Сборка генома *de novo*: мифы и реальность

Сергушичев А. А., Царев Ф. Н.

Практическая школа по биоинформатике

МНЛ «Компьютерные технологии»

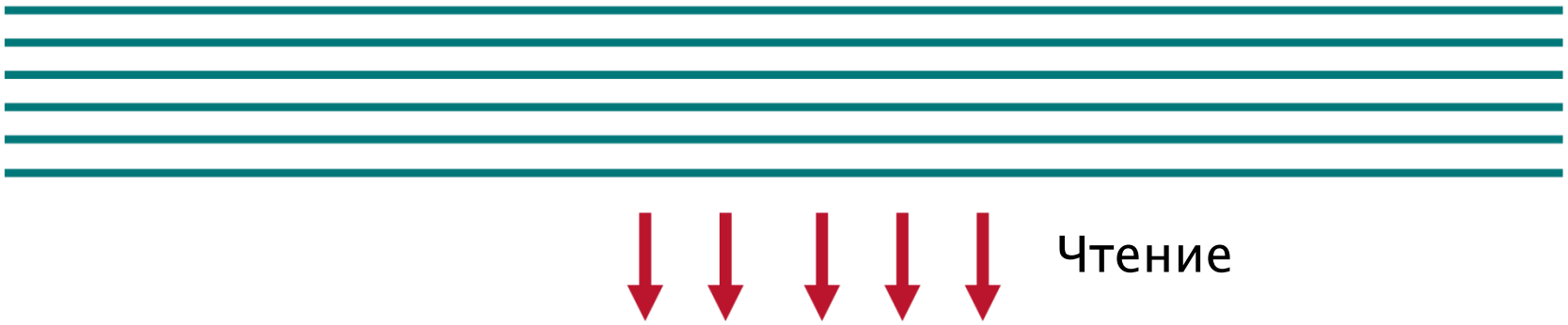
19.02.2014

Чтение и сборка генома

Несколько копий генома

Чтение и сборка генома

Несколько копий генома



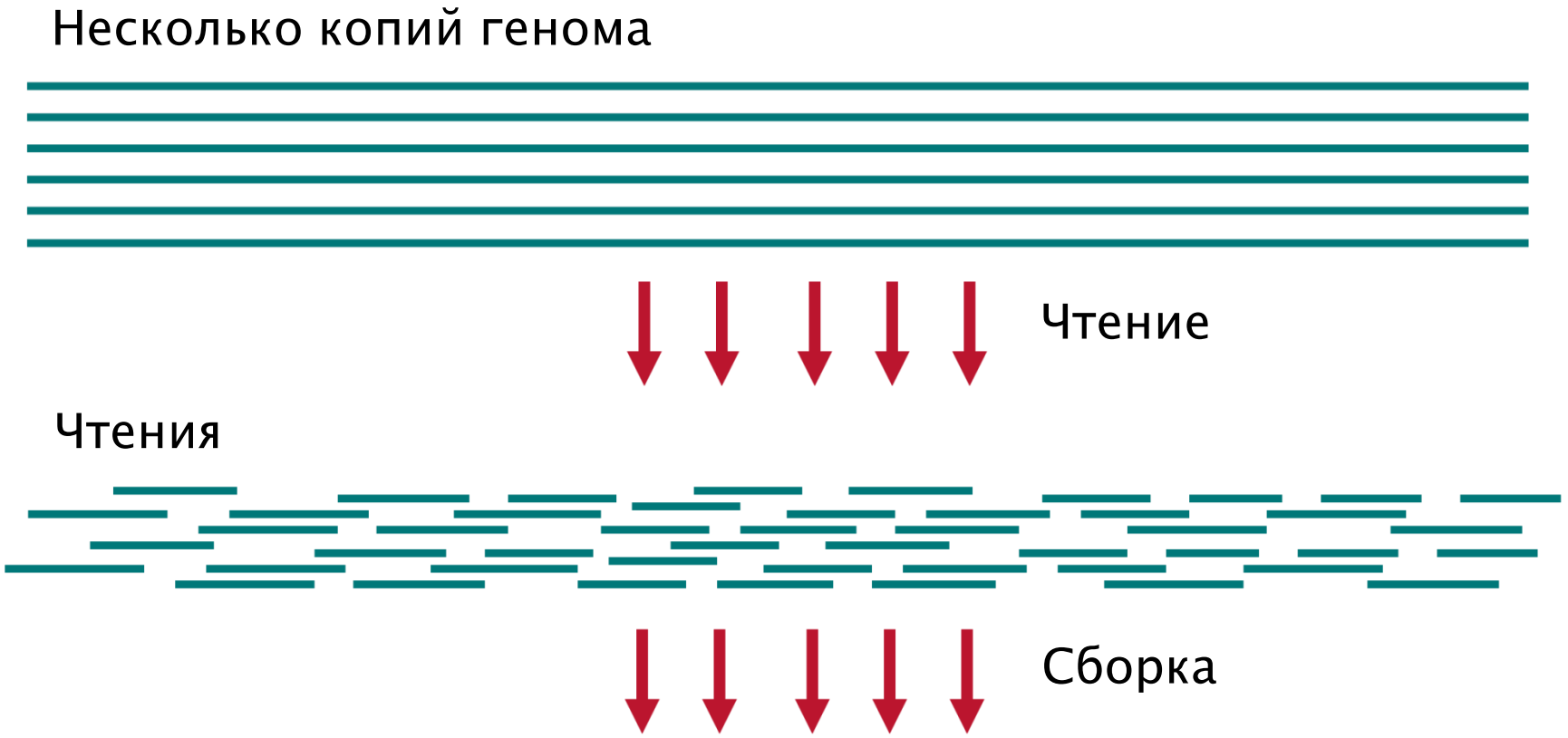
Чтение и сборка генома

Несколько копий генома

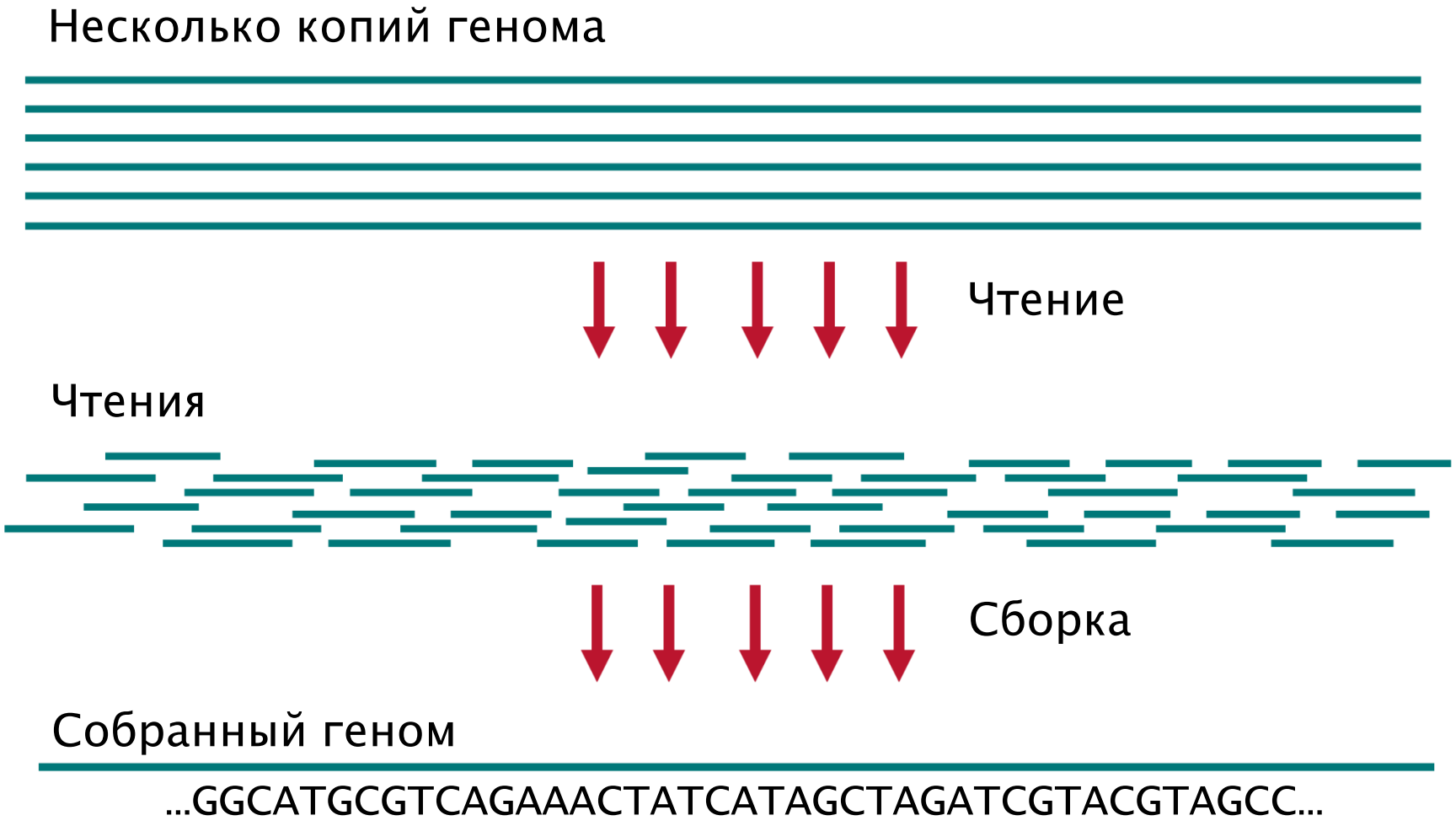
Чтения

Чтения

Чтение и сборка генома



Чтение и сборка генома

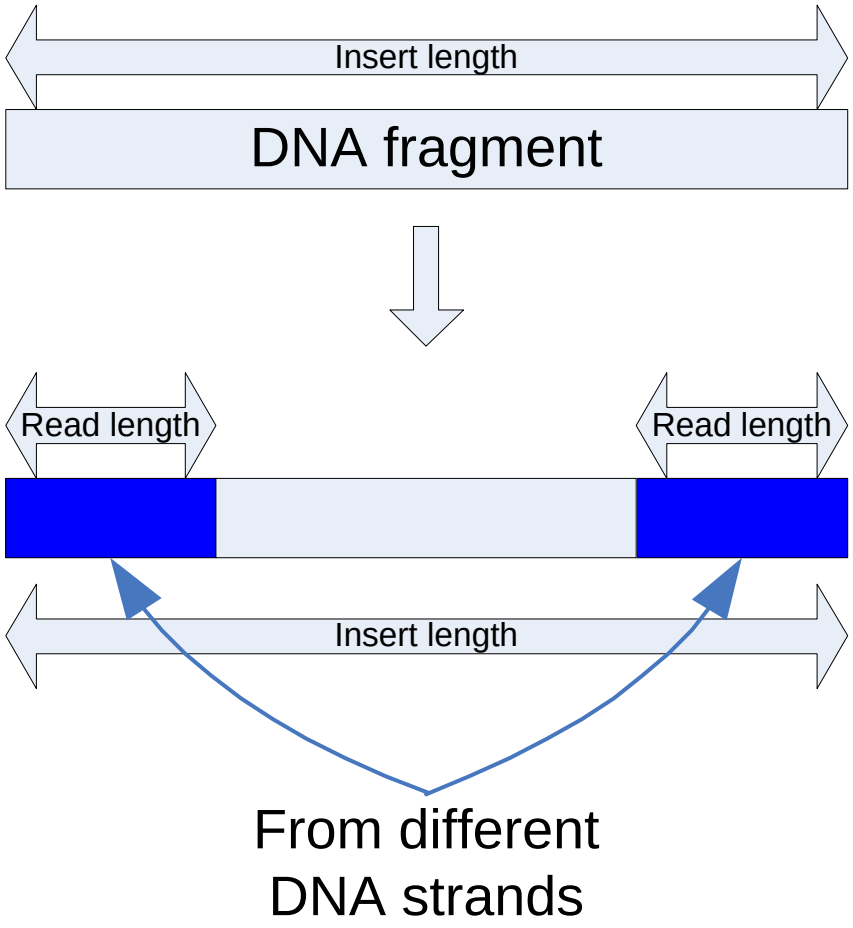


Секвенирование генома

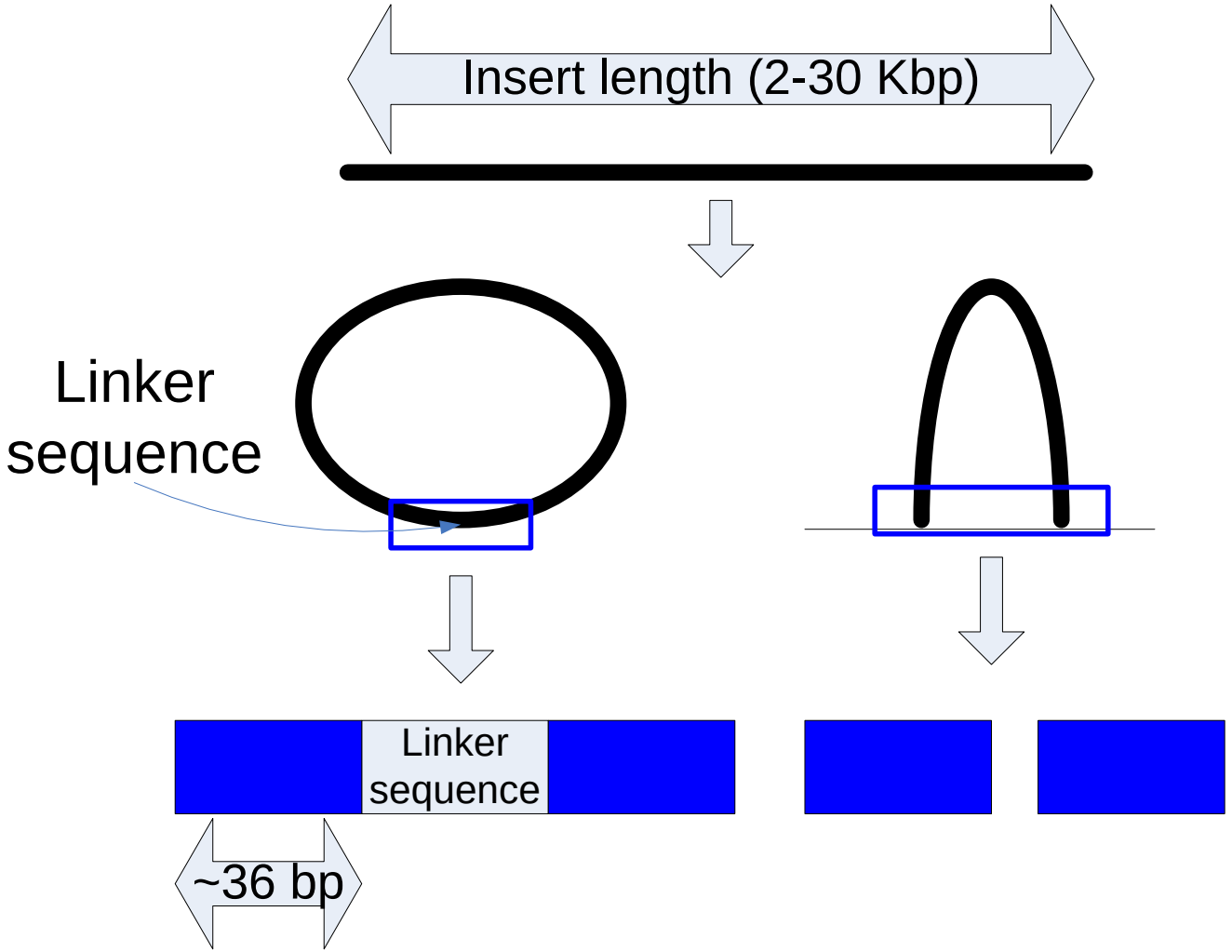
- Специальные устройства секвенаторы
  - Life Technologies
  - Illumina
  - Pacific Biosciences
  - Oxford Nanopore

Парные чтения (paired-end)

- Впервые использованы в секвенаторах Illumina
- Копии генома разрезаются на фрагменты
- Длина фрагмента: 200-500
- Длина чтения: 36-150



Mate-pair чтения



Ошибки в чтениях (1)

- Три типа ошибок:
  - Замена (вместо AGCATA прочитано AGCCTA)
  - Вставка (вместо AGCATA прочитано AGCGATA)
  - Удаление (вместо AGCATA прочитано AGCTA)

Ошибки в чтениях (2)

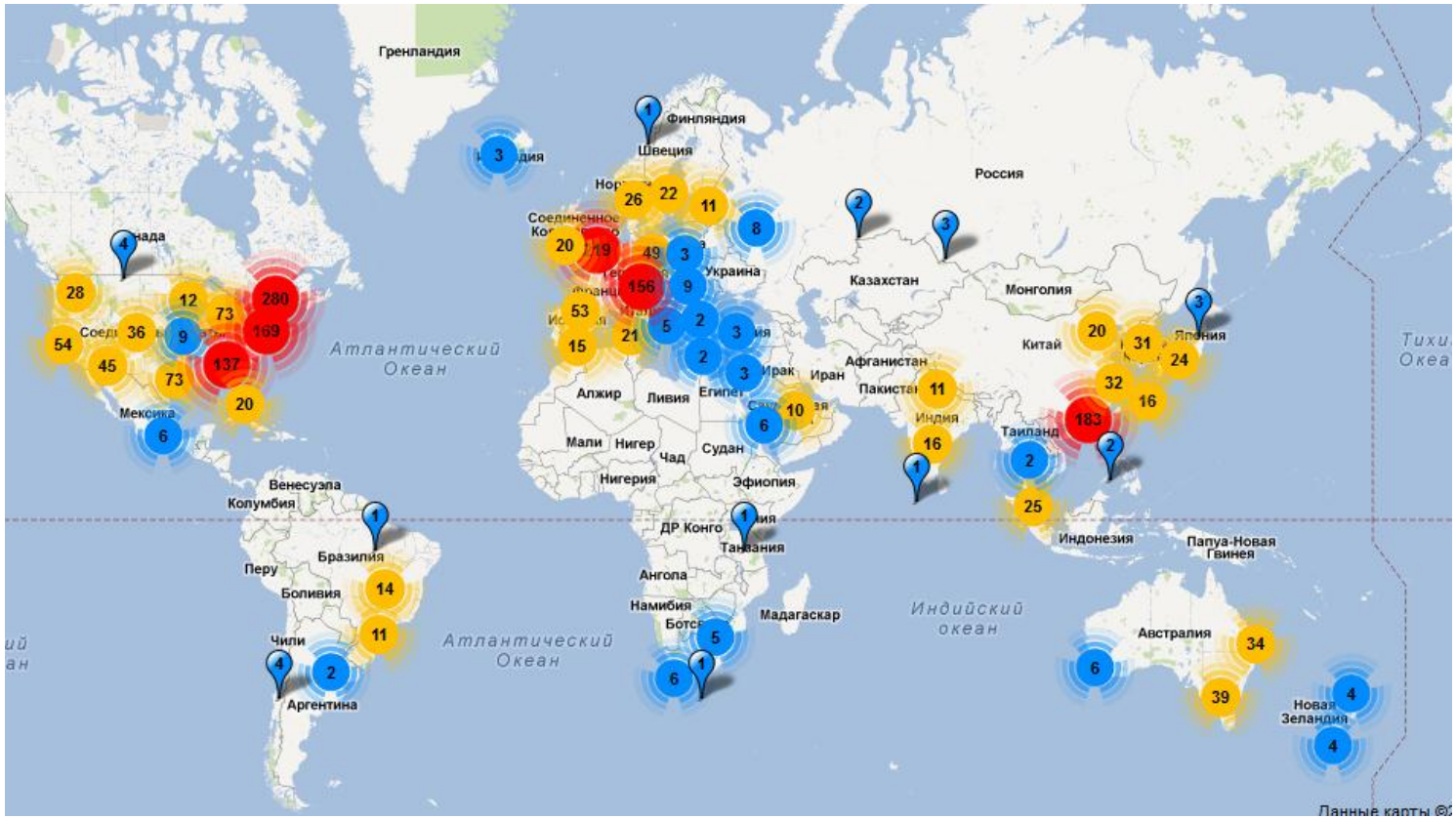
- У разных секвенаторов – разные типы ошибок
- У разных секвенаторов – разные вероятности ошибок (от 0.1% до 10%)

Ошибки в чтениях (3)

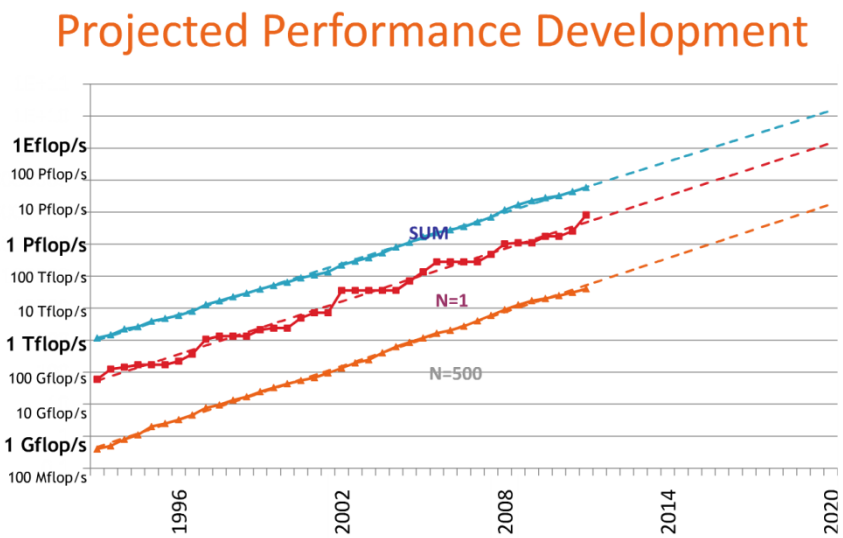
- Химерные чтения – части генома, находящиеся в нем далеко друг от друга, оказались в одном чтении
- Возникают при подготовке библиотек

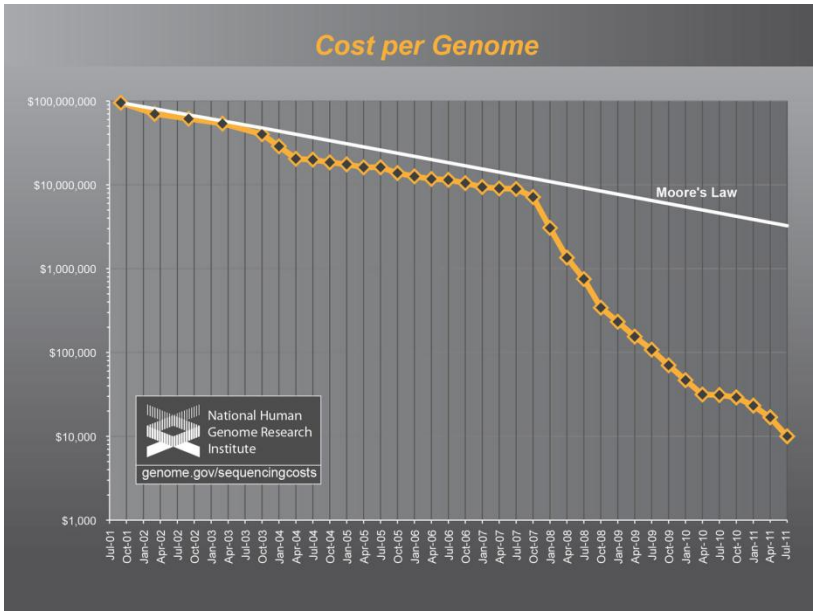
Карта секвенаторов

Объем данным ~15PB/год



Производительность секвенаторов растет быстрее, чем компьютеров



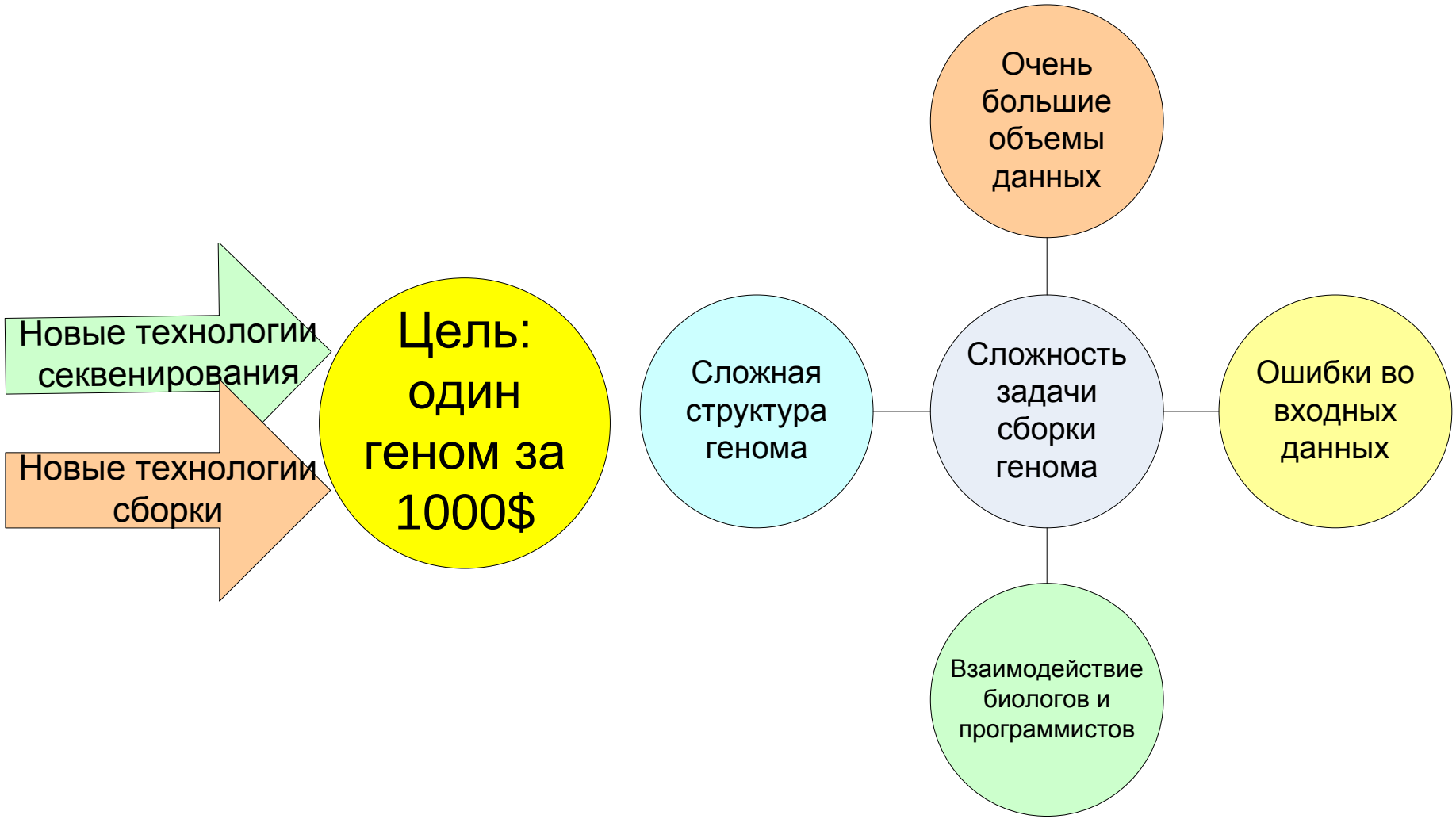


- Закон Мура – производительность компьютеров удваивается каждые 18 месяцев
- Стоимость секвенирования (за Mbp) уменьшается в 10 раз за то же время
- **Требуются новые алгоритмы обработки геномных данных!**

Размеры геномов

Тип	Организм	Размер генома		Комментарий
Virus	Bacteriophage MS2	3,569	3.5kb	Первый прочитанный РНК-геном
Virus	Phage Ф-X174	5,386	5.4kb	Первый прочитанный ДНК-геном
Bacterium	Escherichia coli	4,600,000	4.6Mb	
Plant	Arabidopsis thaliana	157,000,000	157Mb	
Mammal	Homo sapiens	3,200,000,000	3.2Gb	
Fish	Protopterus aethiopicus	130,000,000,000	130Gb	Самый большой известный геном позвоночного
Plant	Paris japonica	150,000,000,000	150Gb	Самый большой известный геном растения

Актуальность и сложность задачи

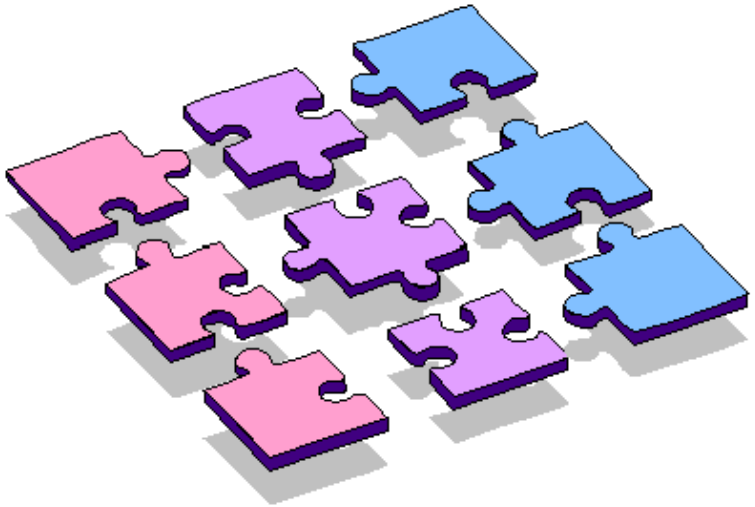


Биоинформатический стиль мышления

- Алгоритм – формальное описание последовательности действий, «рецепт»
- Структура данных – как представить данные в компьютере
- Анализ алгоритма – затраты по времени и по памяти
- Математические модели – формальное описание требований к результату

Задача сборки генома

- Исходные данные – набор чтений
- Результат – геномная последовательность
- Проблема – не знаем из какой части генома прочитано каждое чтение



Сборка генома *de novo*

- Входные данные:
  - Чтения последовательности ДНК
  - Часто – несколько библиотек с различными размерами фрагментов и длинами чтений
  - Типичное покрытие генома: 40x-100x
- Цель – получить как можно больше информации о геномной последовательности

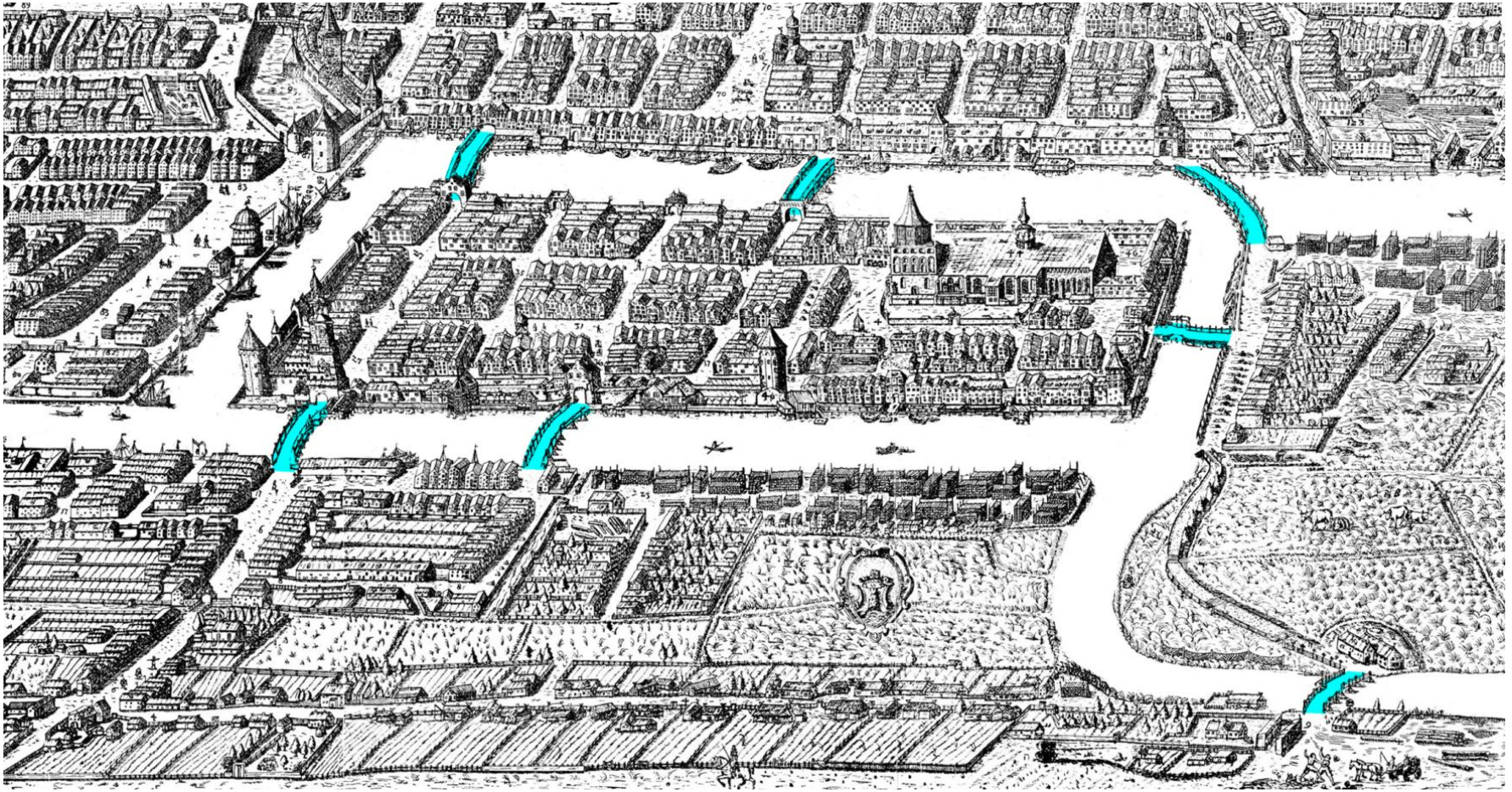
Мифы о сборке генома

- Миф №1. Собрать геном – это просто
- Миф №2. Есть понимание того, что значит «собрать геном»
- Миф №3. Существующие программы для сборки генома хорошо его собирают

Миф №1

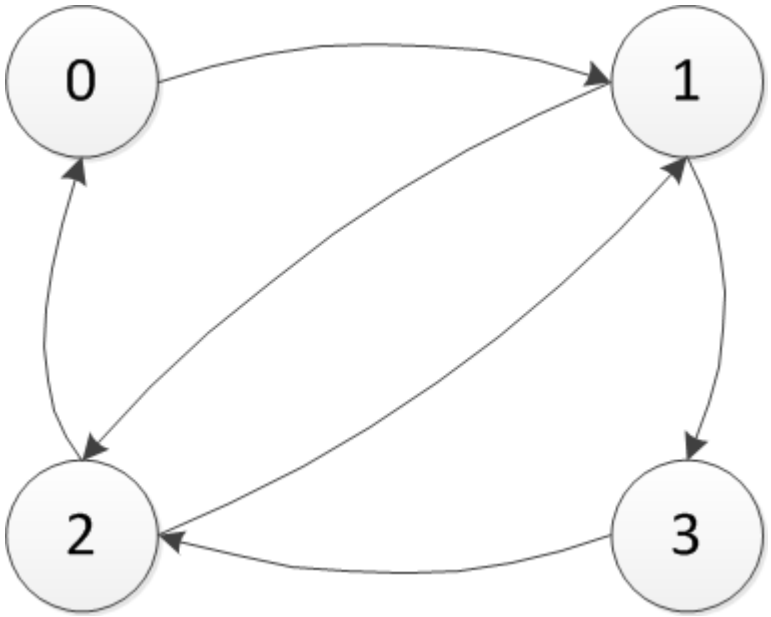
Собрать геном – это просто

Кенигсбергские мосты



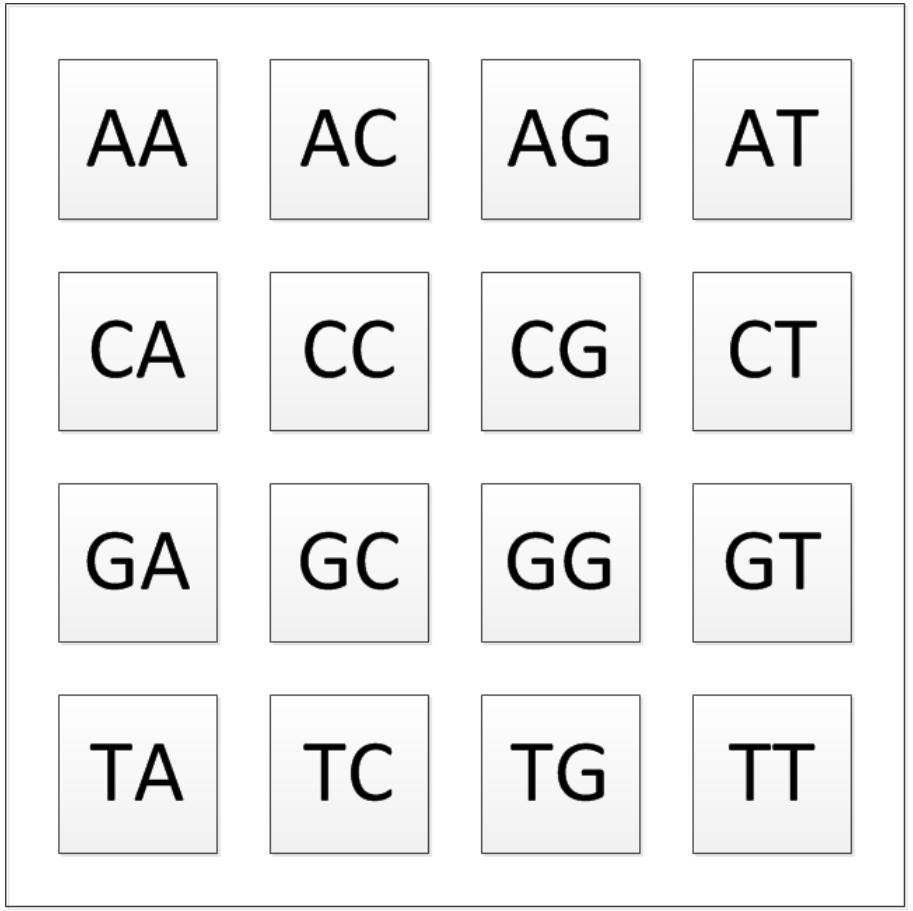
Эйлеров путь в графе

- Путь, который проходит по каждому ребру ровно один раз
- Существует способ быстро определить, есть ли в графе такой путь



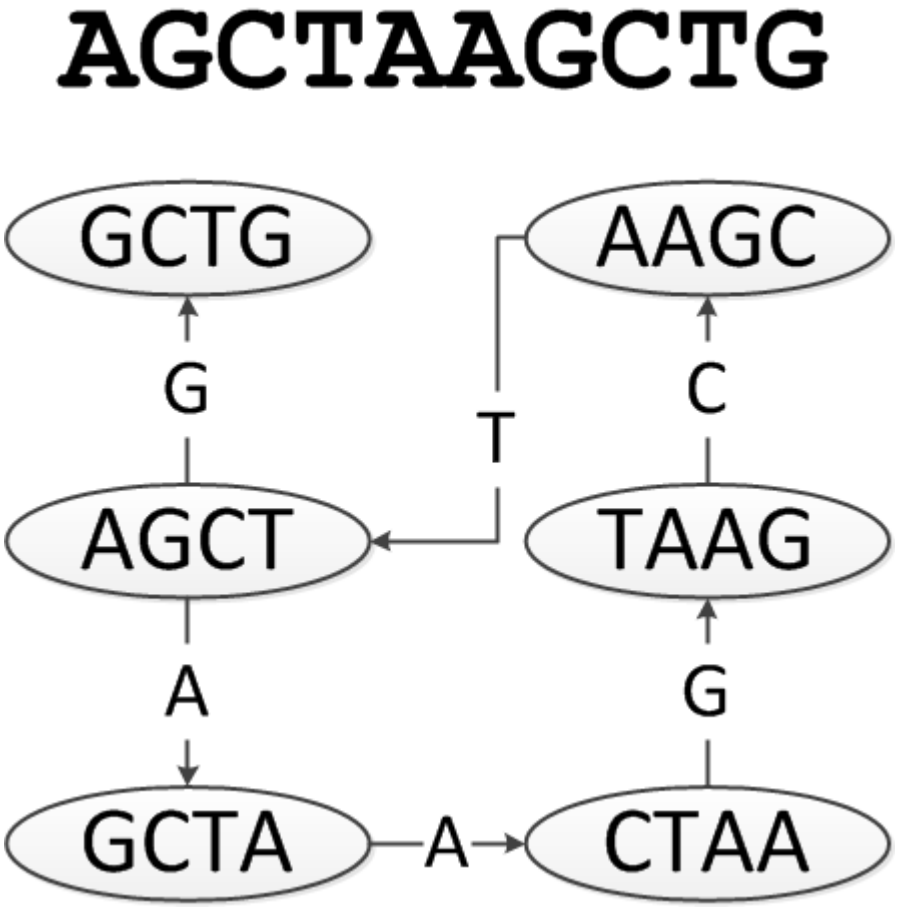
Секвенирование с помощью ДНК-чипов

- С помощью чипа можно определить, содержит ли геном некоторую заданную подстроку
- Зафиксируем длину строки k
- Рассмотрим чип для всех 4^k строк длины k



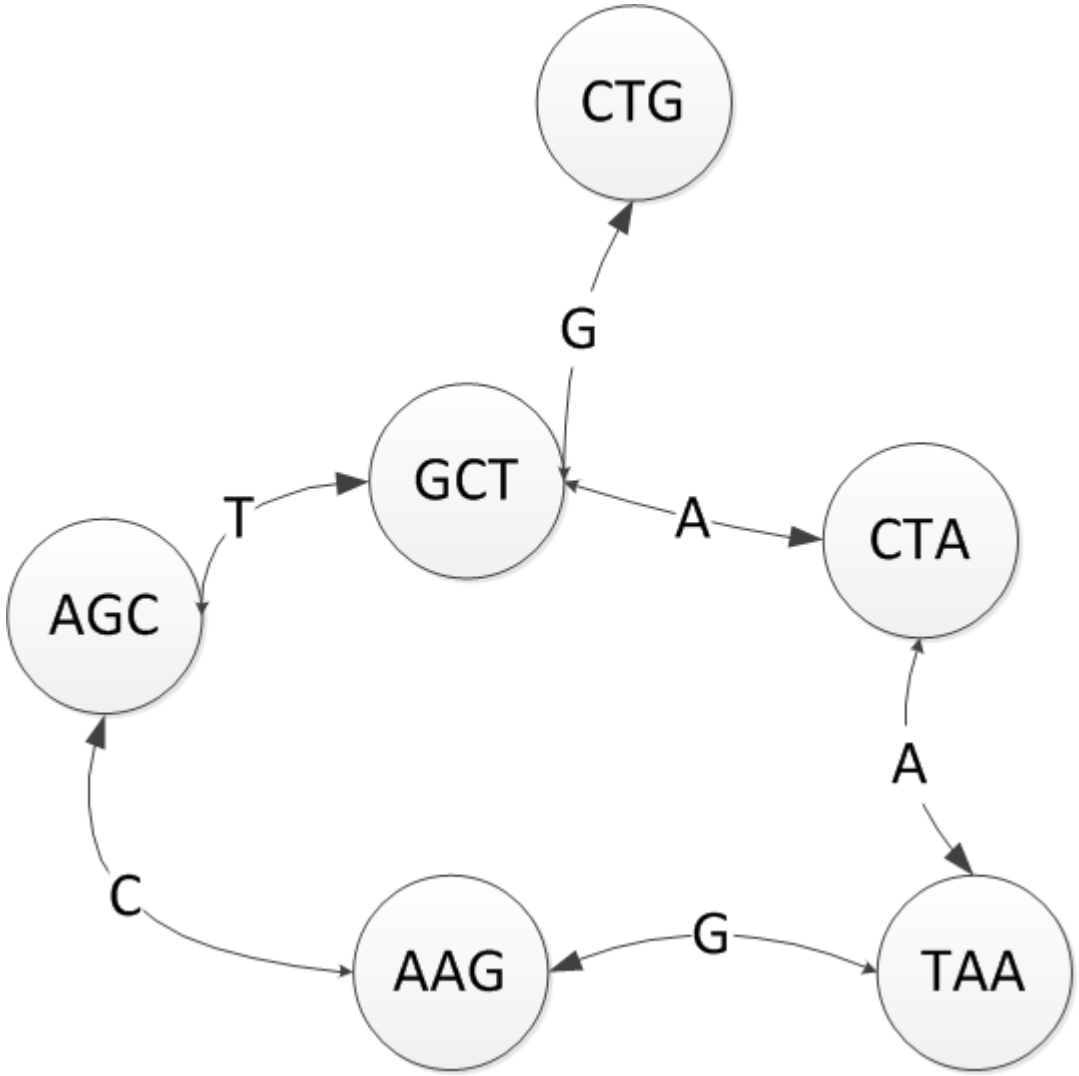
Граф де Брёйна

- Ориентированный граф
- Вершины = строки длины (*k*-1)
- Ребра = строки длины *k*
- Эйлеров путь в этом графе соответствует геномной последовательности

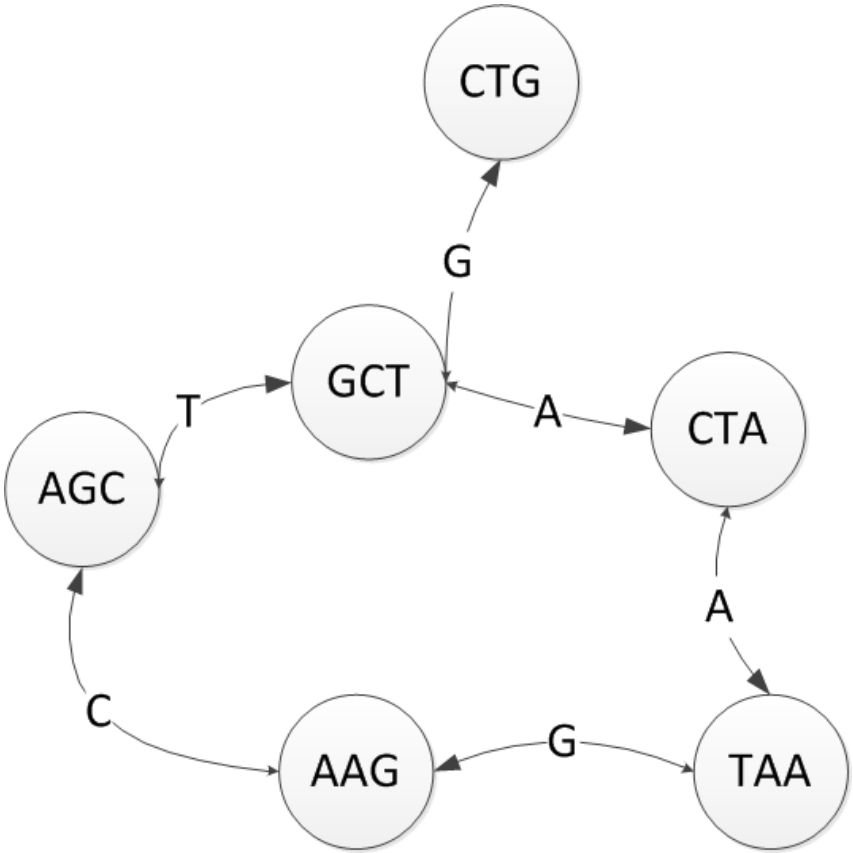


Меньшее *k*

- **AGCTAAGCTG**
- **AGCT**
- **GCTA**
- **CTAA**
- **TAAG**
- **AAGC**
- **AGCT**
- **GCTG**



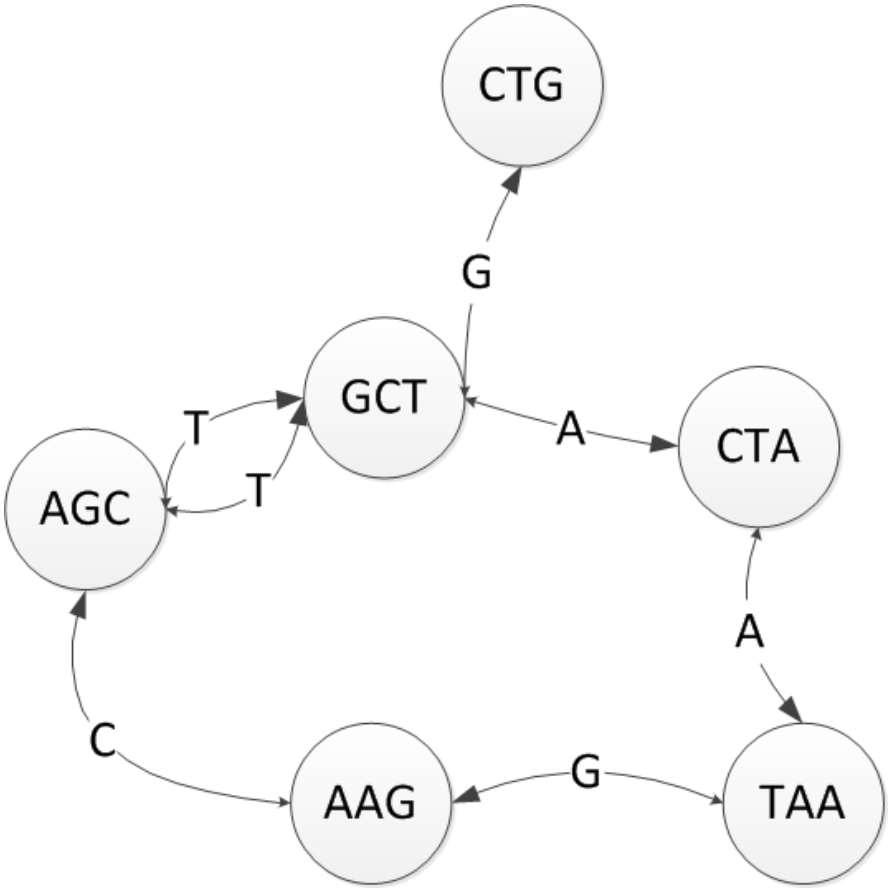
Меньшее *k*



- **GCTAAGCTG**
- Не **AGCTAAGCTG**

- **Проблема возникла из-за повторов!**

Меньшее *k*



- Если знать точное число вхождений, то проблема исправлена

Миф №2

Есть понимание, что значит «собрать геном»

Математические модели сборки генома

- Наименьшая общая надстрока
- Эйлеров путь в графе де Брейна
- Кратчайший суперпуть в графе де Брейна
- Суперпуть в графе де Брейна с кратностями
- Путь в парном графе де Брейна

- Не учитывают ошибки секвенирования!

Наименьшая общая надстрока

- Искомая геномная последовательность – кратчайшая строка, которая содержит чтения в качестве подстрок

- Набор чтений: AATGC, GCATA, CATAG
- Геномная последовательность

AATGCATA

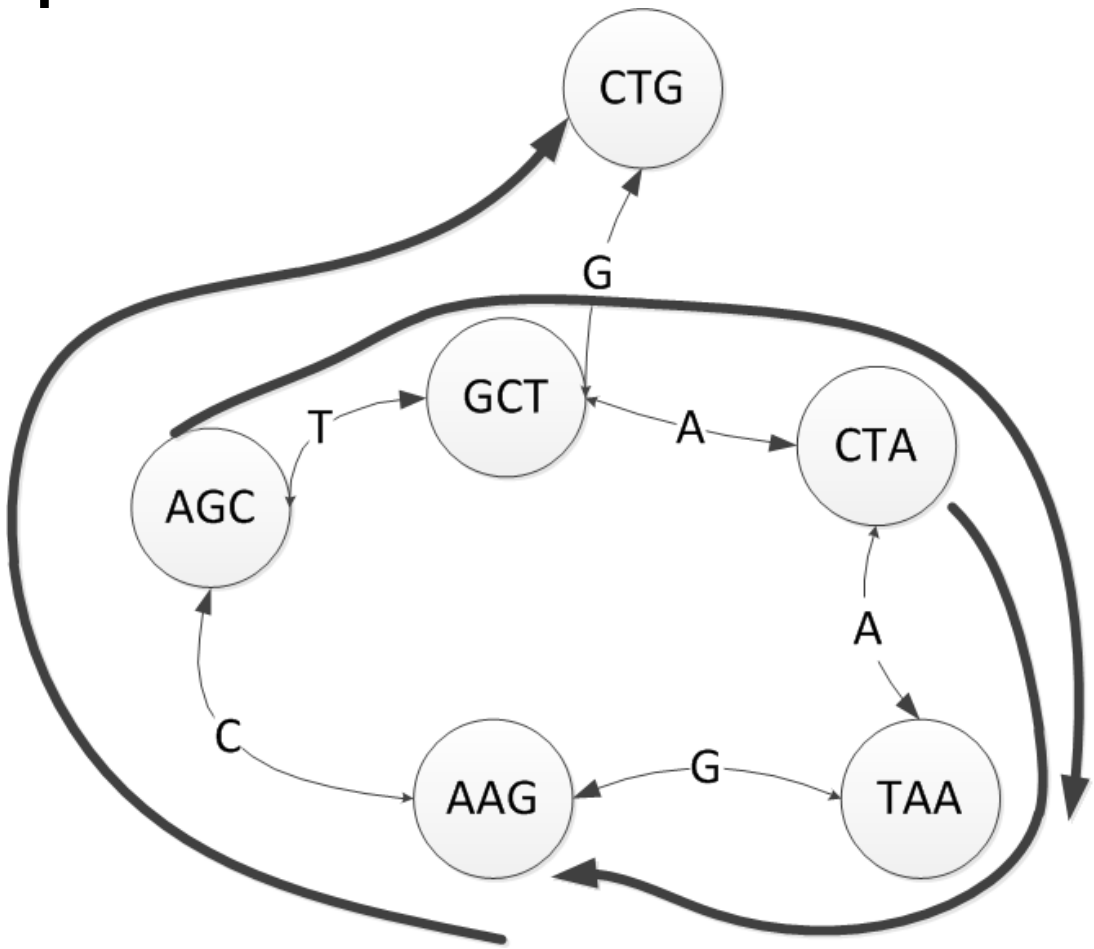
AAT**GC**

GCATA

CATAG

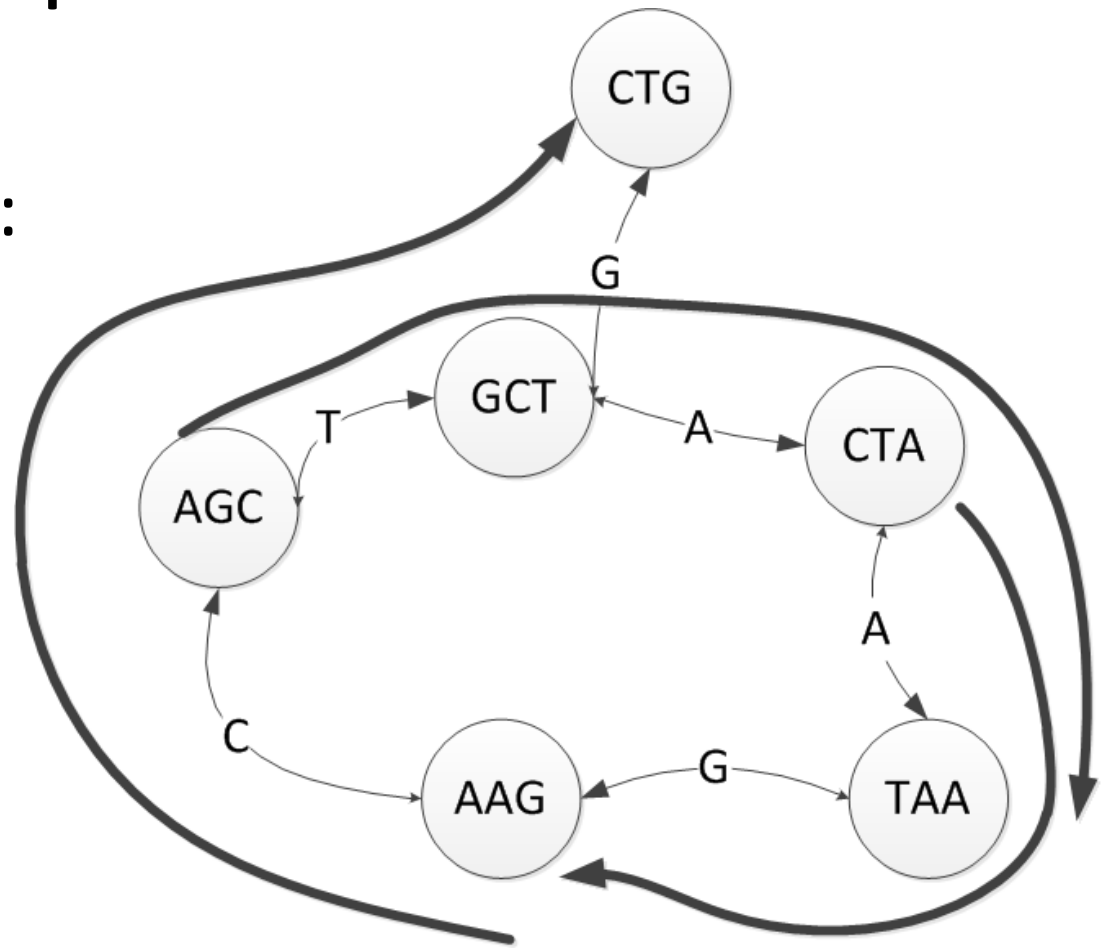
Кратчайший суперпуть в графе де Брейна

- Набор из трех чтений:
  - **AGCTAA**
  - **CTAAG**
  - **AAGCTG**



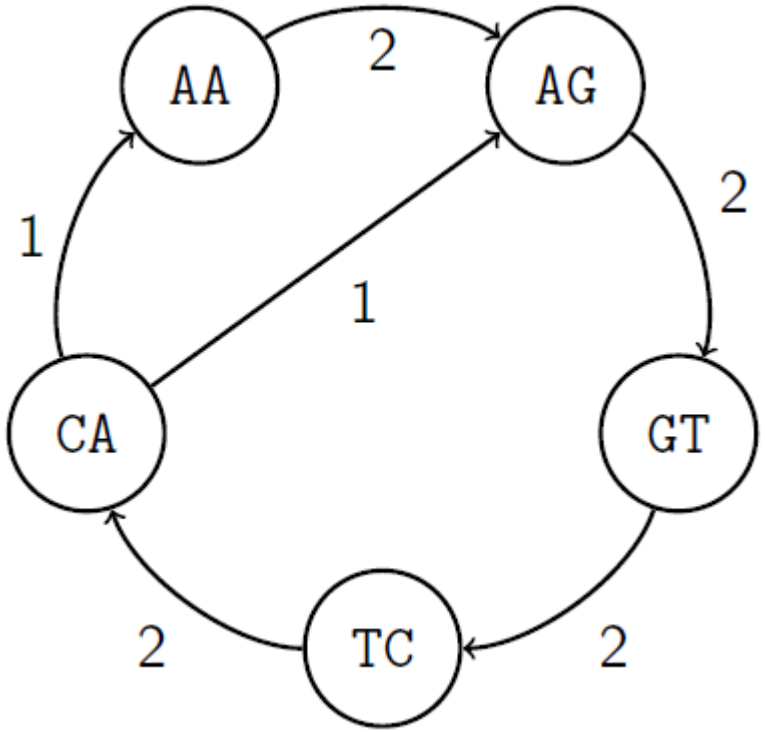
Кратчайший суперпуть в графе де Брейна

- Искомая геномная последовательность:
  - **AGCTAAGCTG**



Суперпуть в графе де Брейна с кратностями

- Кратности – по принципу максимального правдоподобия (Medvedev and Brudno, 2009; Varna et al., 2011)



- Чтения: AAGT, AGTCA, TCAA
- Суперпуть: AAGTCAGTCAAG

Сложность сборки генома для различных моделей

- Наименьшая общая надстрока – труднорешаемая (Gallant et al., 1980)
- Эйлеров путь в графе де Брейна – решается за время, пропорциональное размеру входных данных (Pevzner et al., 1989)
- Суперпуть в графе де Брейна – труднорешаемая (Medvedev et al., 2007)
- Суперпуть в графе де Брейна с кратностями – труднорешаемая (Kapun and Tsarev, 2013)
- Путь в парном графе де Брейна – труднорешаемая (Kapun and Tsarev, 2013)

Миф №3

Существующие программы для сборки генома хорошо его собирают

Как работают сборщики геномов?

- Основаны на эвристических или приближенных алгоритмах
- Собирают не целую геномную последовательность, а контиги и скэффолды
- Распространенные метрики сборки генома мало связаны с качеством сборки

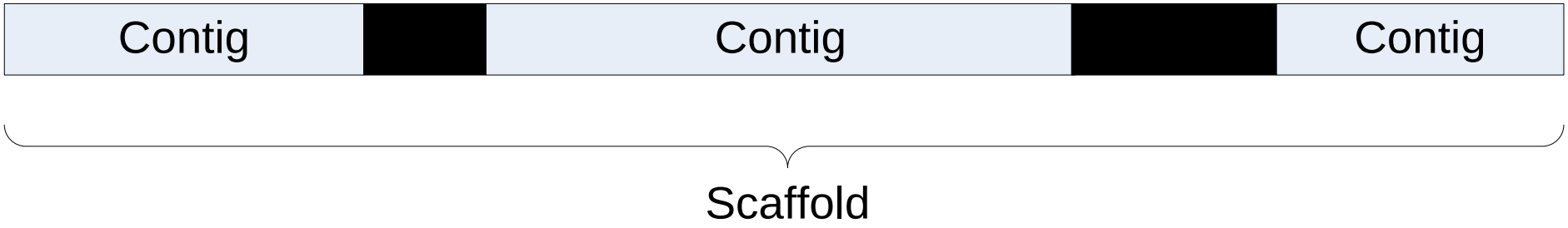
Контиги

- Непрерывная последовательность, которая с большой долей уверенности является частью геномной последовательности

Contig

Скэффолд

- Скэффолд – последовательность контигов, для которых известен их относительный порядок и расстояния между ними



Метрики сборки генома

- Длина кратчайшего контига/скэффолда
- Длина наибольшего контига/скэффолда
- Средняя длина контига/скэффолда

Метрики сборки генома

- N50/N90 – наибольшая длина контига такая, что в контигах не меньшей длины содержится 50/90% *суммарной длины контигов*
- NG50/N90 – наибольшая длина контига такая, что в контигах не меньшей длины содержится 50/90% *суммарной длины генома*
- Аналогично – для скэффолдов

Пример

- Длина контигов:

5, 7, 10, 15, 22, 24, 30, 45

- Длина кратчайшего – 5
- Длина наибольшего – 45
- Средняя длина – (5 + 7 + 10 + 15 + 122+ 24 + 30 + 45) / 8 = 19.75

Пример

- N50 = 24, так как
  - 30 + 45 = 75 < 50% от 158
  - 24 + 30 + 45 = 99 > 50% от 158
- Если длина генома 100, то NG50 = 30
- Если длина генома 200, то NG50 = 22

Выводы

- Есть разрыв между теорией и практикой сборки генома
- Нет сборщика генома, который работает лучше других на всех наборах данных
- Сравнивать сборщики генома можно только на одних и тех же данных
- Один и тот же геном надо пробовать собирать разными сборщиками
- Надо следить за проектами по экспериментальному сравнению различных сборщиков: Assemblathon, GAGE

Спасибо за внимание!