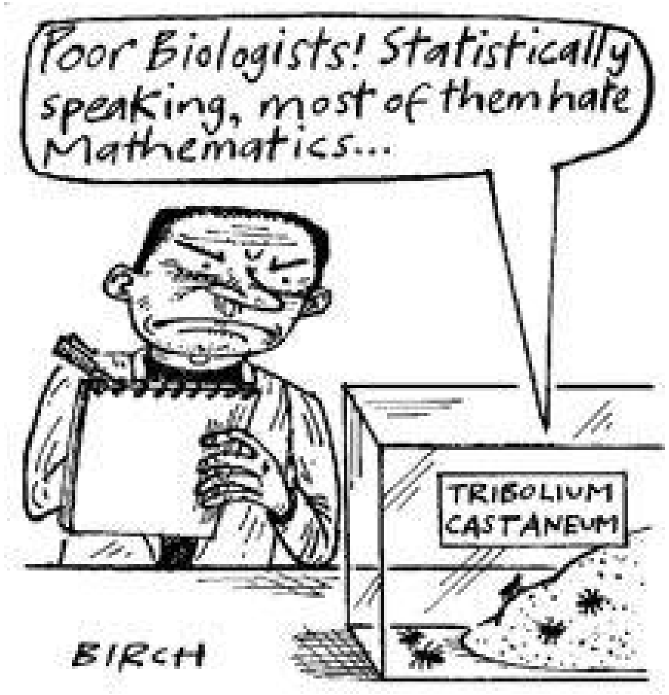
intro@bioinformatics:~$

Александр Предеус
Институт Биоинформатики

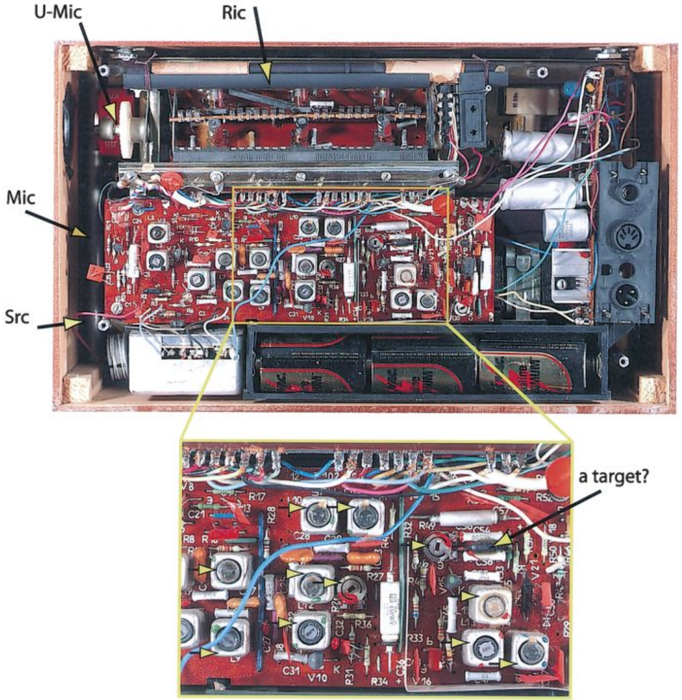
Биология и математика

- нормальное развитие естественной науки подразумевает увеличение количественных моделей
- описательная стадия → научные модели → инженерия
- биология избегала этого очень долго!



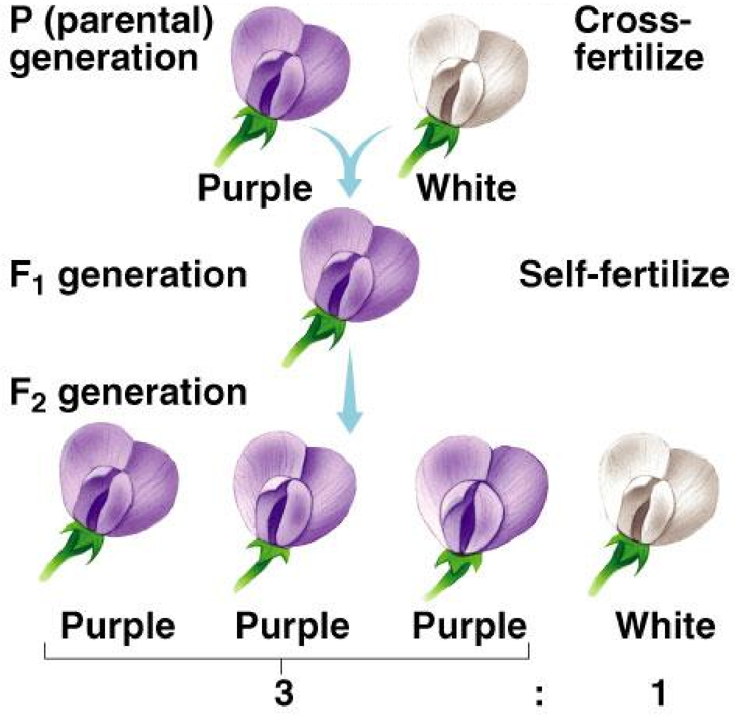
Способен ли биолог починить радио?

- “Can biologist fix a radio - Or, what I have learnt while studying apoptosis”, by Yury Lazebnik



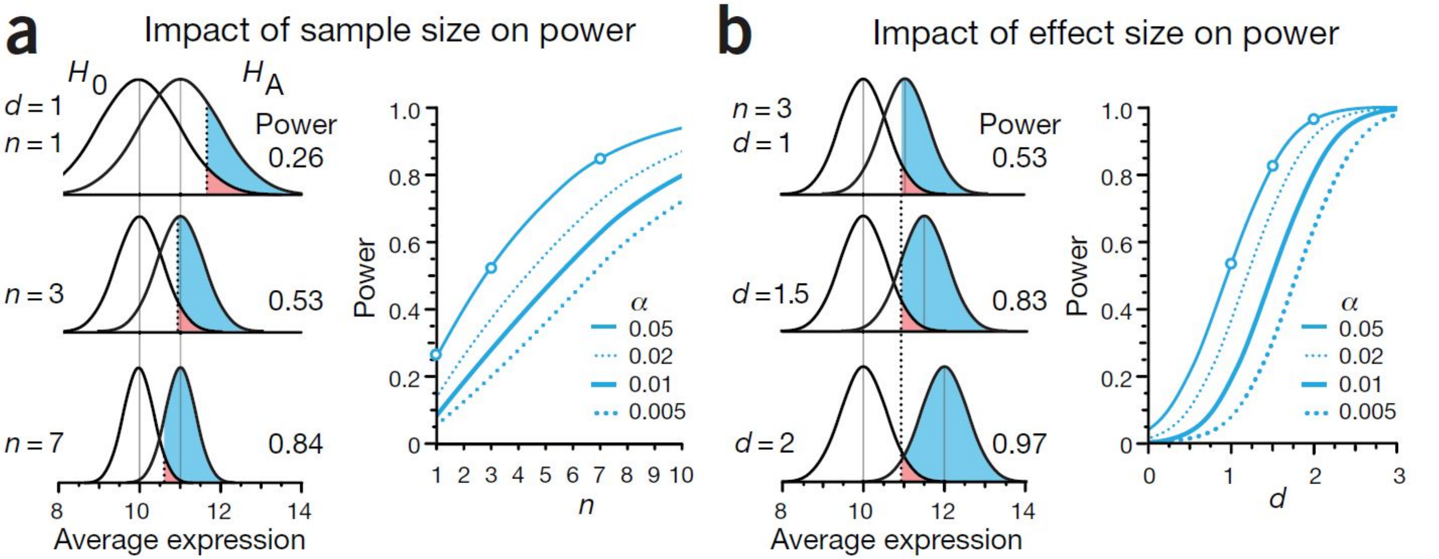
Эволюция математики в биологии

- любая описательная наука требует знания статистики
- статистический анализ - мощнейший инструмент во всех науках, где построение точных моделей затруднено

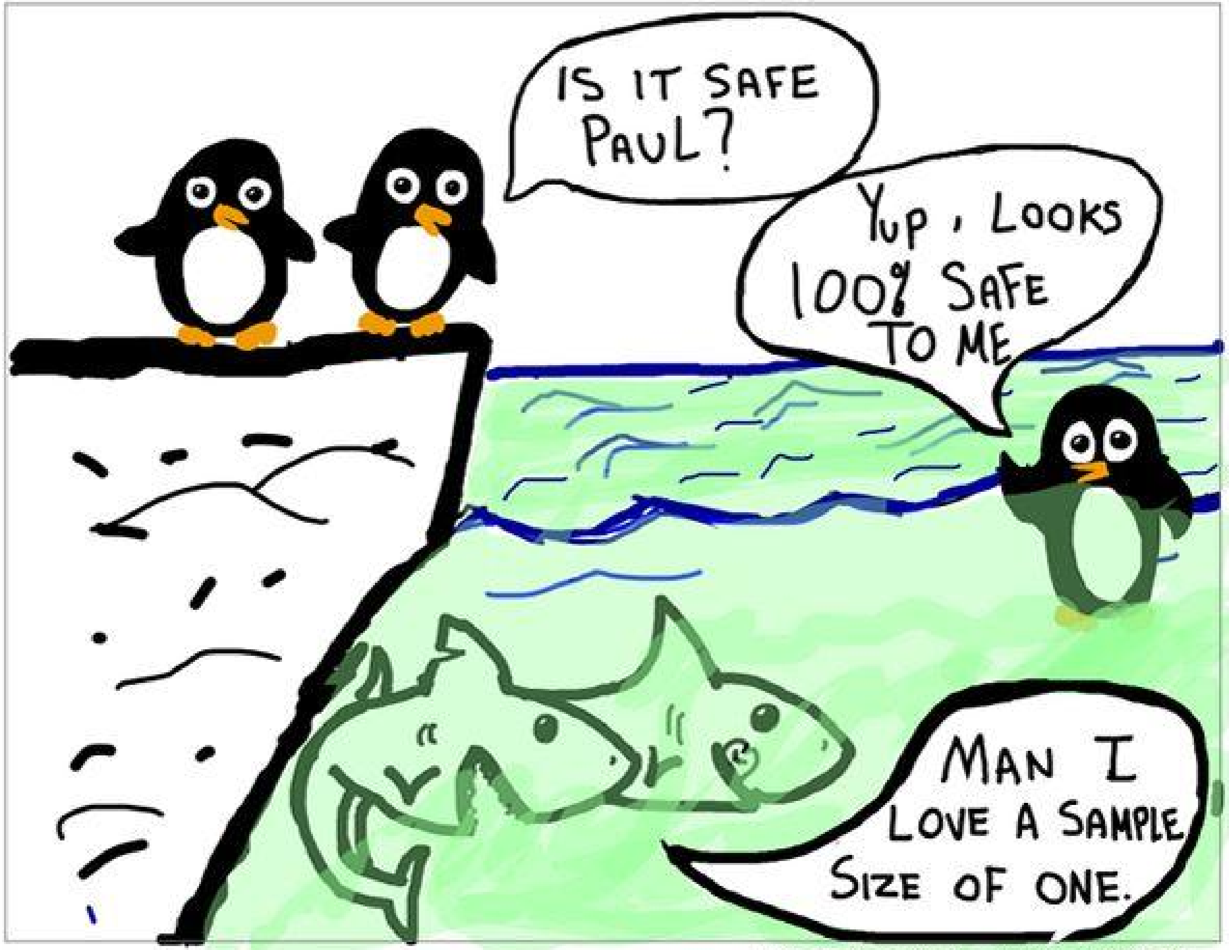


Значимость результатов

- размер выборки и эффекта имеет решающее значение в интерпретации любого количественного эксперимента



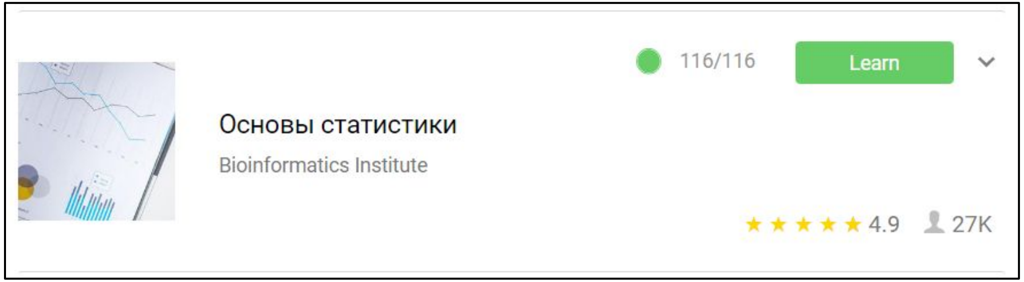
https://www.nature.com/collections/qghhqm/pointsofsignificance

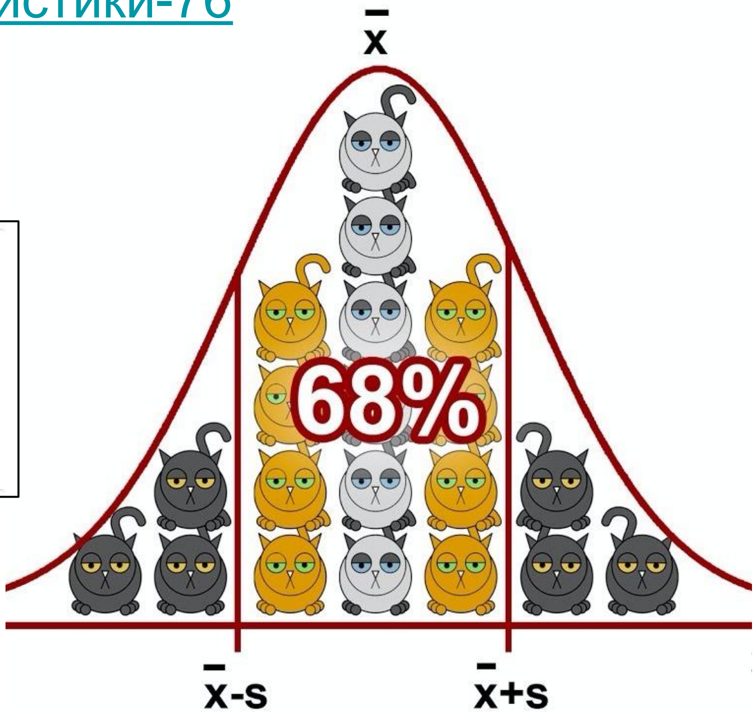


#DataDoodle by @WholeWhale

Статистика и котики

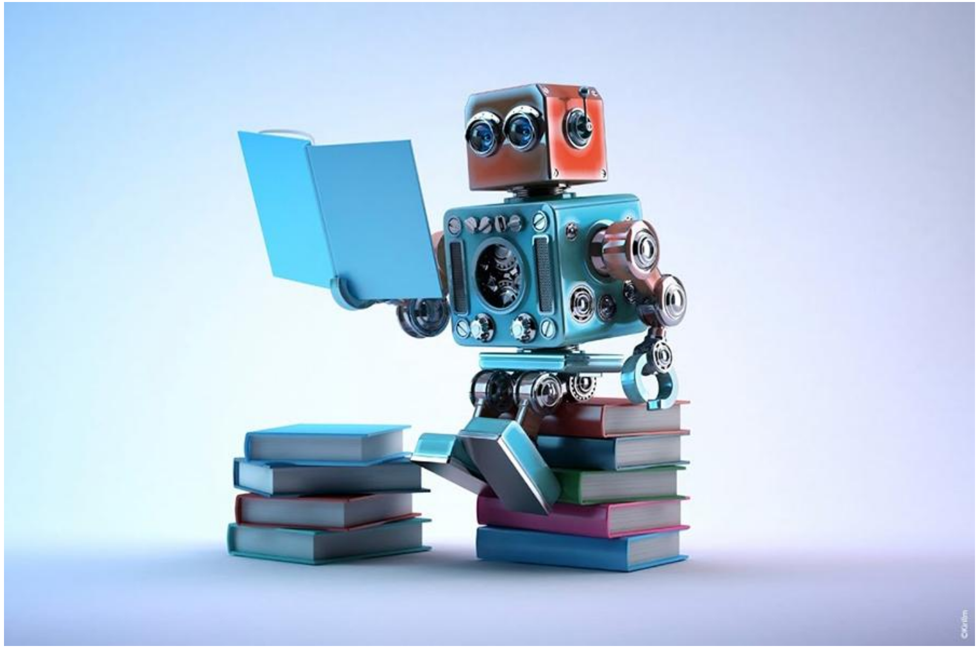
- мало кто любит статистику, зато все любят котиков (с)
- https://stepik.org/course/Основы-статистики-76





Машинное обучение

- классификация
- регрессия
- кластеризация
- уменьшение размерности
- нейросети
- поиск аномалий



Эволюция “биоинформатики”

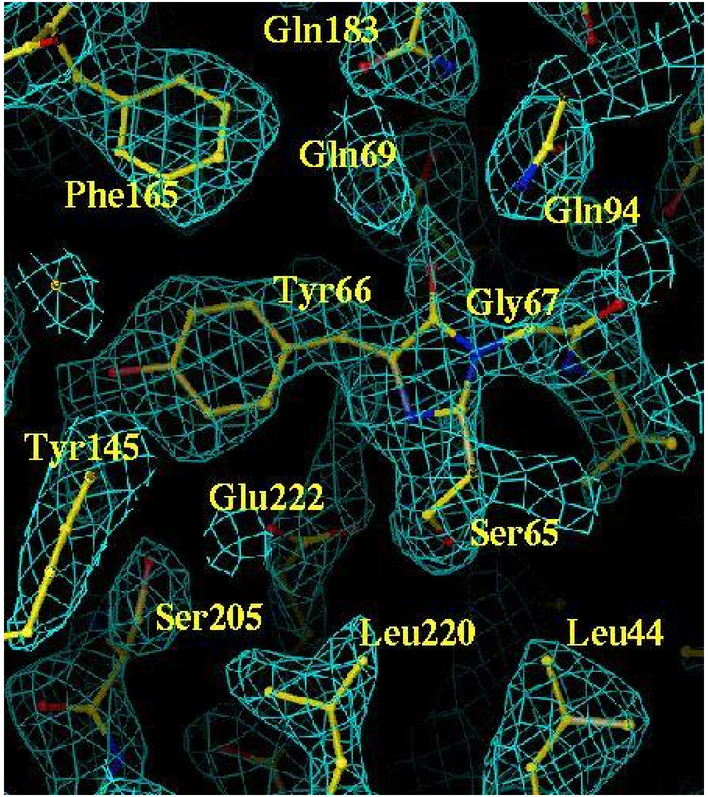
- математическая биология
    - дифференциальные уравнения, кинетические модели, и т.д.
- структурная биоинформатика
    - ака биофизика
    - рост связан с бумом кристаллографии протеинов в 90-е
- геномная биоинформатика
    - рост связан с бумом секвенирования
- системная биология
    - графы (сети), байесовские модели, анализ потоков, и т.д.
    - попытки интегрировать различные данные

Структурная биоинформатика

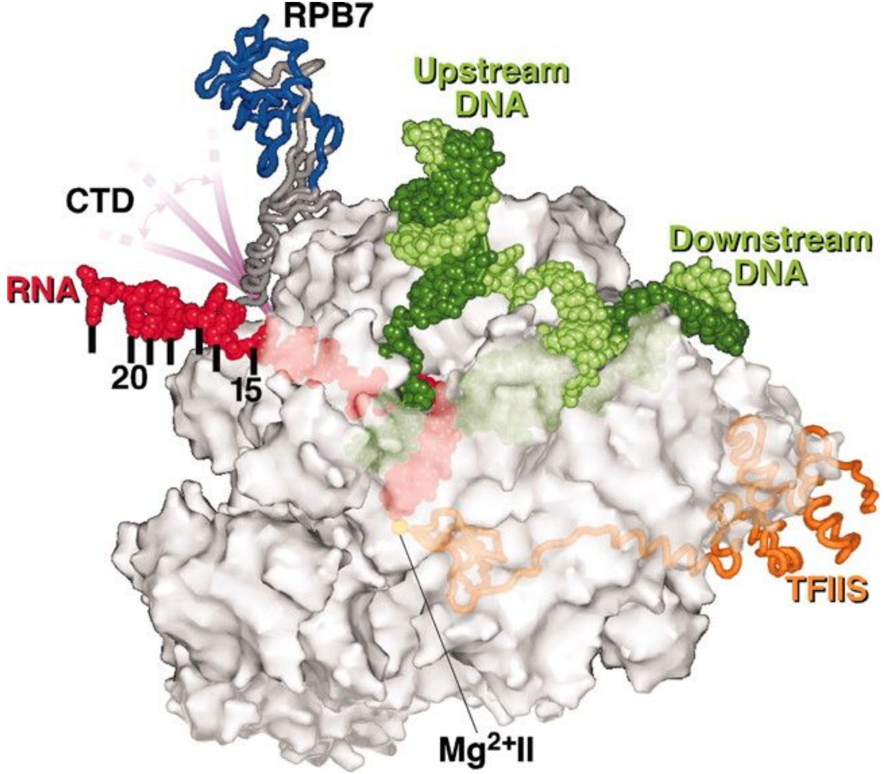
- предсказание и фолдинг неизвестных структур белков
- симуляция поведения белков и комплексов белок-ДНК
- “крупнозернистые” симуляции больших биологических объектов (органелл, вирусов, клеток, участков мембраны)
- докинг малых молекул (лекарств)

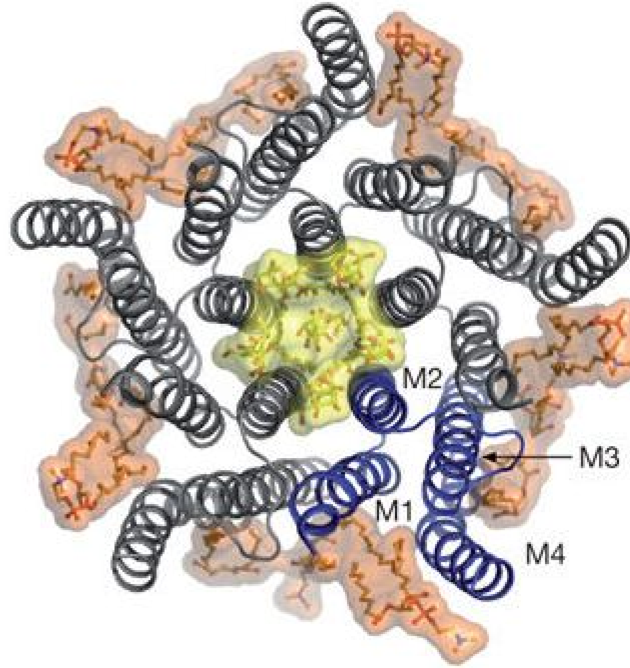
Кристаллография протеинов

- монокристалл протеина облучается рентгеновскими лучами
- дифракционная картина позволяет восстановить позиции индивидуальных атомов с высокой достоверностью



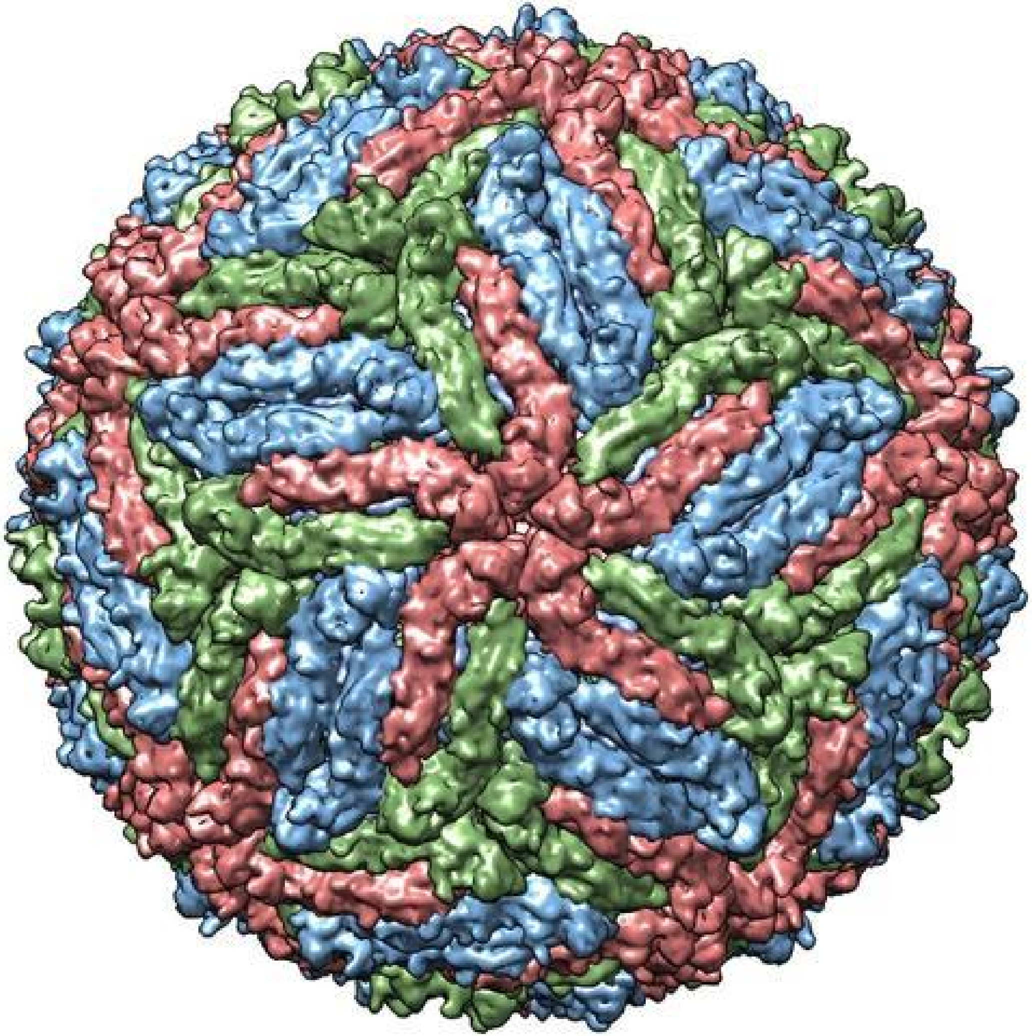
Структуры, потрясающие воображение





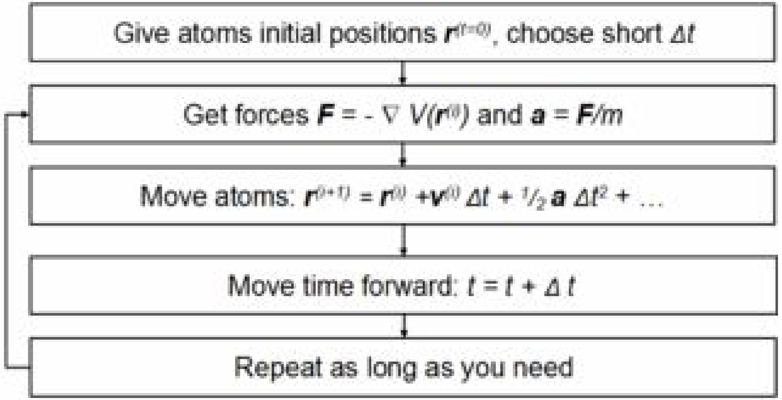
Cryo-EM

- не требует кристаллизации!
- разрешение становится все лучше
- незаменим для определения структур огромного размера
- например, вируса Зика

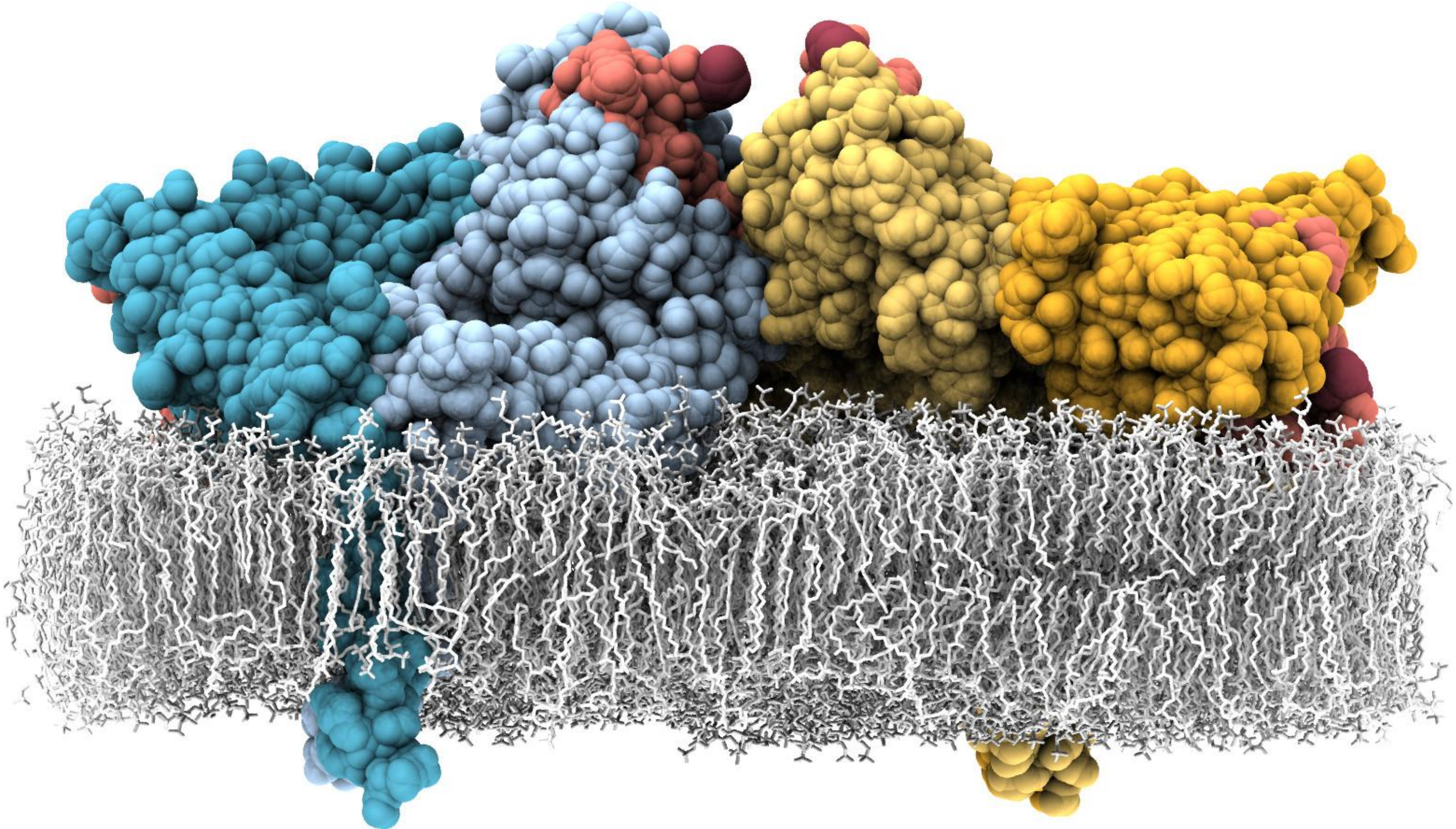


Молекулярная динамика

- решение ньютоновских дифференциальных уравнений движения

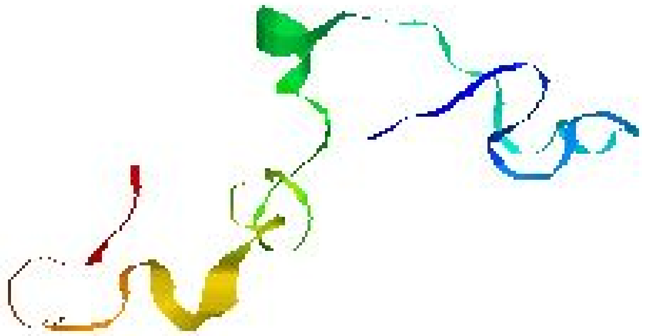


Симуляции миллионов атомов



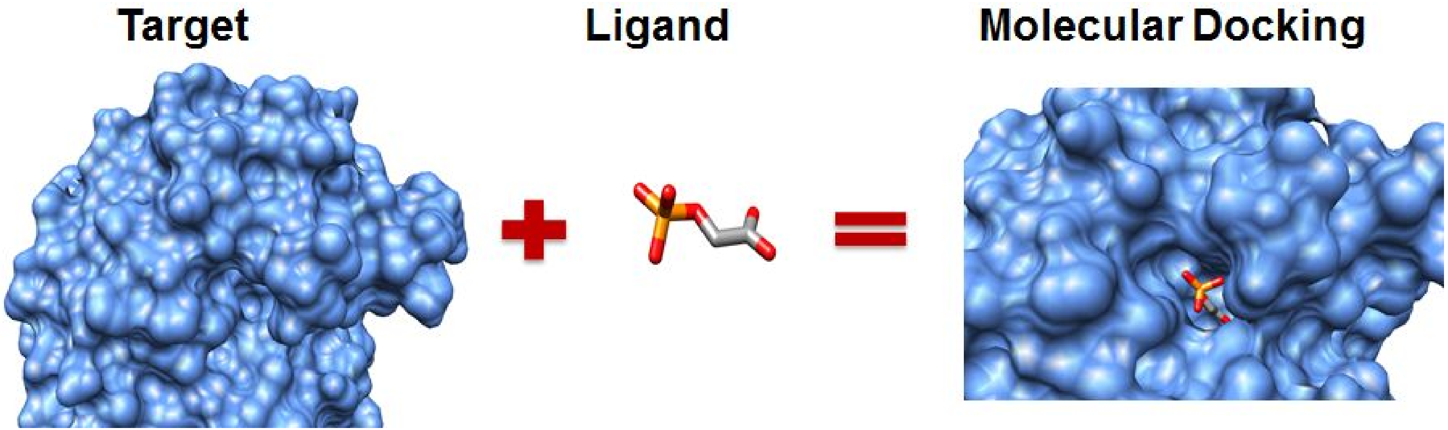
Фолдинг протеинов

- отличается от молекулярной динамики целью и подходами
- молекулярная динамика может использоваться для дальнейшего улучшения структуры



Докинг малых молекул

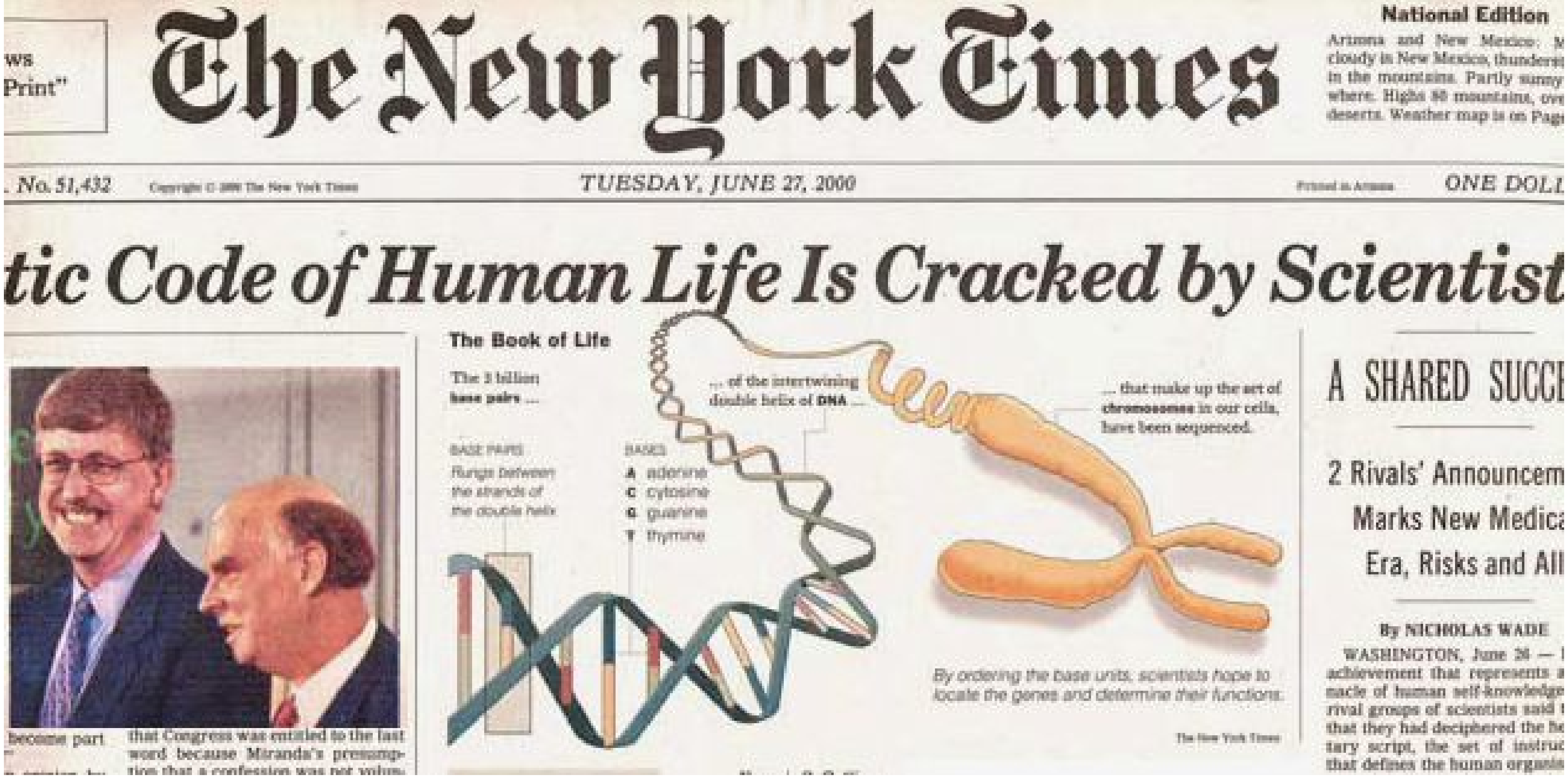
- большинство современных лекарств ингибируют ферменты, блокируя активный центр
- предсказание активных молекул может сильно снизить цену разработки новых лекарств



Геномная биоинформатика

- прогресс связан с прогрессом в секвенировании
- секвенирование - определение последовательности (ДНК, РНК, протеина)
- задачи:
  - поиск нуклеотидных и протеиновых последовательностей в базах данных (*blast*)
  - сборка геномов *de novo* (*De Bruijn graphs*)
  - выравнивание на собранные геномы и транскриптомы (*FM-index, Burrows-Wheeler transform*)

Проект “геном человека”



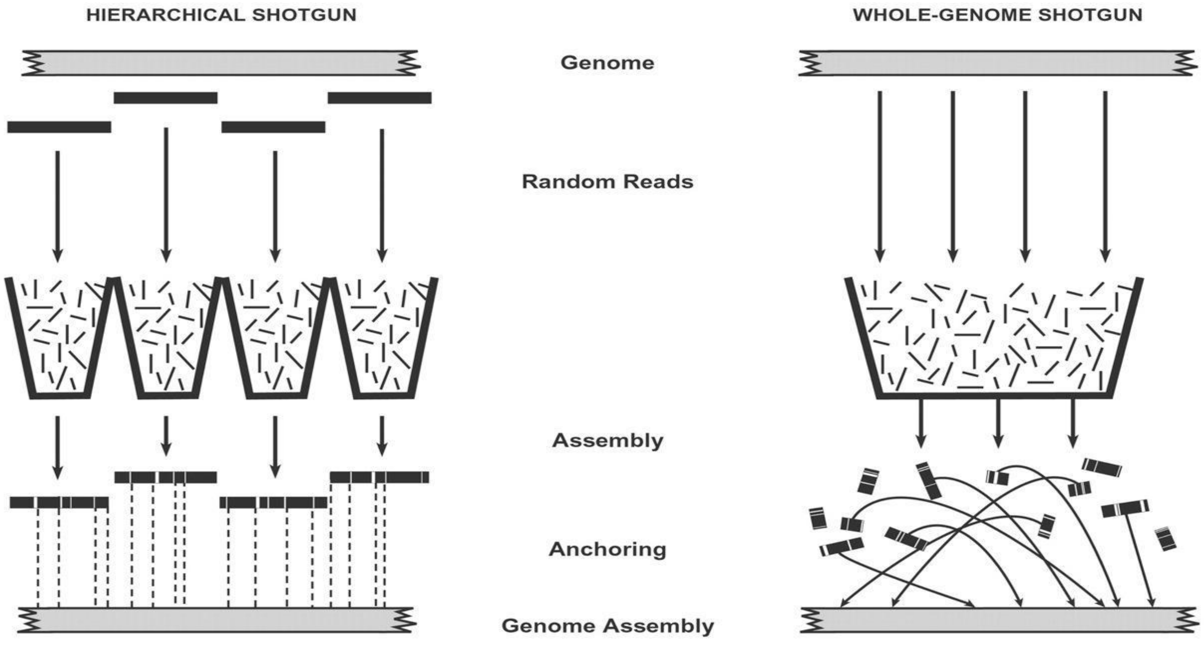
Университеты vs. Celera Genomics

Francis Collins

Eric Lander

Craig Venter

Соревнование стратегий

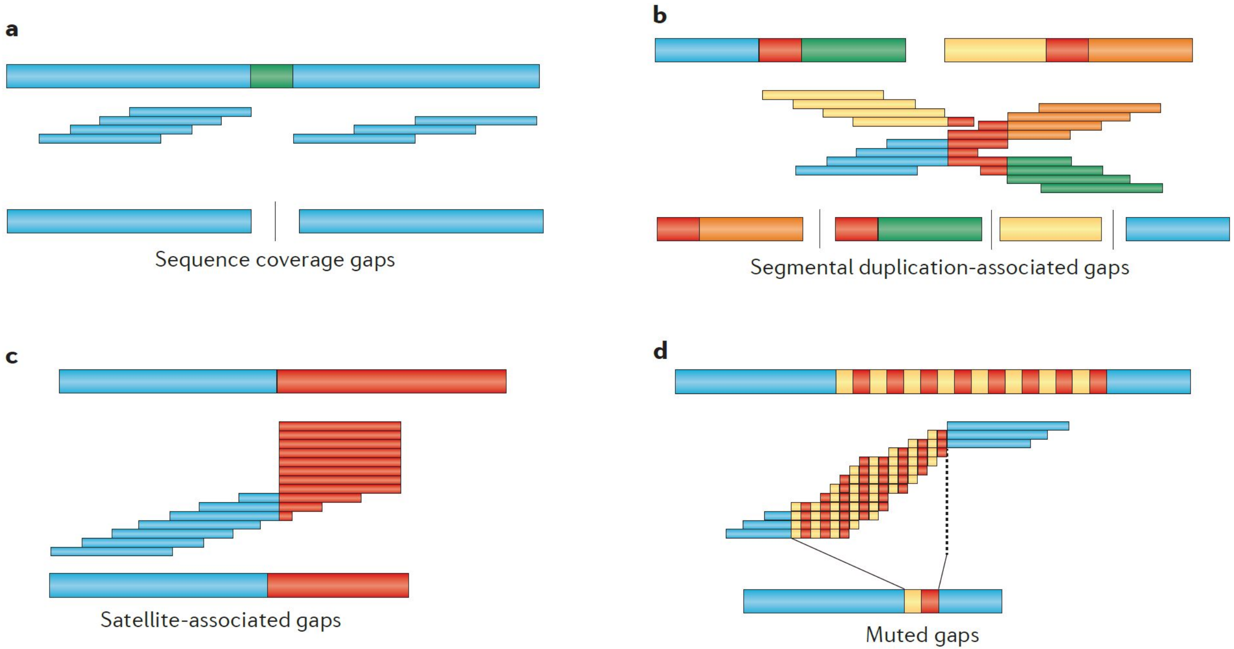


Public (Universities)
1990-2001 (2003)
3 billion dollars

Celera Corporation
1999-2001 (2003)
300 million dollars

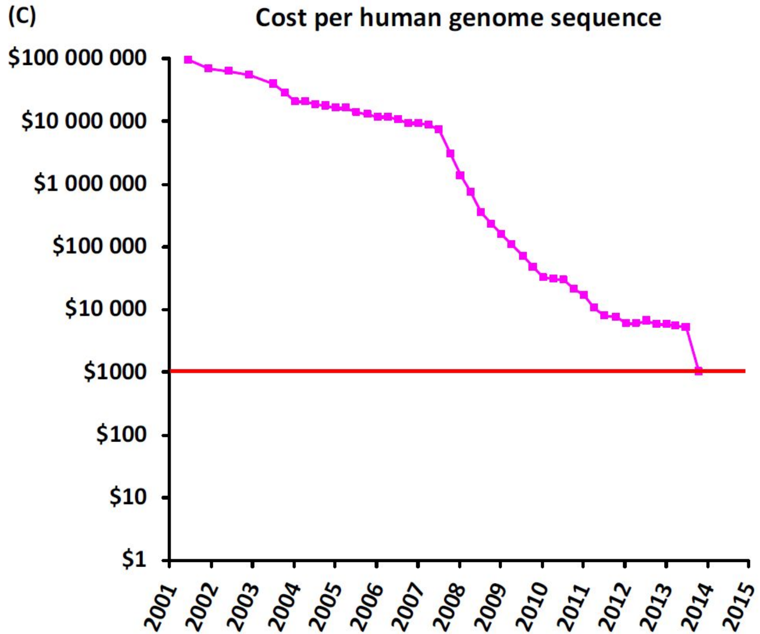
Возможно ли?

- Сомнение в возможности сборки коротких прочтений в геном
- Трудности – повторы, плохое покрытие



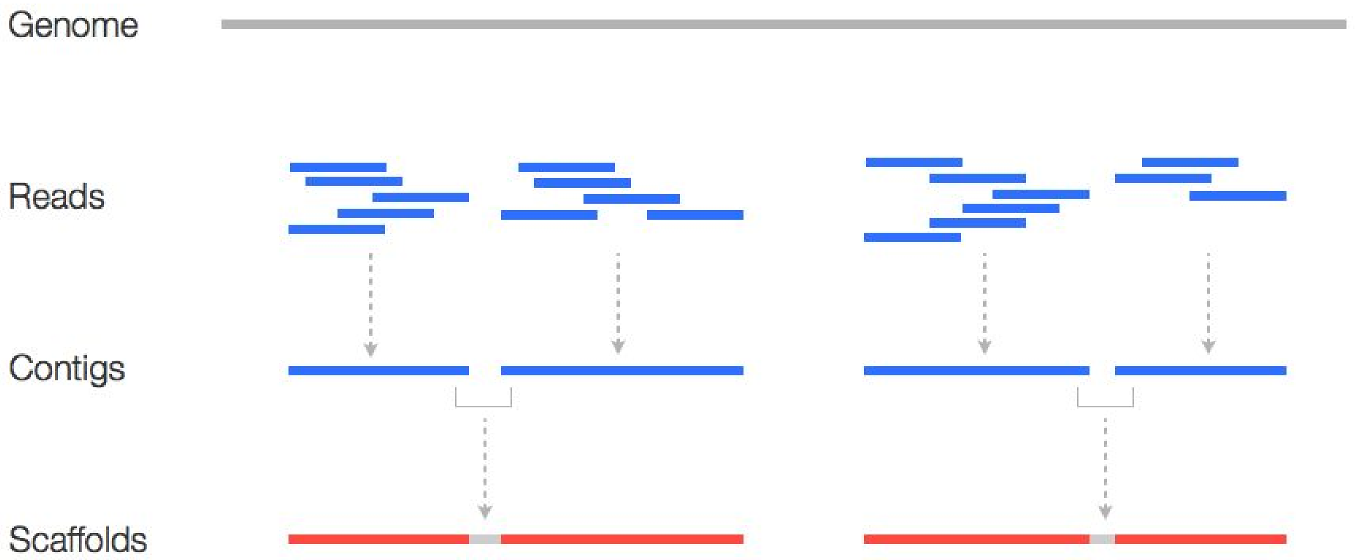
Секвенирование: революция 2000-х

- Секвенирование генома человека упало в цене от 100 миллионов до ~1000 $



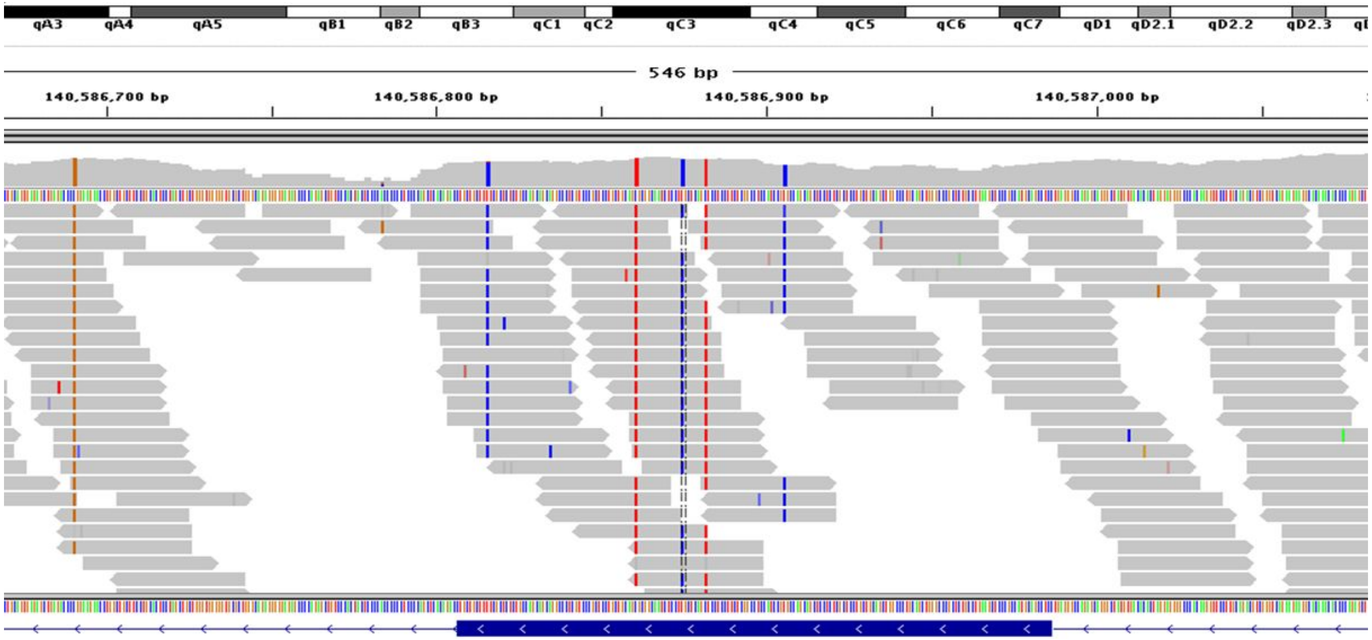
Тип 1: неизученные виды (сборка)

- Сборка геномов (de-novo genome assembly)
- Сборка транскриптомов (de-novo transcriptome assembly)



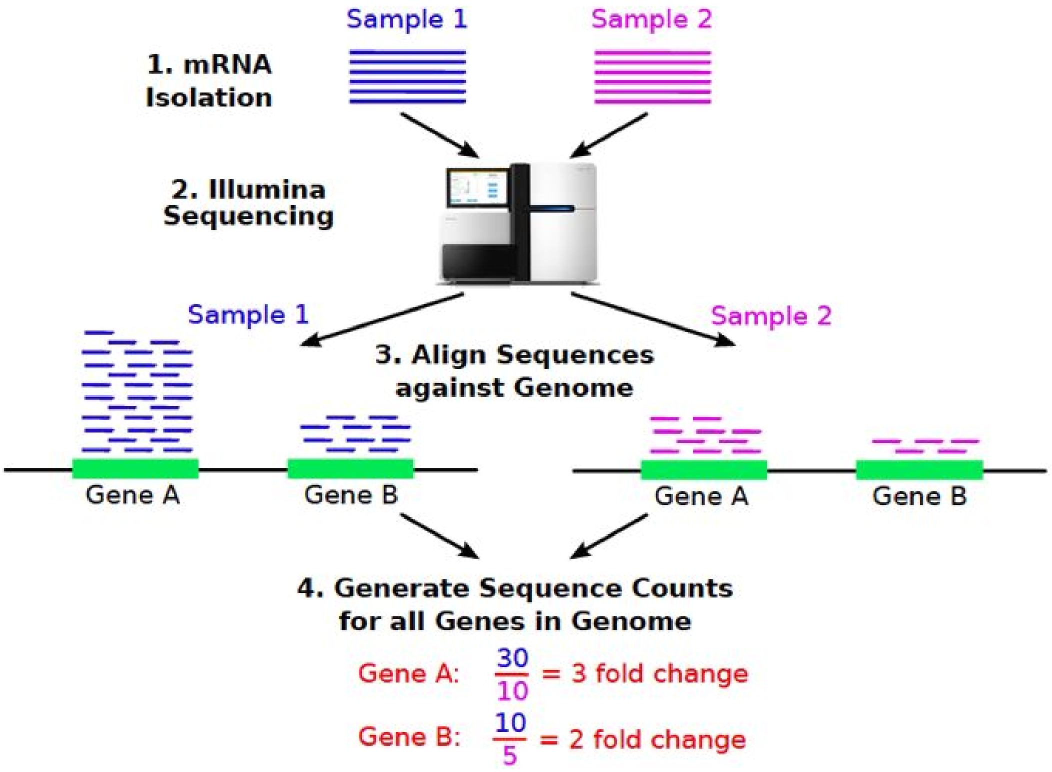
Тип 2: Уточнение последовательности (выравнивание)

- Ресеквенирование геномов
- Ресеквенирование частей генома



Тип 3: Количественный анализ (выравнивание)

- Транскриптом (RNA-seq)
- Позиционный (ChIP-seq, genome footprinting)



Что (и кого) мы секвенируем?

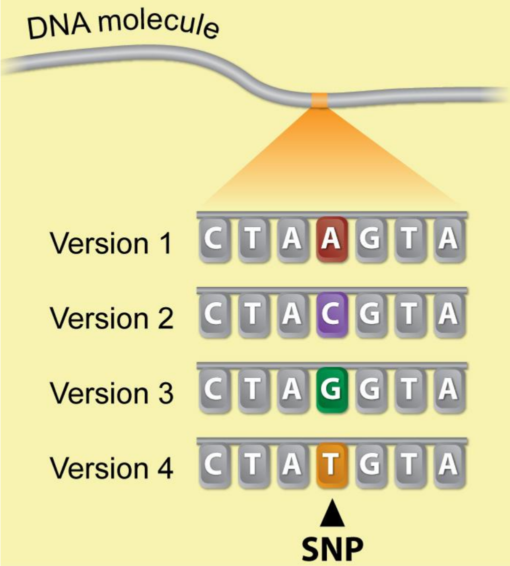
- Плазмиды и конструкты
- Вирусы
  - новые
  - количество (вирусная нагрузка)
  - варианты (штаммы)
- Прокариоты
  - новые
  - варианты (штаммы)
- Эукариоты
  - новые
  - модельные животные
- Люди
  - пациенты
  - исследование здоровых волонтеров

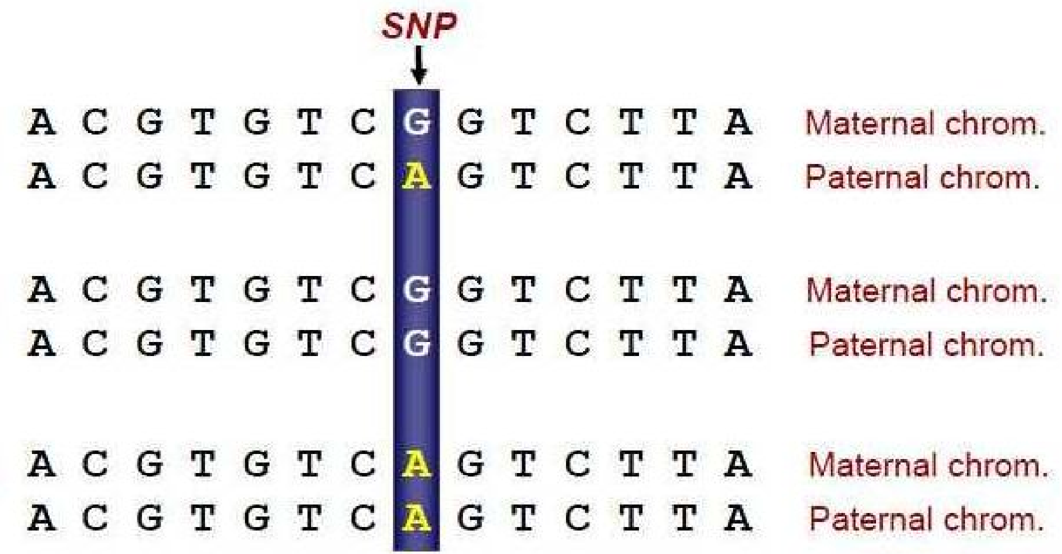
Примеры применения NGS

- исследование вариации в геноме человека
- определение присутствующих организмов (“баркодинг”)
- анализ дифференциальной экспрессии - сравнить те же ткани у больных и здоровых людей
- поиск участков ДНК, активных в определенных тканях и условиях

Вариативность в геномах

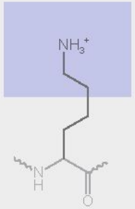
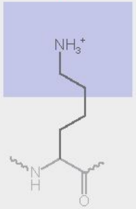
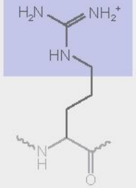
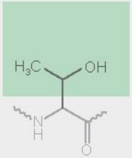
- SNP (SNV) - Single Nucleotide Polymorphism (Variant)
- Может быть гетеро/гомозиготный





Типы однонуклеотидных замен

- Эффект замены радикально зависит от ее типа

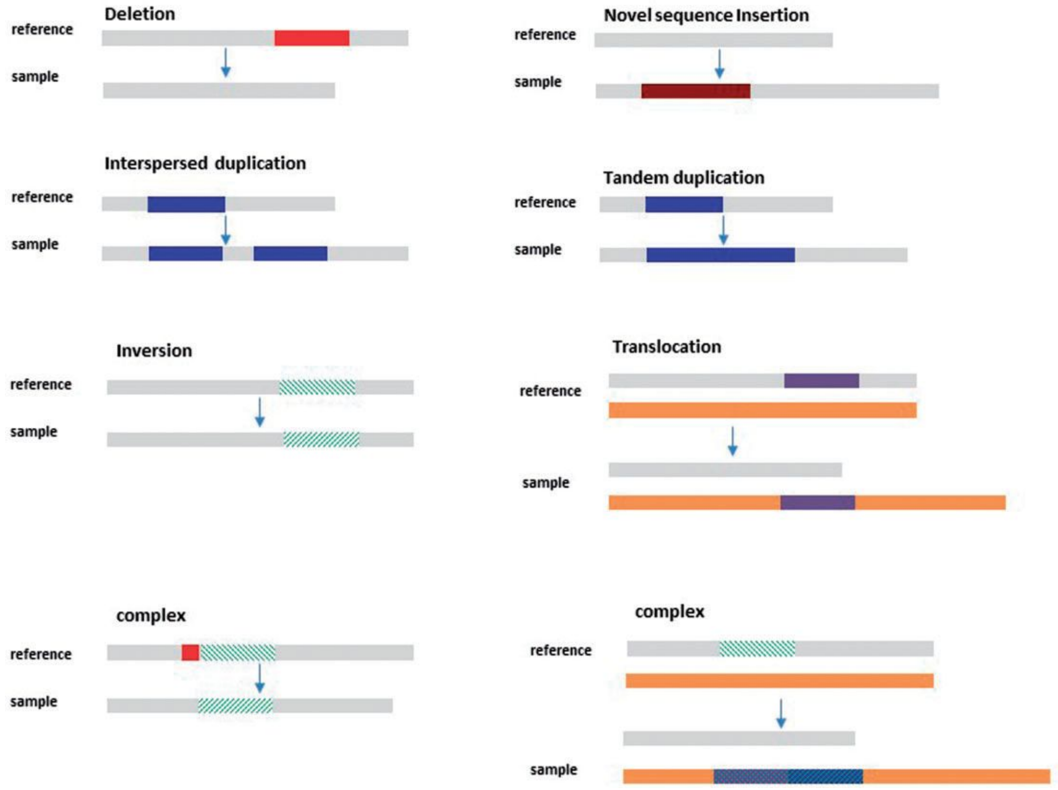
		Point mutations			
	No mutation	Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	**Lys**	**Lys**	**STOP**	**Arg**	**Thr**
					

basic
polar

Вариативность в геномах

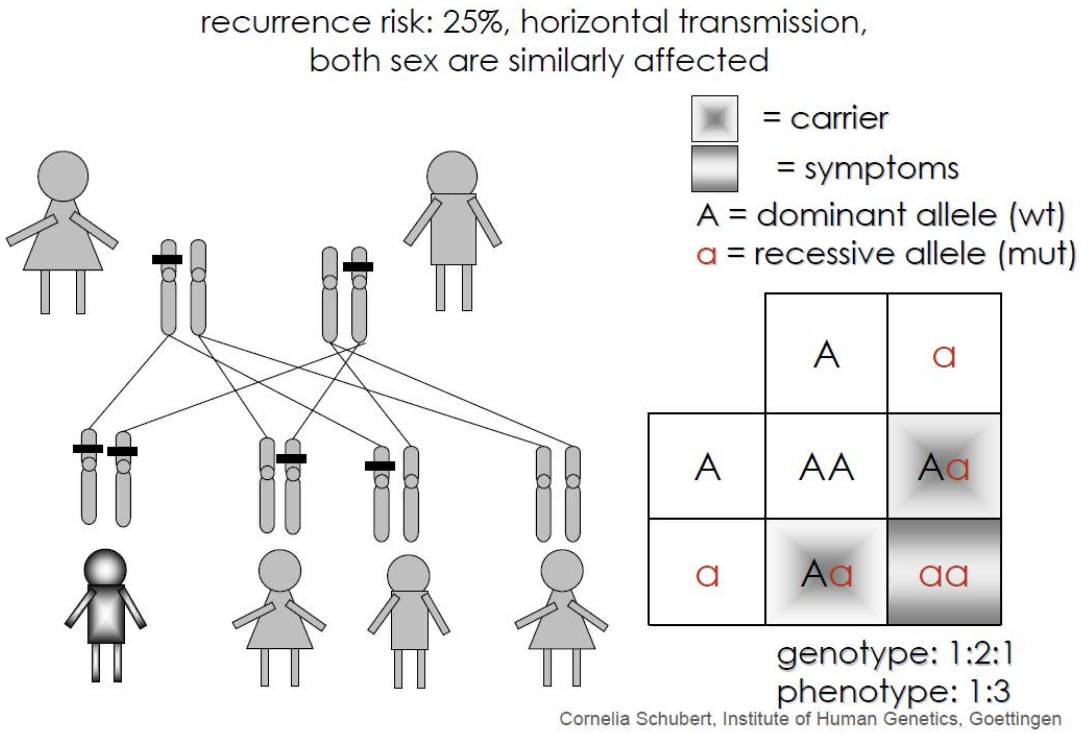
- “Инделы” - инсерции и делеции
- Крупные структурные вариации

```
Indel examples
wild-type sequence
ATCTTCAGCCATAAAAGATGAAGTT
3 bp deletion
ATCTTCAGCCAAAGATGAAGTT
4 bp insertion (orange)
ATCTTCAGCCATATGTGAAAGATGAAGTT
```



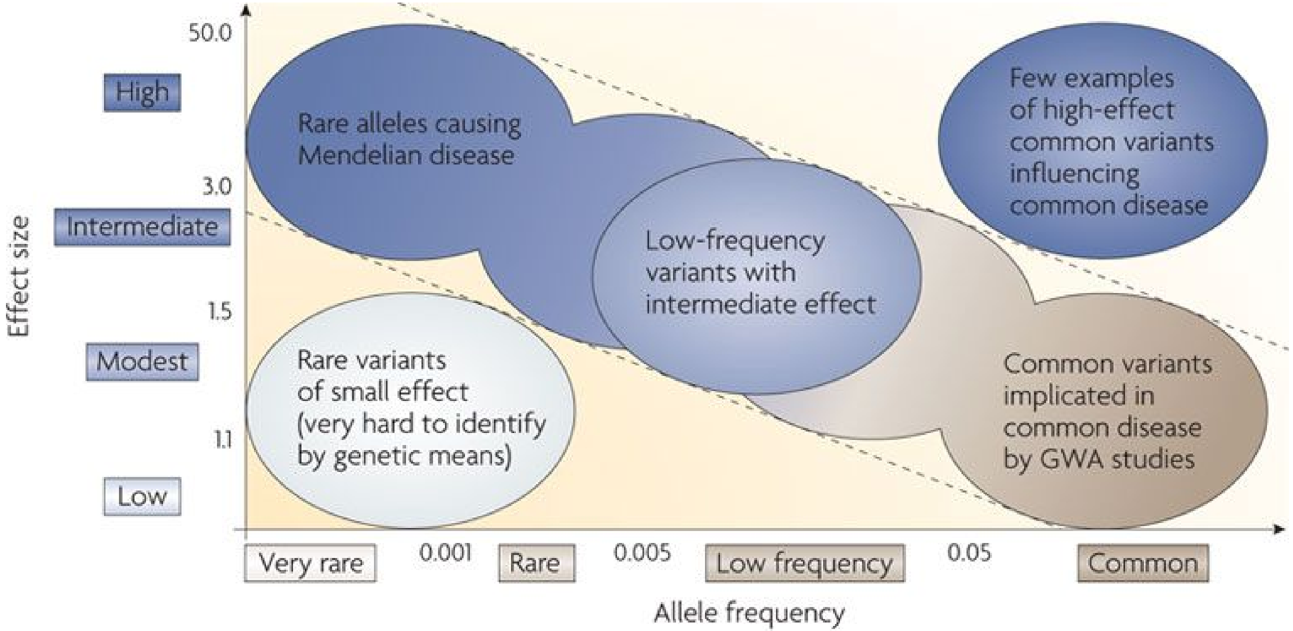
Медицинский NGS

- Около 7000 редких заболеваний, 80% генетических
- Общая частота - около 10%
- Менделевские болезни - от 2 до 5%



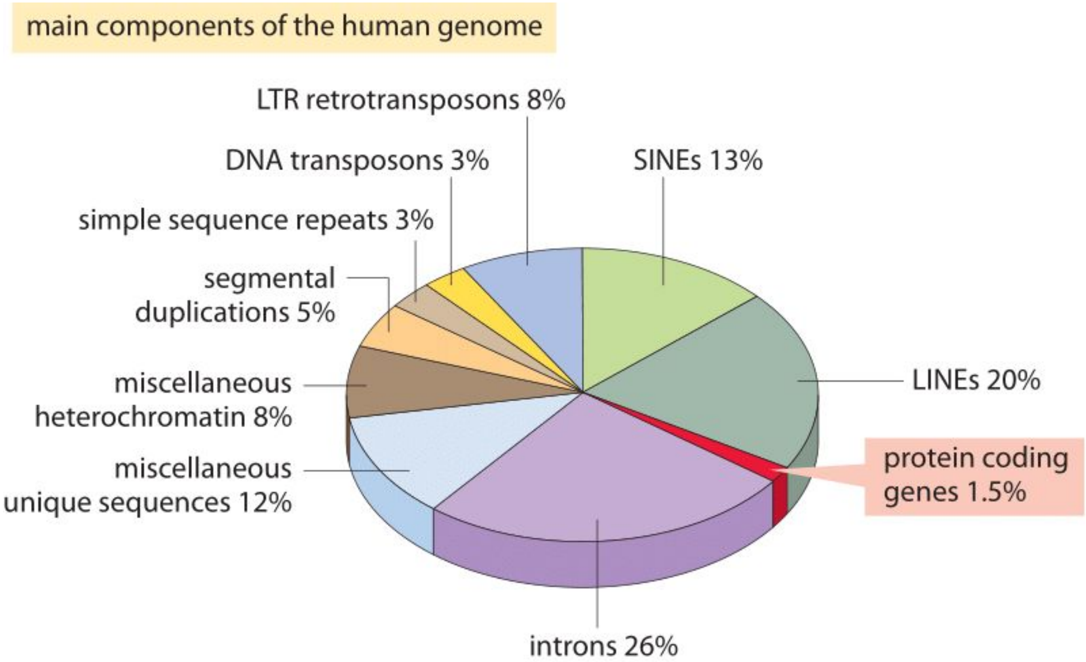
Частота и влияние вариантов

- Большинство сильно-патогенных вариантов - редкие
- Эффект частых вариантов, как правило, очень мал



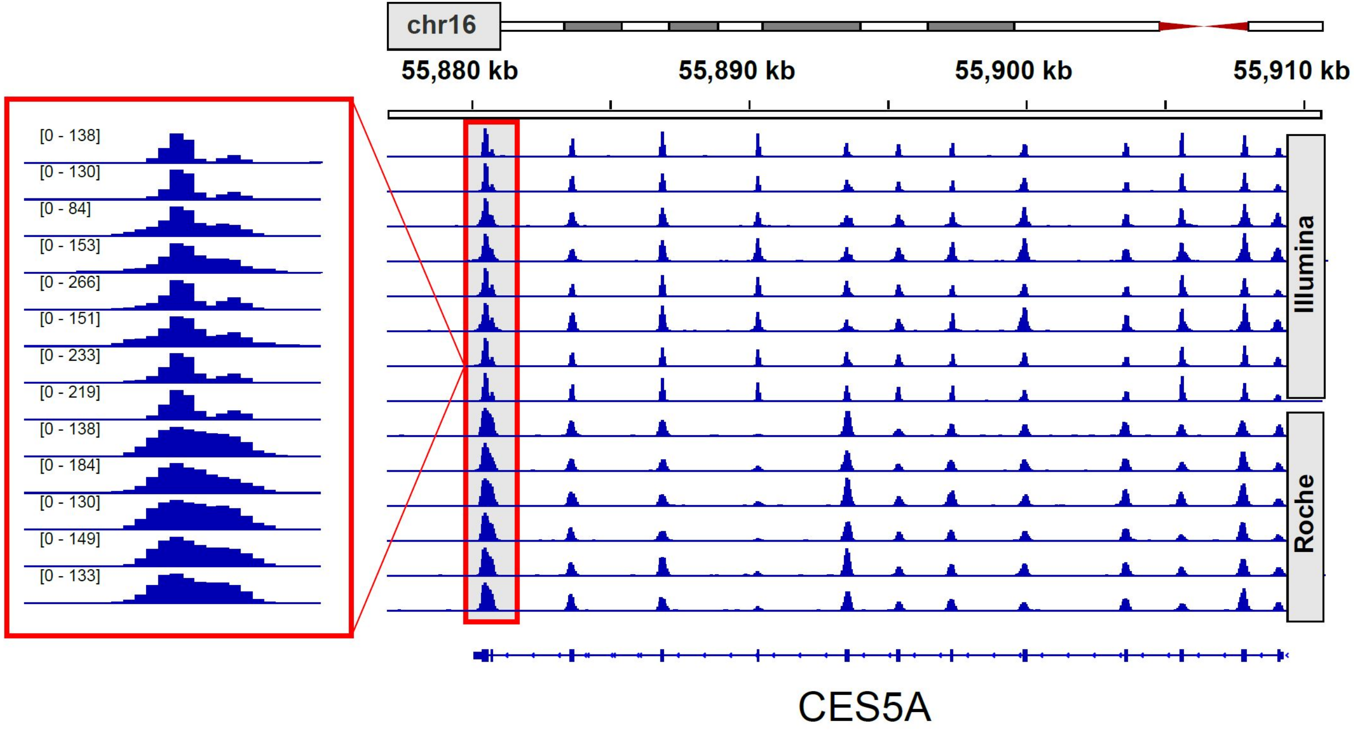
Гены в геноме человека

- кодирующие последовательности составляют всего 1.5% от всего генома (45 Мб из 3 Гб)



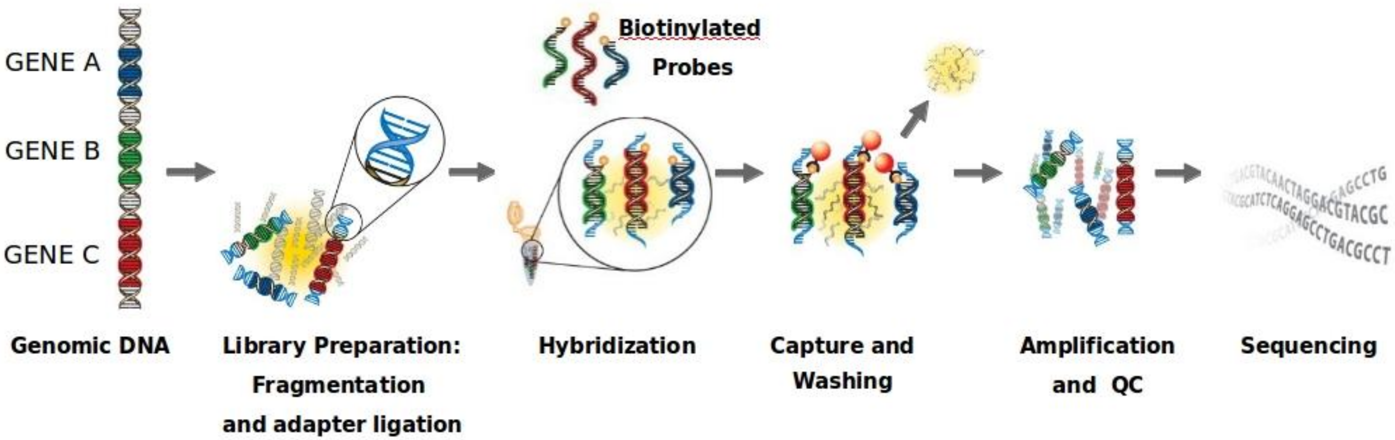
Что такое экзом?

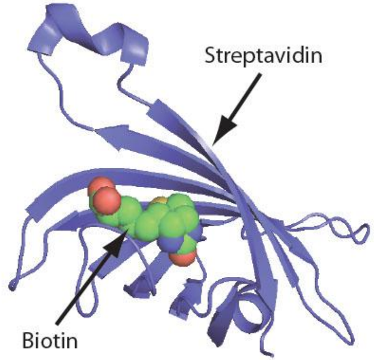
- Фокус только на кодирующих регионах (экзонах)



Обогащение ДНК

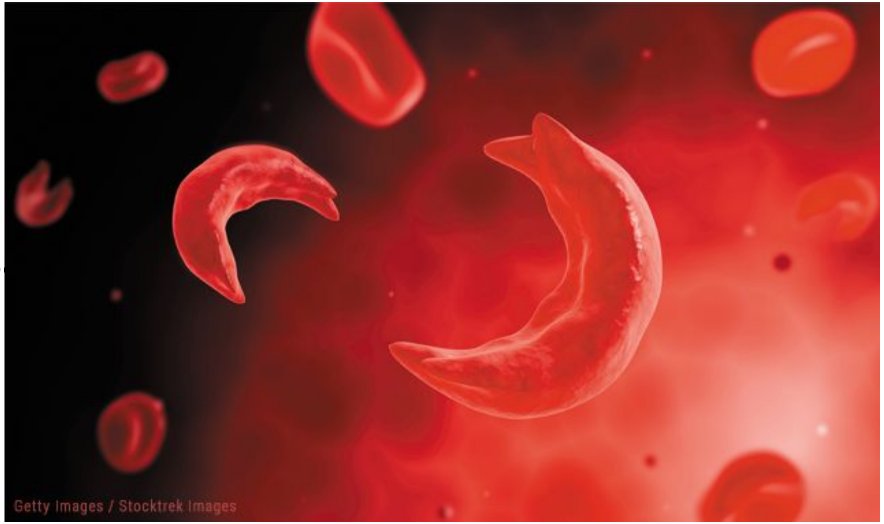
- Выделяются только нужные нам фрагменты
- Набор фрагментов может быть практически любым!



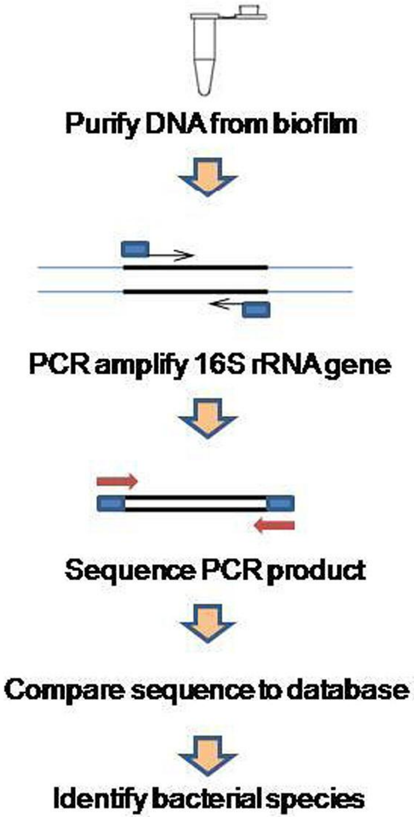


Ресеквенирование ДНК

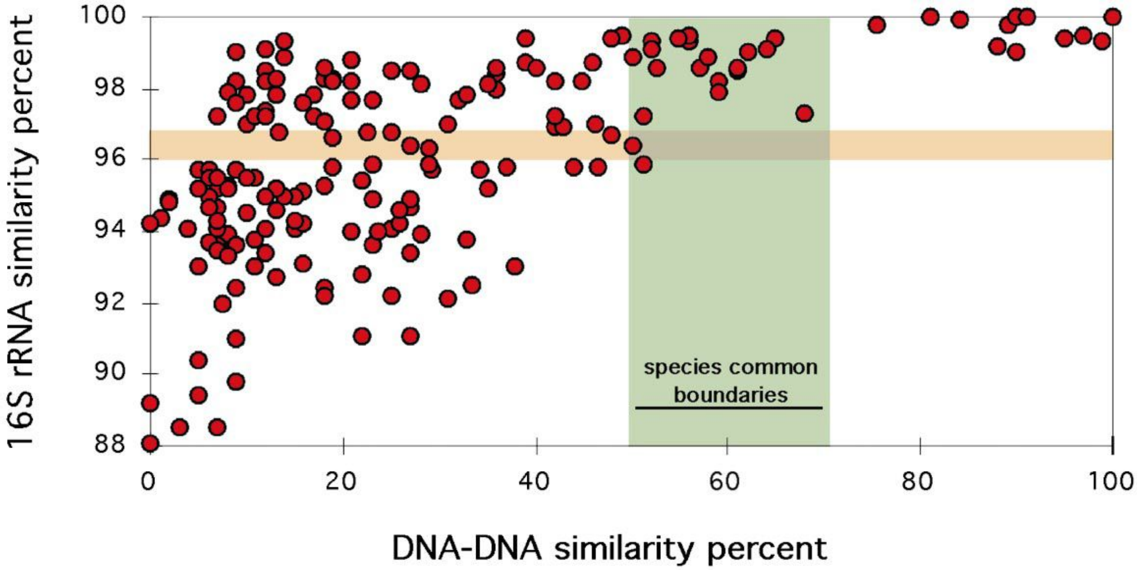
- Таргетная панель (1-500 генов)
- Клинический экзом (5000 важных генов) - CES
- Полный экзом (20000 генов) - WES
- Полный геном - WGS
- Являются необходимостью и де-факто стандартом при генетических (наследуемых) заболеваниях



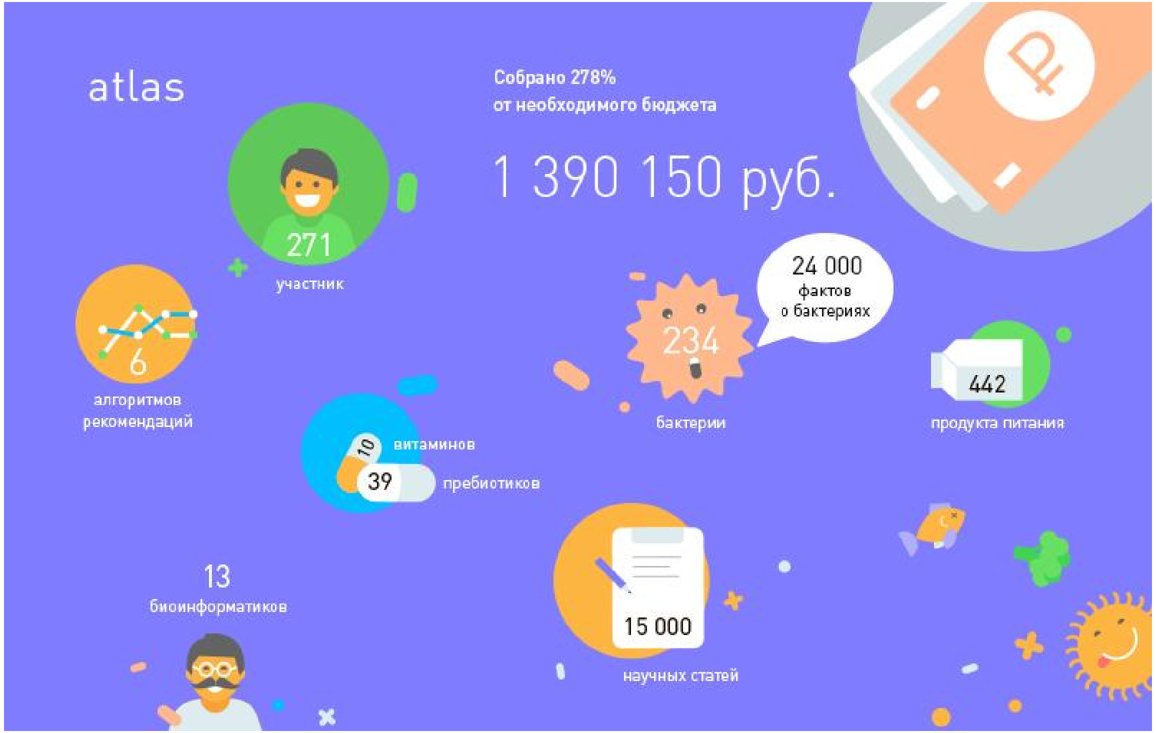
16S-секвенирование



- Секвенирование ампликона

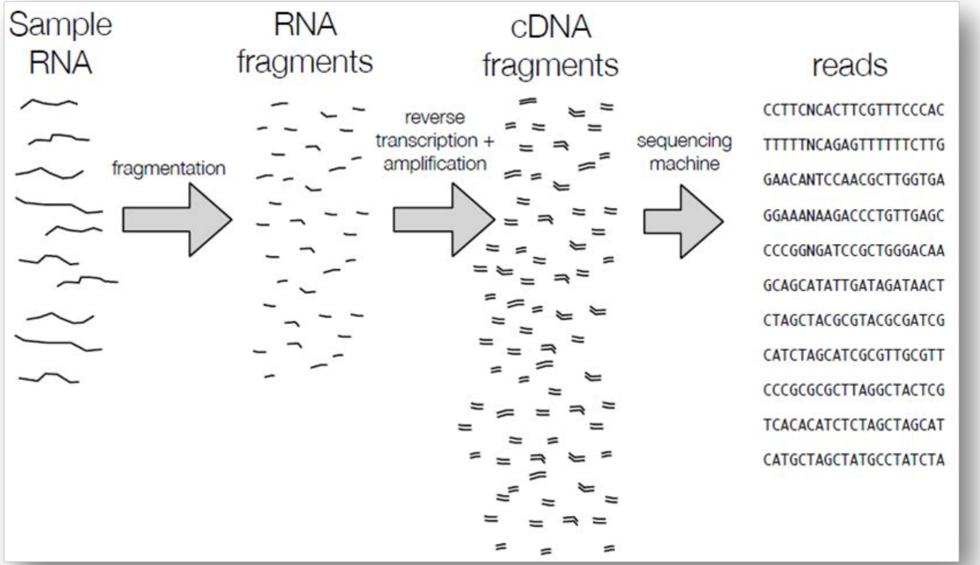


Oh My Gut!


atlas
Собрано 278%
от необходимого бюджета
1 390 150 руб.
271
участник
6
алгоритмов
рекомендаций
24 000
фактов
о бактериях
234
бактерии
442
продукта питания
10
витаминов
39
пребиотиков
13
биоинформатиков
15 000
научных статей

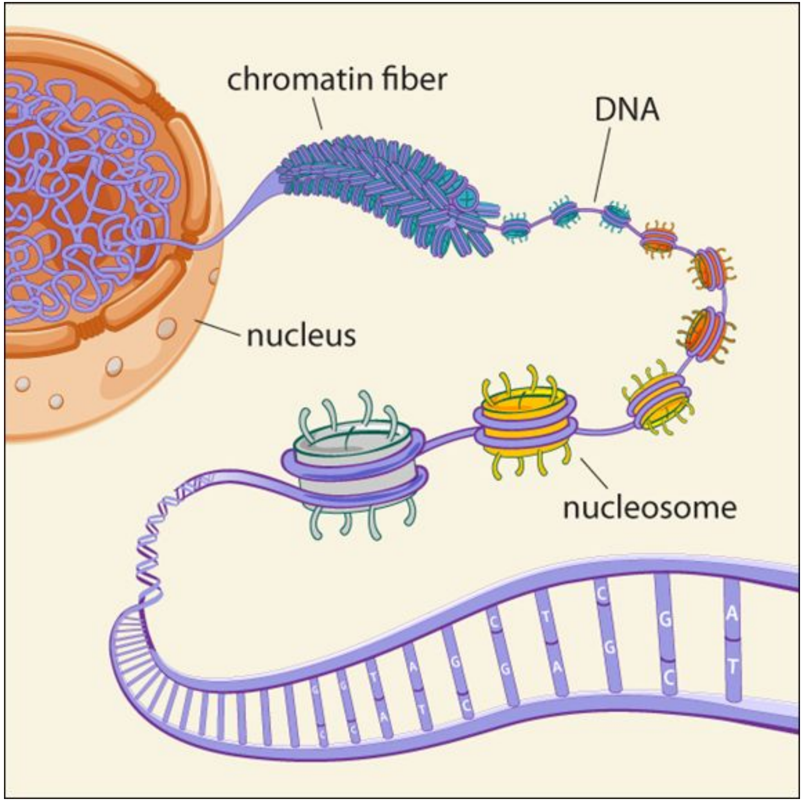
Секвенирование транскриптома

- De novo - если сборка генома слишком трудна, или нужна разница между тканями
- На хорошо аннотированных организмах - связь с фенотипом или условиями
- Цели:
    - Количественный анализ экспрессии
    - Поиск новых транскриптов



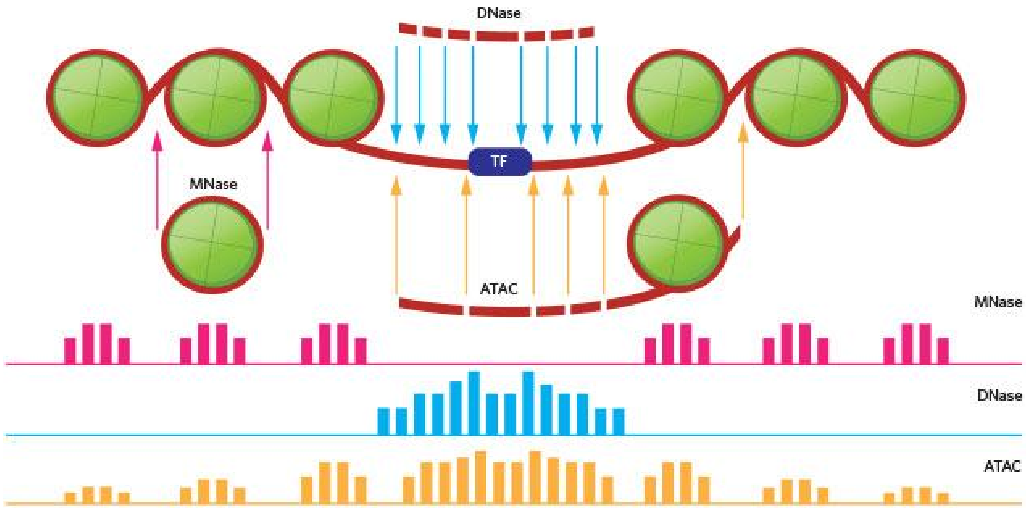
Хроматин в эукариотах

- Три уровня упаковки:
  - ”бусы на нитке” - нуклеосомы
  - 30 нм фибрила
  - Плотно упакованная структура - гетерохроматин
- Нуклеосомы содержат гистонные протеины
- Длина ДНК hs - около 2 м, размер клетки - порядка микрометров



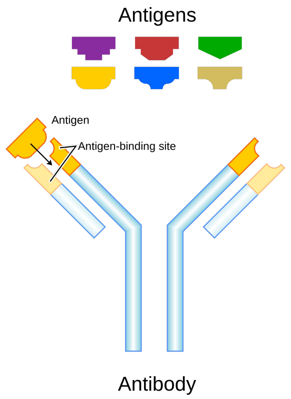
Genome footprinting

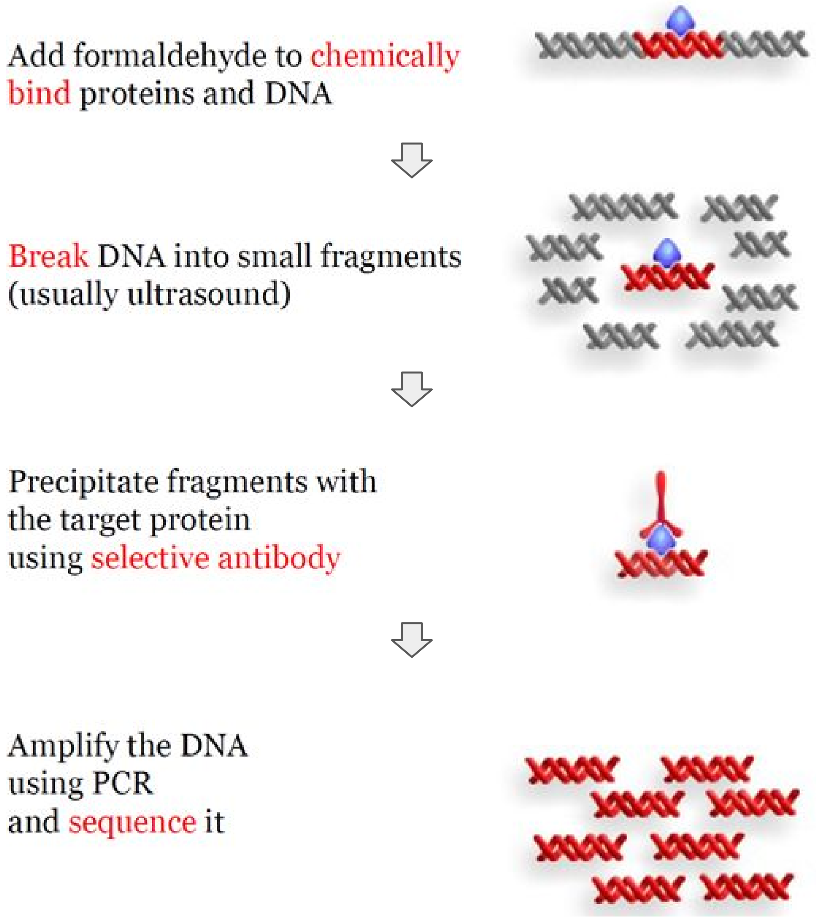
- Смысл: поймать наиболее “развернутые” участки хроматина
- DNase-seq, FAIRE-seq, ATAC-seq



ChIP-Seq

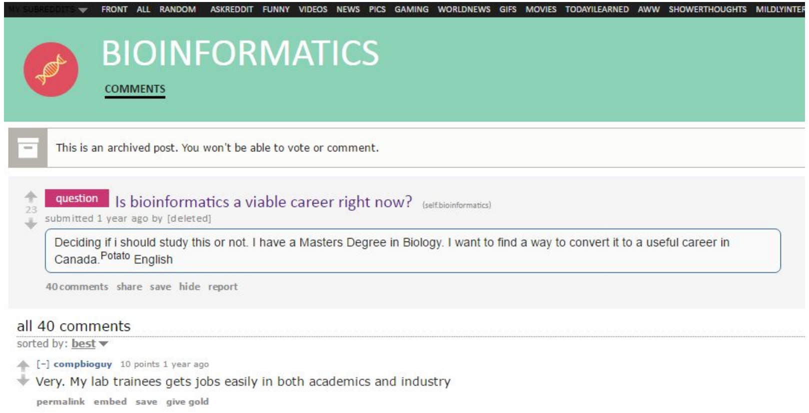
- Смысл: поймать участки ДНК, связанные с определенным протеином
- Делается с помощью селективных антител

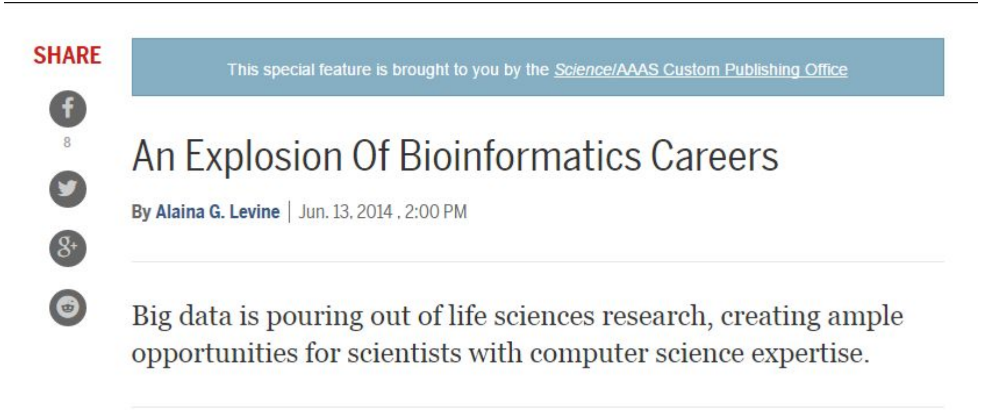




Куда податься биоинформатику?

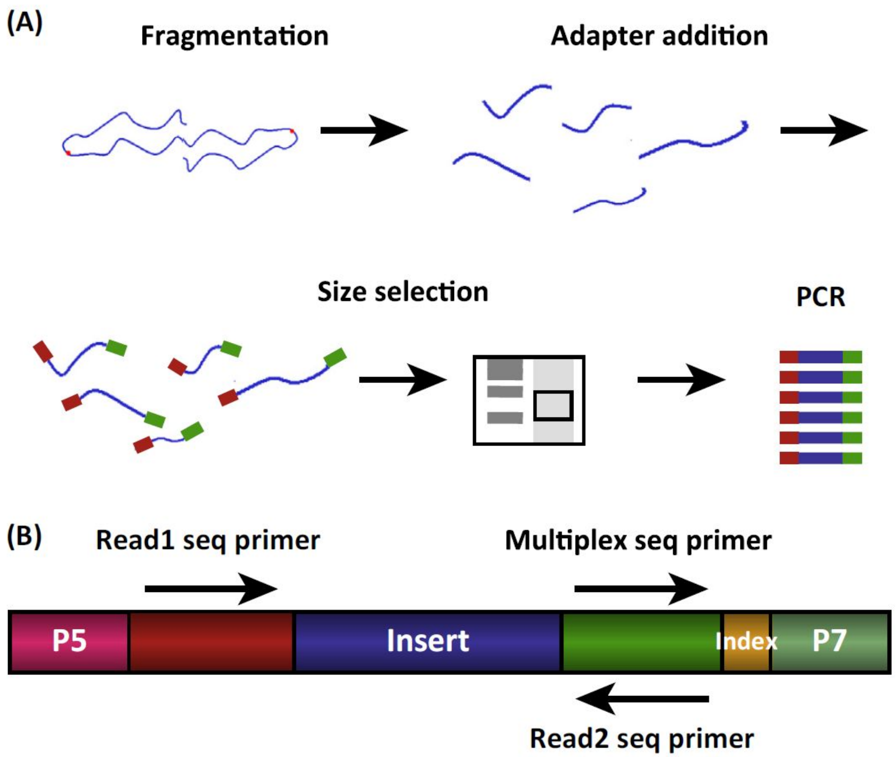
- биоинформатика - крайне быстро растущий рынок
- в том числе и в России
- большинство потребителей услуг - в медицине и науке





Типичный NGS-эксперимент

- фрагментация ДНК
- лигация адаптеров
- выбор фрагментов нужного размера
- амплификация
- секвенирование



Биоинформатика IRL

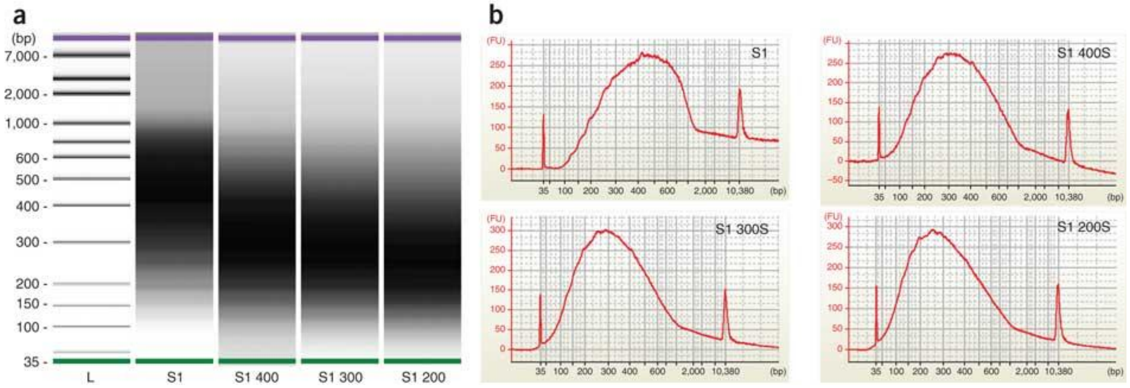
- Практически работа биоинформатика ~ NGS
- Работа с NGS ~ quality control (QC)
- Плюс, грамотное применение подходящих пайплайнов

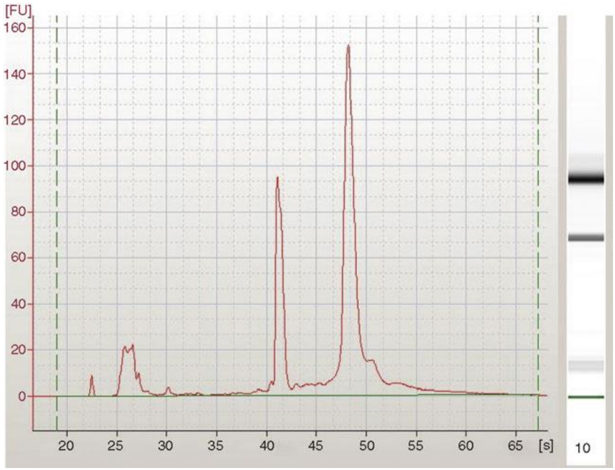
Постадийный QC

- QC эксперимента и приготовления библиотеки
    - Биологические наблюдения (живые клетки, гомогенность)
    - Качество экстракции ДНК/РНК, ее состояние
    - Распределение фрагментов по размеру - Bioanalyzer
    - PCR на нужные фрагменты (до секвенирования)
- QC секвенатора
    - На Illumina - плотность кластеров, basecalls, состояние прибора
- QC полученных прочтений
- QC после сборки и выравнивания

Типичные ранние проблемы

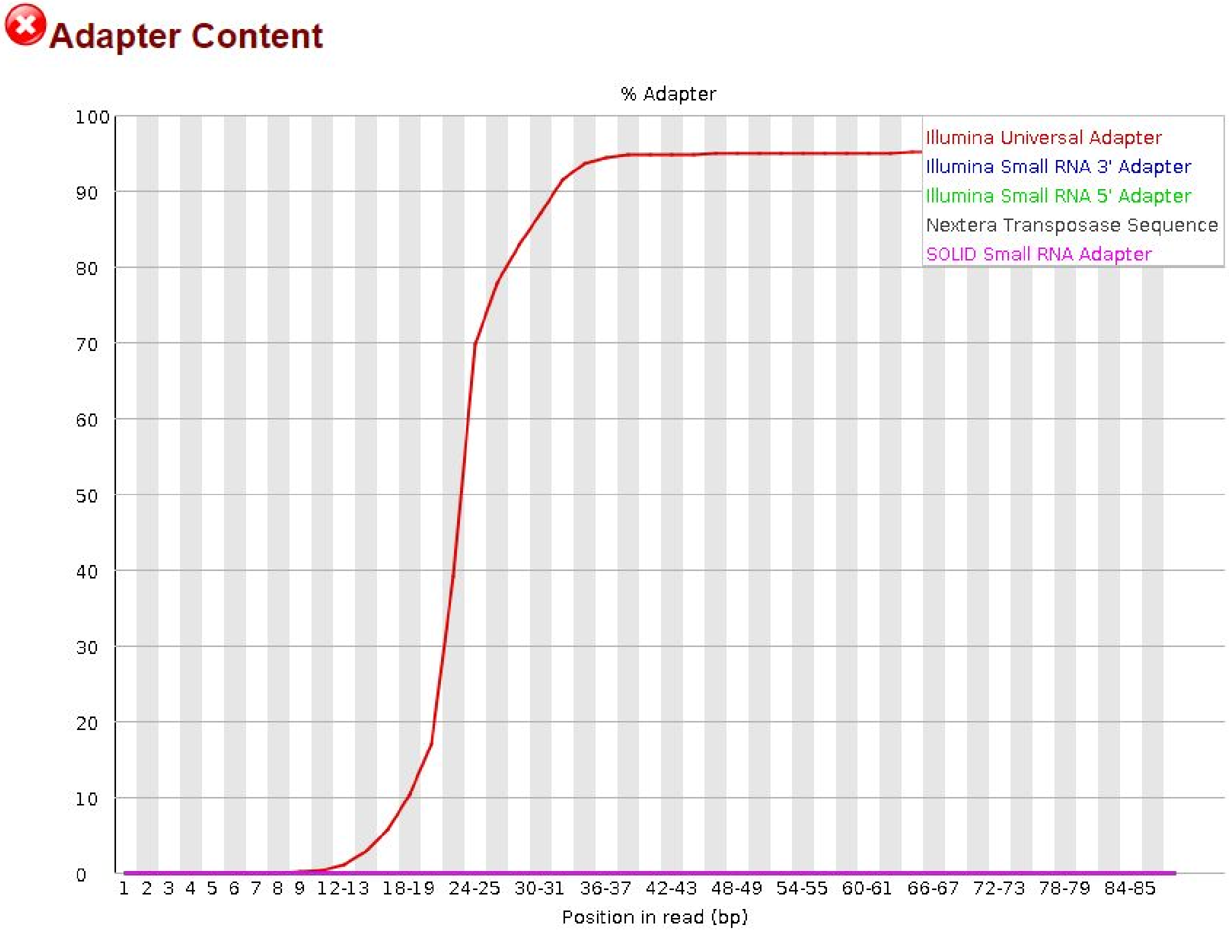
- РНК или ДНК деградировал
- Чрезмерная/недостаточная соникация или дробление энзимом
- Ошибки в количествах реактивов (плохие реактивы)
- Библиотека “низкой сложности” (мало уникальных фрагментов)





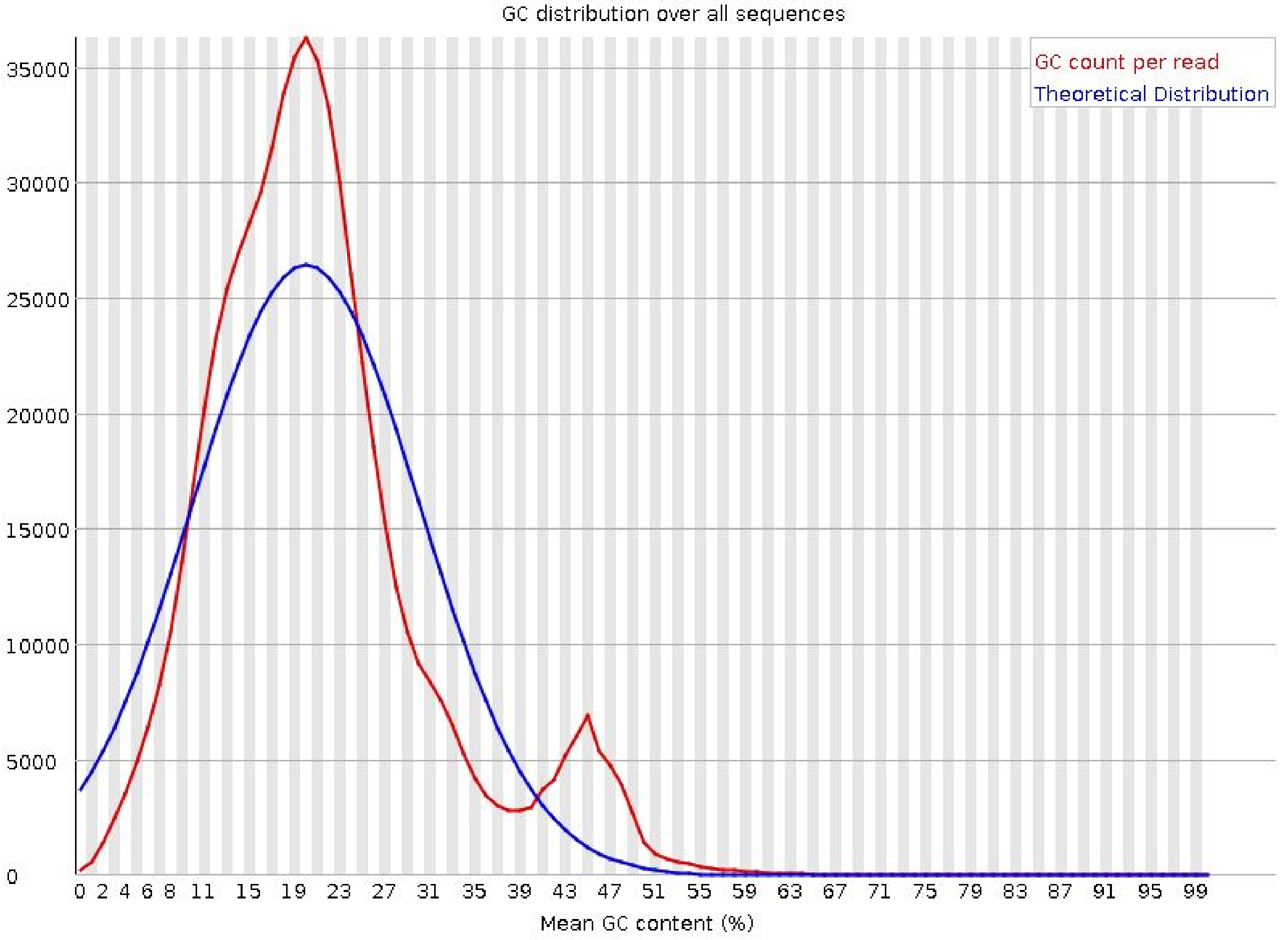
Пример 1

- Слишком короткие фрагменты
- Секвенирование доходит до конца и идет в адаптер



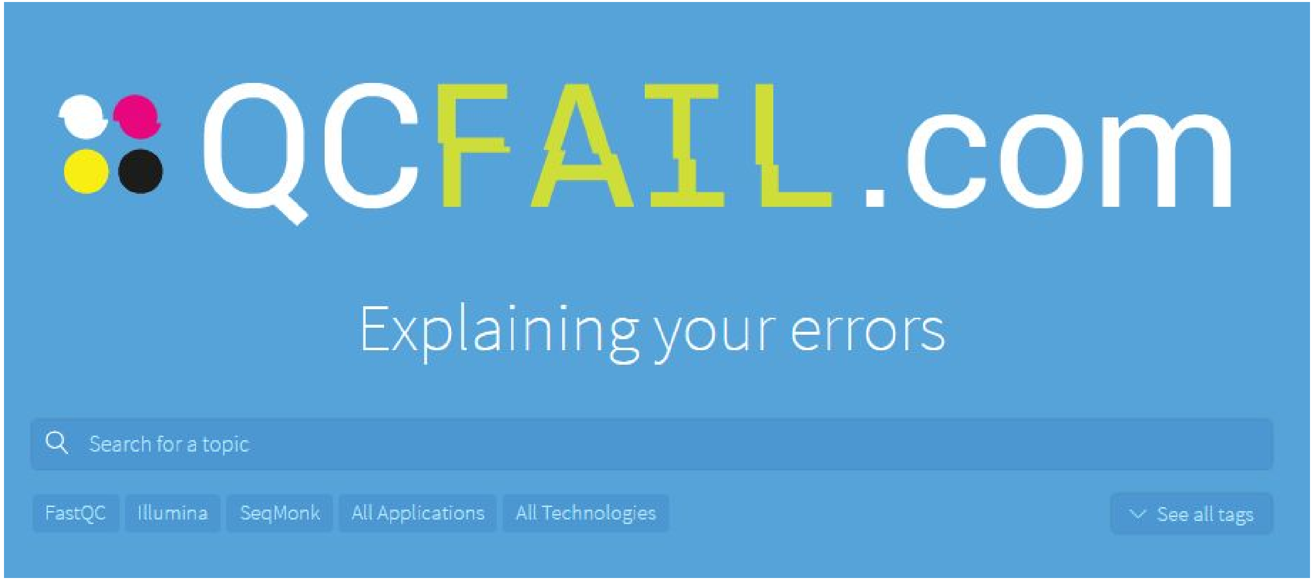
Пример 2

- загрязнение бактериальным геномом



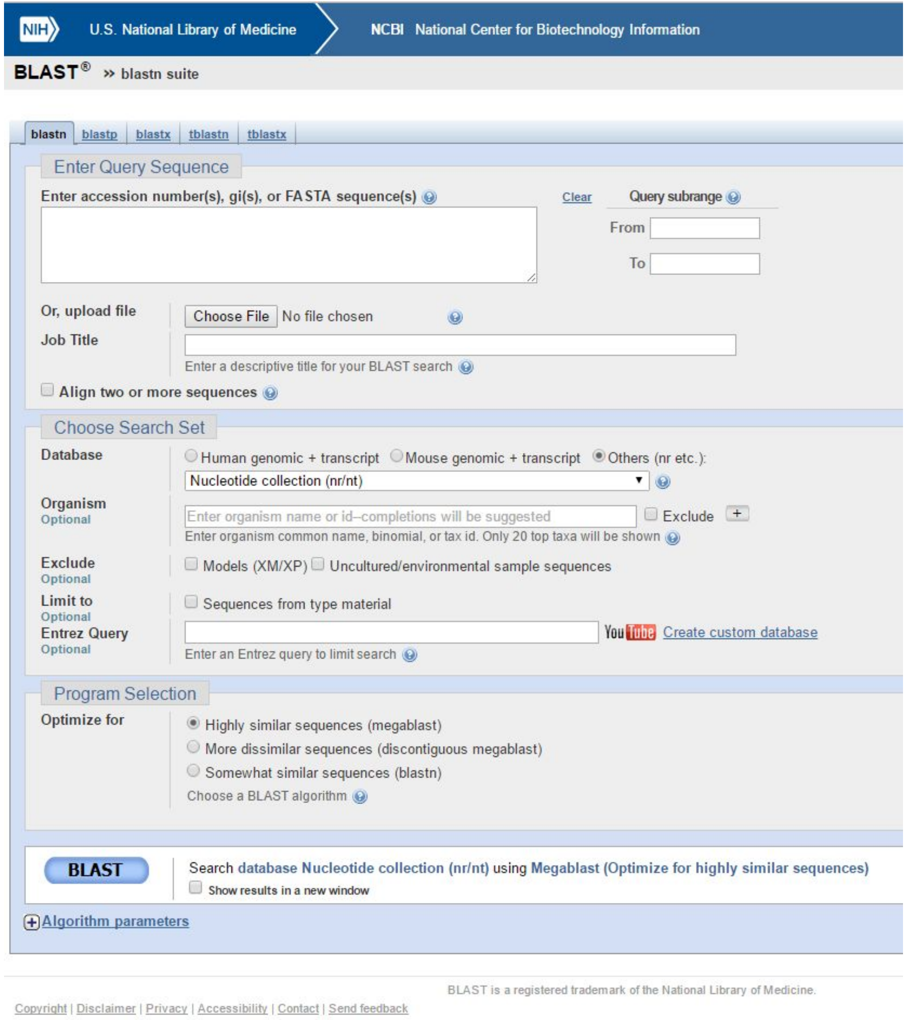
Примеры жизненных неудач

- QCFail: https://sequencing.qcfail.com/



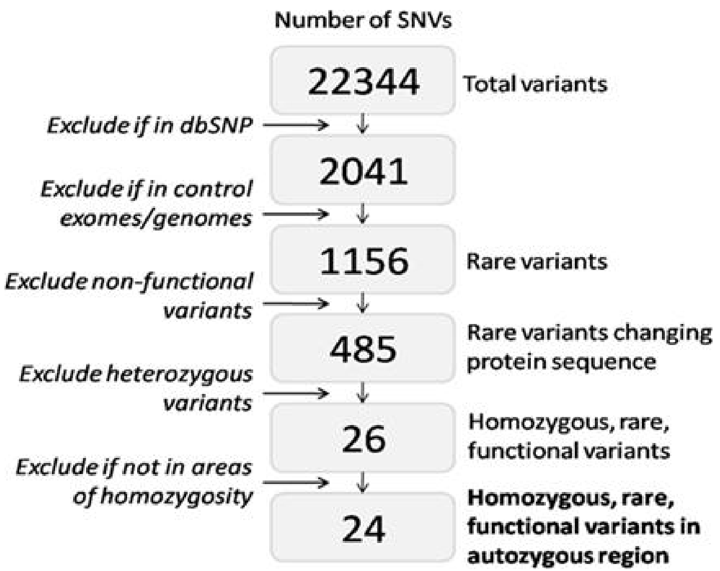
Непонятные прочтения

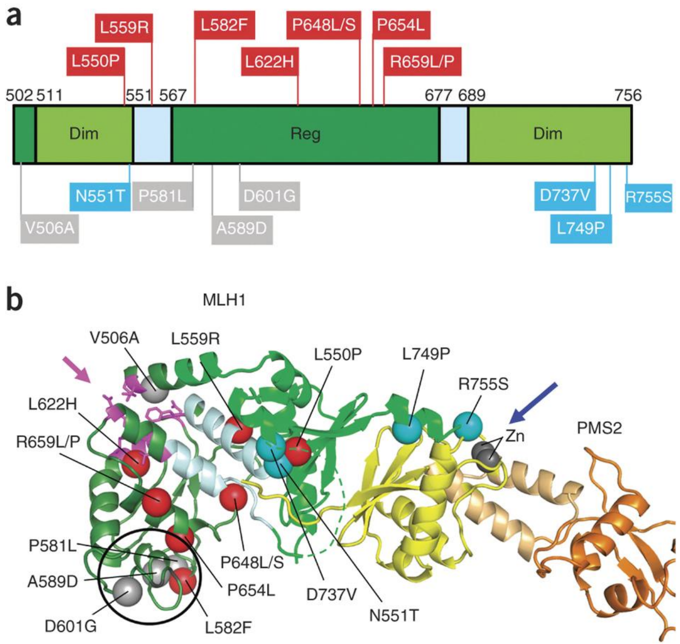
- BLAST - понять, какой биологический вид
- Методика - чтобы понять, откуда они взялись



Чем занимается мед-биоинформатик?

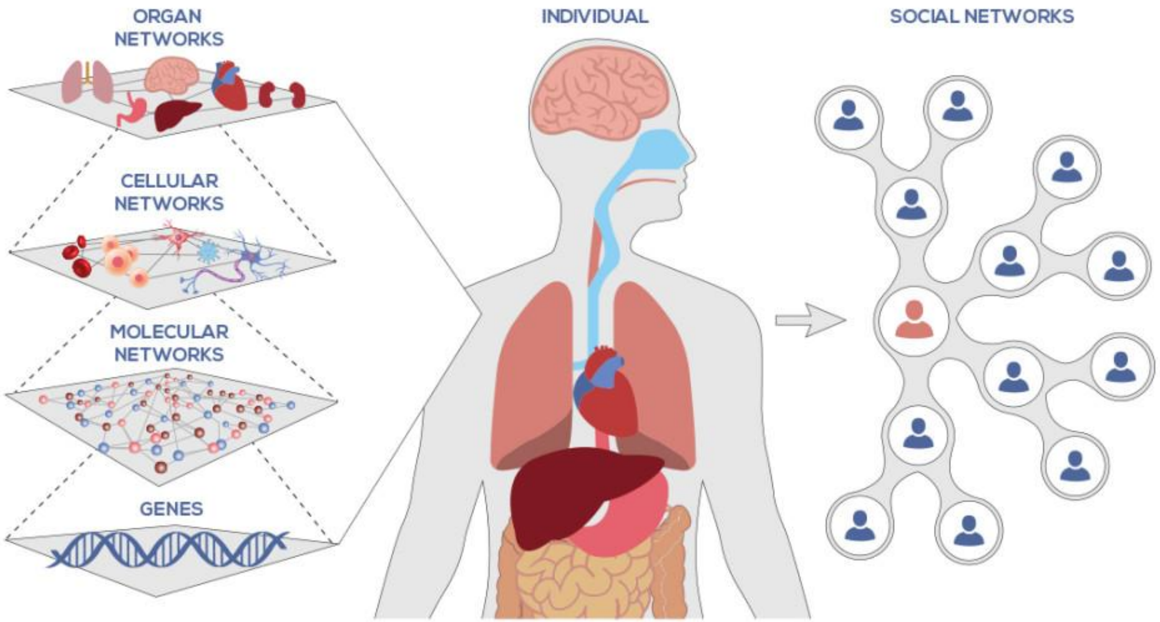
- общение с врачом-генетиком
- контроль качества образцов, анализ покрытия, кандидаты
- анализ и приоритезация мутаций





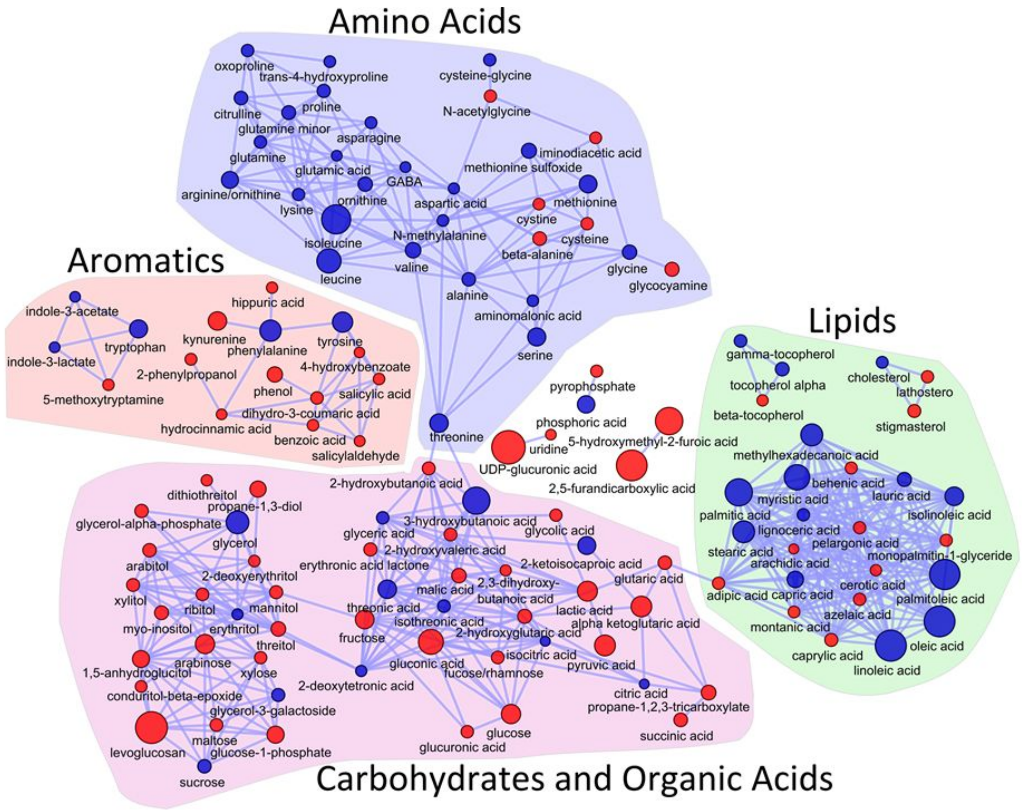
Системная биология

- попытки объединить разрозненные данные в целостную картину
- попытки построить более точные математические модели



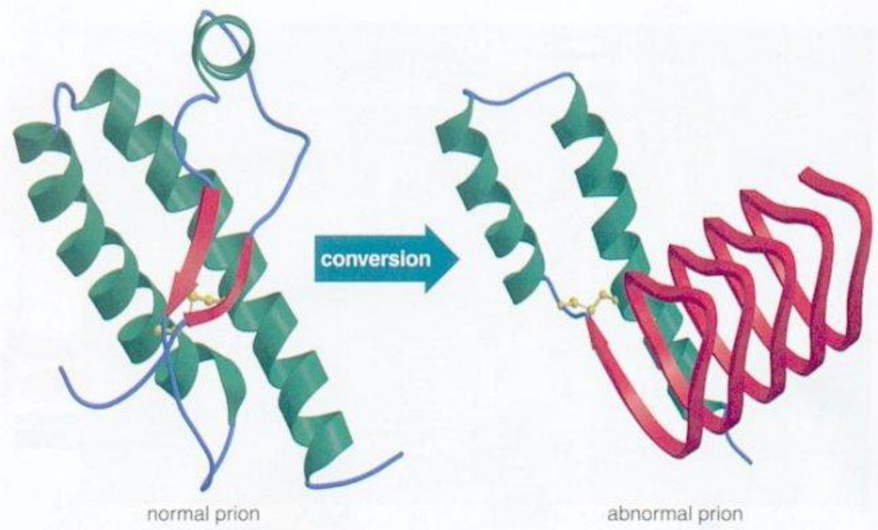
Интеграция омиксных технологий

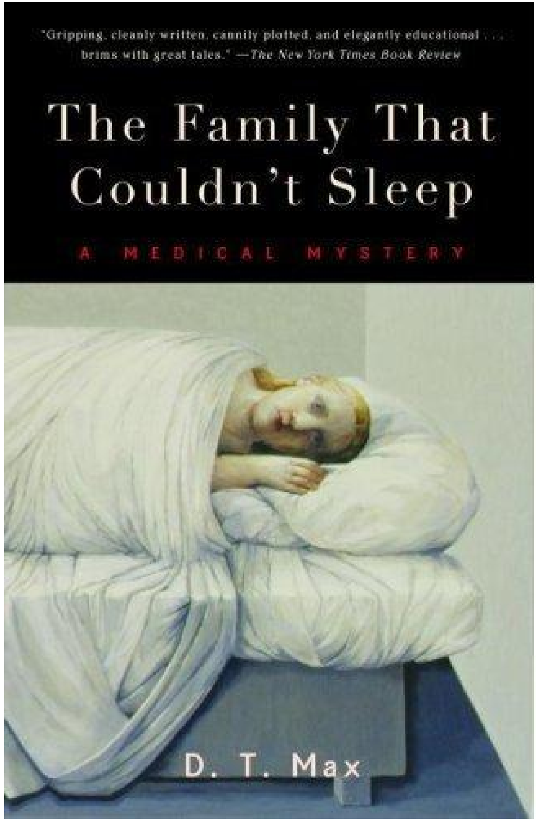
- геномика
- транскриптомика
- протеомика
- метаболомика
  - липидомика
  - гликомика



Fatal Familial Insomnia (FFI)

- аутосомно-доминантное прионное заболевание
- известно всего 25 семей
- знаменитая семья из Венето, страдающая от заболевания уже более 200 лет!





История другой семьи

- Мать умерла от деменции в 52 года
- Дочь оказалась носителем мутации
- Оба бросили свои карьеры (в 2013-м году) и были приняты в аспирантуру Broad Institute
- cureFFI.org

CureFFI.org

About Archives Contact

Does this mean I'll definitely get the disease?

Jan 20, 2016 • ericminikel • Cambridge, MA

I am excited to announce today the publication of our new study, "Quantifying prion disease penetrance using large population control cohorts" [full text, PDF, perspective piece by Robert Green]. I wanted to write a blog post to announce it and to offer up the comments section (bottom of this post) as a place for post-publication peer review. I also wanted to take this opportunity to tell the history of this study, a profoundly personal endeavor that has grown and taken shape over the entire four year course of my re-training as a scientist.

На данный момент

- Организовали небольшую лабораторию в Broad Institute
- Собрали деньги на клинические испытания
- Успели еще очень много всякого!

Спасибо за внимание!