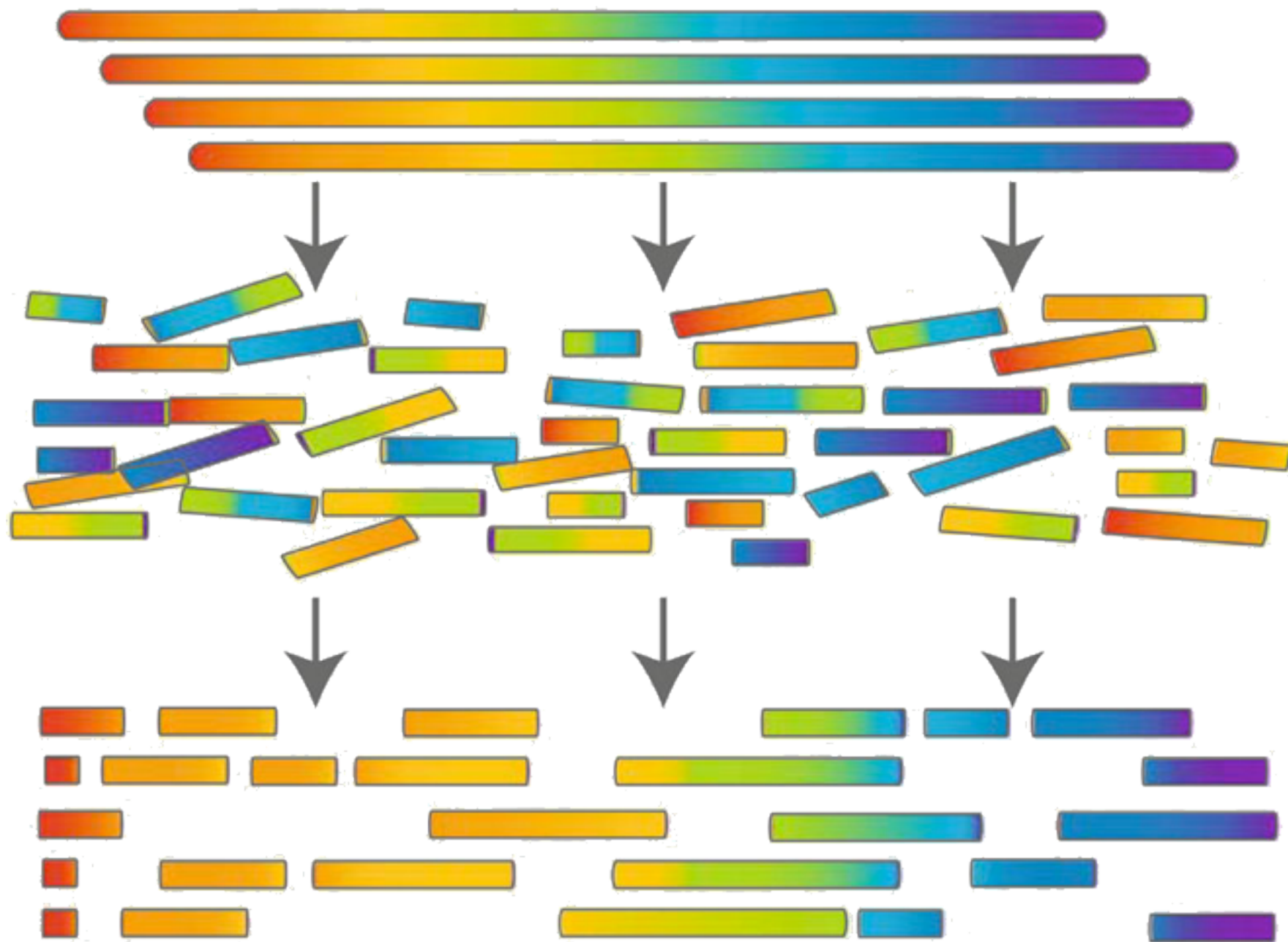


Сборка фрагментов (вычислительная):

Собрать геном из этих ридов
с помощью алгоритмов.

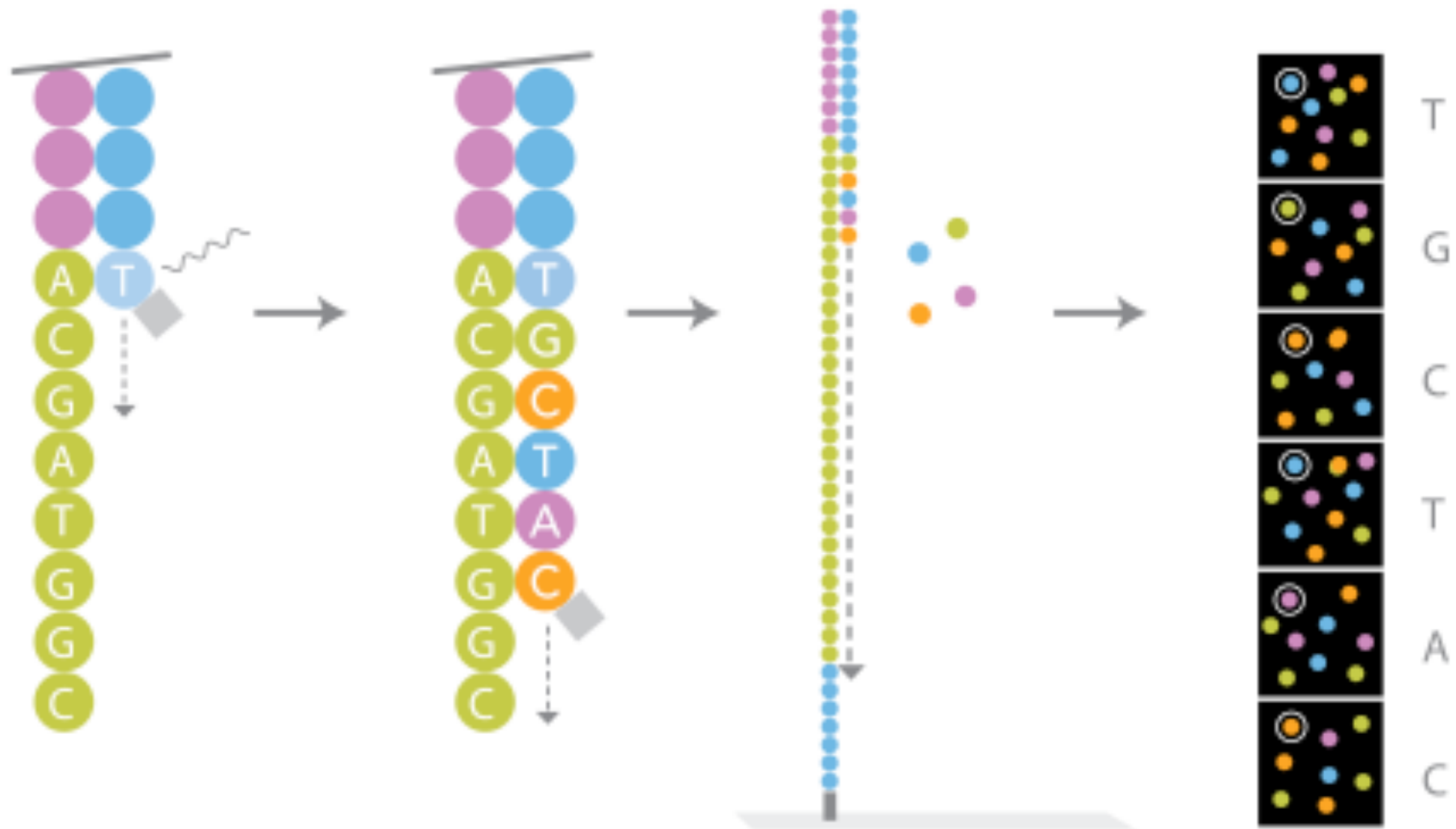


Секвенирование



ATGTTCCGATTAGGAAACCTATCTGTAAGTGTTCATTCAAGTAAAAGGAGGAAATATAA

Next Gen Sequencing



Short Read Alignment



Задача

Даны:

шаблоны P суммарной длины N ,
текст T длины M .

Найти все позиции вхождения P в T .

Бор

Бор

Хранит словарь строк.

Бор

Хранит словарь строк.

Бор

Хранит словарь строк.

Trie, префиксное дерево.

Бор

Хранит словарь строк.

Trie, префиксное дерево.

Бор

Хранит словарь строк.

Trie, префиксное дерево.

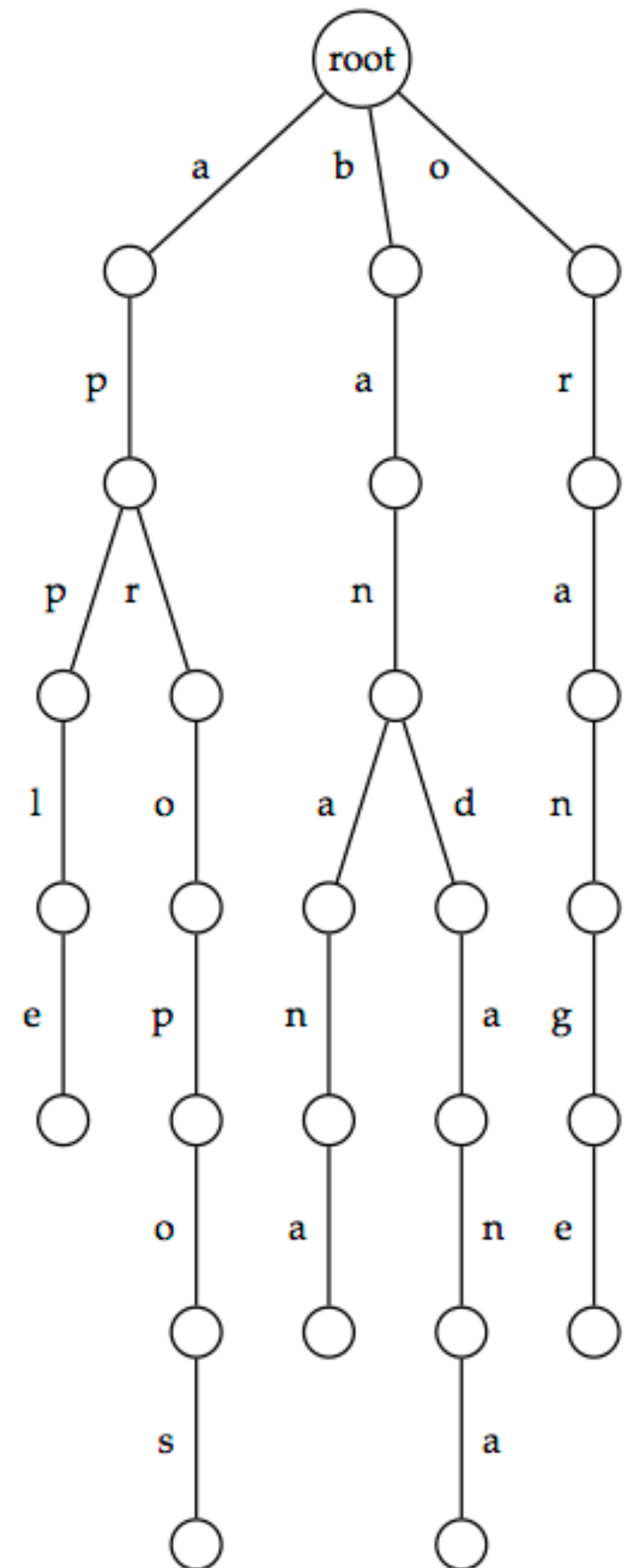
Пример: apple, apropos,
banana, bandana, orange.

Бор

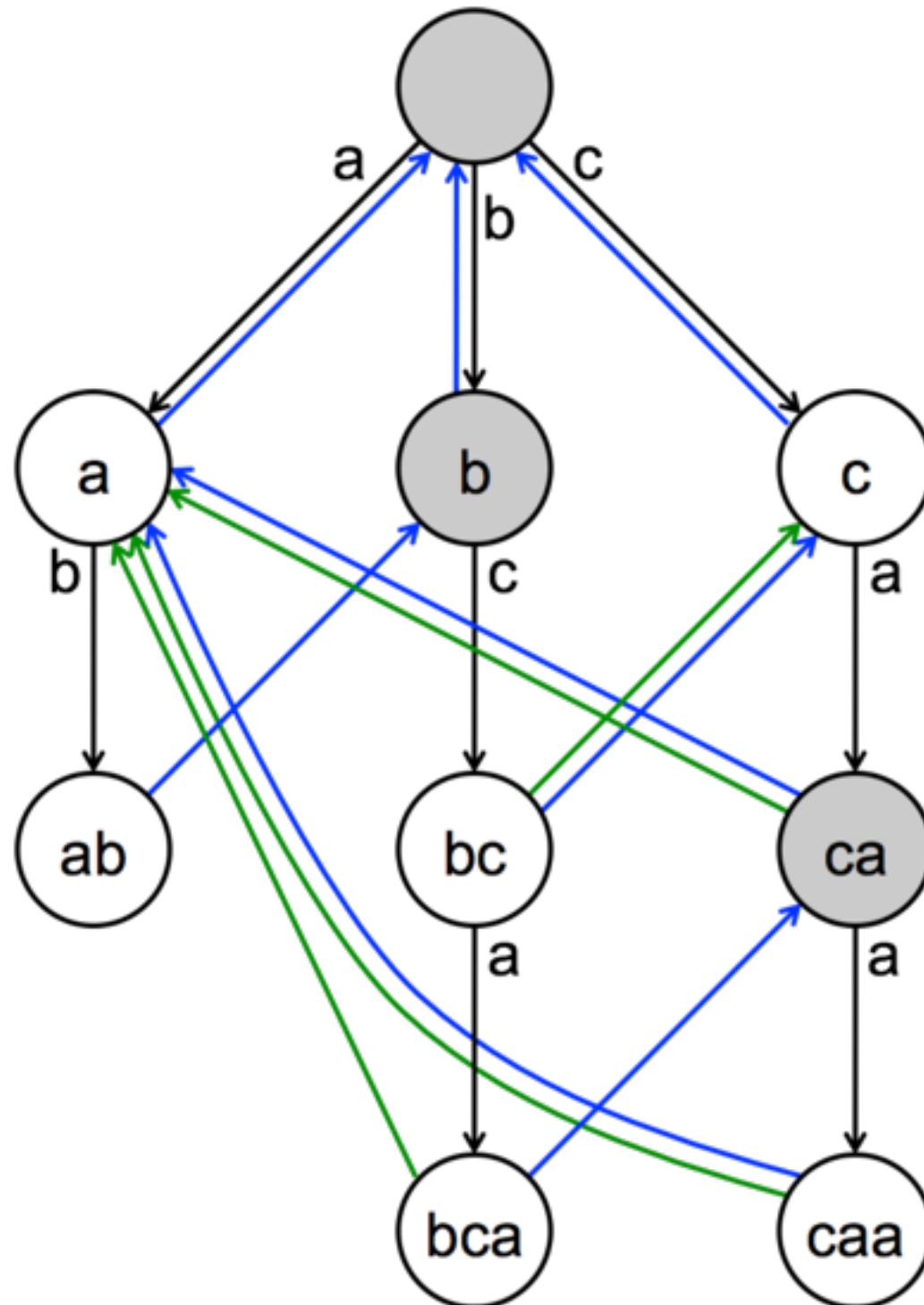
Хранит словарь строк.

Trie, префиксное дерево.

Пример: apple, apropos,
banana, bandana, orange.



Ахо – Корасик



http://e-maxx.ru/algo/aho_corasick

Новая задача

Даны:

шаблон P длины N ,

текст T длины M .

Можно заранее обработать T .

Найти все позиции вхождения P в T .

Что мы узнали

DNA Sequencing

Бор

Ахо – Корасик