федеральное государственное автономное образовательное учреждение высшего образования
Первый Московский государственный медицинский университет им. И.М. Сеченова
Министерства здравоохранения Российской Федерации
(Сеченовский Университет)

Институт фармации им. А.П. Нелюбина
Кафедра биотехнологии

Методические материалы по дисциплине:

Геномика и протеомика

основная профессиональная образовательная программа высшего профессионального образования - программа специалитета

06.05.01 Биоинженерия и биоинформатика

ФГАОУ ВО Первый МГМУ им. И. М. Сеченова Минздрава России (Сеченовский Университет)

Кафедра биотехнологии ИФ

Москва, 2025

Протеом и аминокислоты, источники, методы выделения и очистки белков

Протеомика

- **Протеомика** – область молекулярной биологии, изучающая белки организмов, их количественный анализ, строение, функции и взаимодействие

- **Объект изучения** протеомики – **протеом –** совокупность всех белков, которые экспрессируются в данной клетке, ткани, организме в данный момент времени

Белок

Белок – высокомолекулярный полимер, состоящий из соединенных пептидными связями 20 α-аминокислот (преимущественно в L форме) , с числом аминокислотных остатков от 50 и массой не менее 10 000 Да

Например, рибонуклеаза – 13 700 Да, коннектин – 2 993 442.763 *Да*

	Molecular weight	*Number of residues*	*Number of polypeptide chains*
Cytochrome *c* (human)	13,000	104	1
Ribonuclease A (bovine pancreas)	13,700	124	1
Lysozyme (chicken egg white)	13,930	129	1
Myoglobin (equine heart)	16,890	153	1
Chymotrypsin (bovine pancreas)	21,600	241	3
Chymotrypsinogen (bovine)	22,000	245	1
Hemoglobin (human)	64,500	574	4
Serum albumin (human)	68,500	609	1
Hexokinase (yeast)	102,000	972	2
RNA polymerase (*E. coli*)	450,000	4,158	5
Apolipoprotein B (human)	513,000	4,536	1
Glutamine synthetase (*E. coli*)	619,000	5,628	12
Titin (human)	2,993,000	26,926	1

Виды белков (1)

Группа белков	Функция	Структурные особенности	Примеры
Ферменты	катализируют более 4 000 биохимических реакций, часто с высокой специфичностью и огромной эффективностью	активный сайт может иметь всего несколько аминокислотных остатков, находящихся в непосредственном контакте с субстратом, и обычно только 3-4 остатка участвуют в реальном катализе	ферменты энергопродуцирующих циклов; ферменты, участвующие в поддержании и передаче генетической информации в клетке; пищеварительные ферменты
Структурные белки structural proteins	организуют внутриклеточные структуры; внеклеточно обеспечивают механическую поддержку клеткам и тканям	часто многосубъединичные белки, в которых отдельные субъединицы взаимодействуют друг с другом, образуя волокна	внутри клеток тубулин образует микротрубочки; актин образует актиновые нити, поддерживающие плазматическую мембрану; в ядре гистоны образуют октамерные белковые ядра, вокруг которых ДНК обматывается, образуя структуры, известные как нуклеосомы

Виды белков (2)

Группа белков	Функция	Структурные особенности	Примеры
Строительные белки scaffold proteins	удерживают вместе белки, которые являются частью сигнального или каталитического пути	многодоменные; многие, по-видимому, эволюционно различные типы	Hsp70 и Hsp90 регулируют последовательное сворачивание белком
Транспортные белки	переносят небольшие молекулы или ионы; те, которые встроены в мембраны, переносят молекулы через мембраны	изменение конформации участка связывания обычно сопровождает связывание перемещаемой молекулы или иона	альбумин в крови переносит липиды; гемоглобин в эритроцитах переносит кислород; трансферрин переносит железо; белковые кальциевые насосы переносят Ca^{2+} в мышечные клетки, чтобы вызвать сокращение мышц

Виды белков (3)

Группа белков	Функция	Структурные особенности	Примеры
Белки для хранения	служат депо для хранения малых молекул и ионов в некоторых клетках или хранилищами аминокислот для синтеза других белков	часто напоминают ферменты, имея несколько мест связывания, но не обладают каталитической функцией	овальбумин в яичном белке птиц и казеин в молоке млекопитающих являются источниками аминокислот для эмбрионов или новорожденных
Сигнальные белки	обеспечивают связь между клетками и внутри них	часто высокоспецифичны в своих взаимодействиях с клеточными поверхностями или внутриклеточными структурами	инсулин контролирует уровень глюкозы в крови; эпидермальный фактор роста стимулирует рост клеток и деление в эпителиальных клетках

Виды белков (4)

Группа белков	Функция	Структурные особенности	Примеры
Рецепторные белки	обычно мембранные белки, которые улавливают сигналы (из окружающей среды или в процессе развития) и передают их внутрь клетки, чтобы вызвать соответствующие клеточные реакции	обычно димеризуются в мембране в ответ на сигналы	рецептор инсулина опосредует клеточный ответ на глюкозу, взаимодействуя с инсулином
Регуляторные белки	регулируют многочисленные реакции, такие как транскрипция, связываясь с ДНК или белками транскрипционного механизма, и трансляция, связываясь с компонентами трансляционного механизма	обычно они содержат как участки для связывания с мишенью, такой как ДНК, РНК или белки, так и участки, связывающие регуляторы	репрессор лактозы у бактерий связывается с участками ДНК, кодирующими ферменты, которые участвуют в утилизации лактозы; многочисленные факторы транскрипции стимулируют транскрипцию специфических генов у эукариот в ответ на воздействие окружающей среды

Виды белков (5)

Группа белков	Функция	Структурные особенности	Примеры
Прочие белки с высокоспецифичной функцией	функции очень разнообразны в зависимости от организма и окружающей среды	часто имеют необычные аминокислоты	белки, защищающие от замерзания организмы, живущие в морозной среде; стрессовые белки, позволяющие организмам переносить высокие температуры или соленость; белки клея, прикрепляющие морские организмы к скалам; белки, выполняющие защитные функции против чужеродных организмов, например, антитела

Аминокислоты

α - аминокислоты - производные карбоновых кислот, у которых водородный атом у α - углерода замещен на аминогруппу -NH2

β α

$R\text{-}CH_2\text{-}CH_2\text{-}COOH$

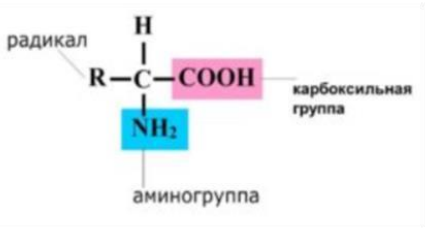
жирная кислота

α

$R - CH_2 - CH(NH_2) - COOH$

α- аминокислота

α- аминокислоты
- из базовой формулы и радикала



Аминокислоты

Три отличительных характеристики протеиногенных аминокислот:

1. Радикал расположен у α-углеродного атома
2. α-аминокислоты имеют L-конфигурацию
3. Почти половина протеиногенных аминокислот не синтезируется в организме, а должна поступать с пищей

Число аминокислот = n	Возможное число пептидов = n!
2	2
4	24
10	3 628 800
20	$2 \cdot 10^{12}$

Аминокислоты

Cα асимметричен, т.к с ним связаны 4 разные химические группы , поэтому для аминокислоты существуют две возможные конфигурации (энантиоизомеры). Почти все встречающиеся в природе α-аминокислоты имеют L-конфигурацию

D-аминокислоты могут образовываться при посттрансляционной модификации (опиоидные D-метионин и D-аланин входят в состав гептапептидов кожи филломедуз, некоторые пептидные антибиотики бактериального происхождения и т.д.)

D-аминокислоты входят в состав пептидов и их производных, образующихся путём нерибосомного синтеза в клетках грибов и бактерий.

D-аминокислоты могут образовываться в структурных белках при неферментативной рацемизации в процесе старения (D-аспартат)

Аминокислоты - классификация

По путям биосинтеза

Семейство аспартата:

аспартат, аспарагин, треонин, изолейцин, метионин, лизин

Семейство глутамата:

глутамат, глутамин, аргинин, пролин

Семейство пирувата: *аланин, валин, лейцин*

Семейство серина: *серин, цистеин, глицин*

Семейство пентоз: *гистидин, фенилаланин, тирозин, триптофан*

Семейство шикимата иногда включает *фенилаланин, тирозин, триптофан*

Аминокислоты - классификация

По радикалу (!) – имеет значение для изолирования

Нейтральный : *глицин*

Неполярные: *аланин, валин, изолейцин, лейцин, пролин*

Полярные незаряженные (заряды скомпенсированы) при pH=7 :
серин, треонин, цистеин, метионин аспарагин, глутамин

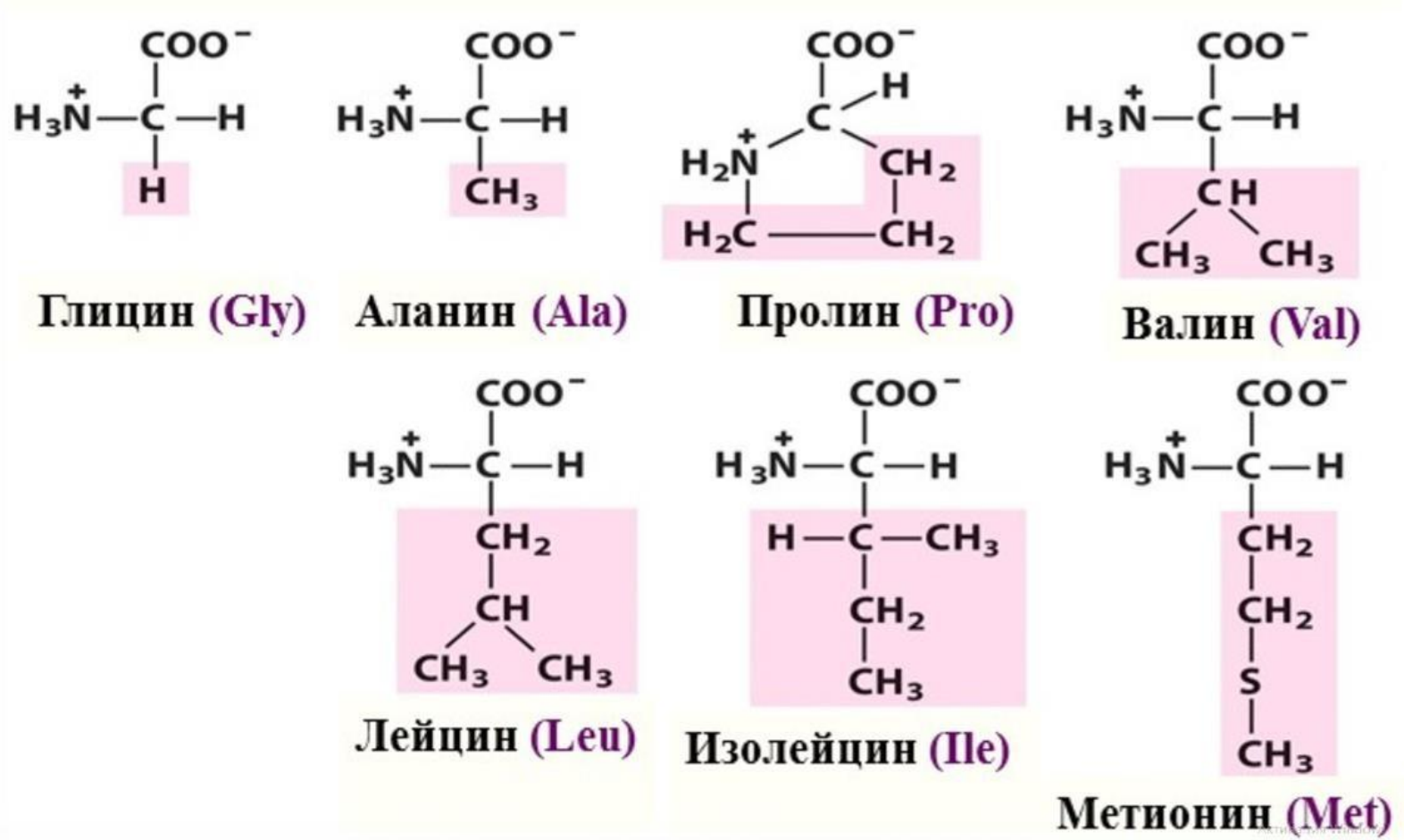
Ароматические: *фенилаланин, триптофан, тирозин*

Полярные заряженные отрицательно при pH=7: *аспартат, глутамат*

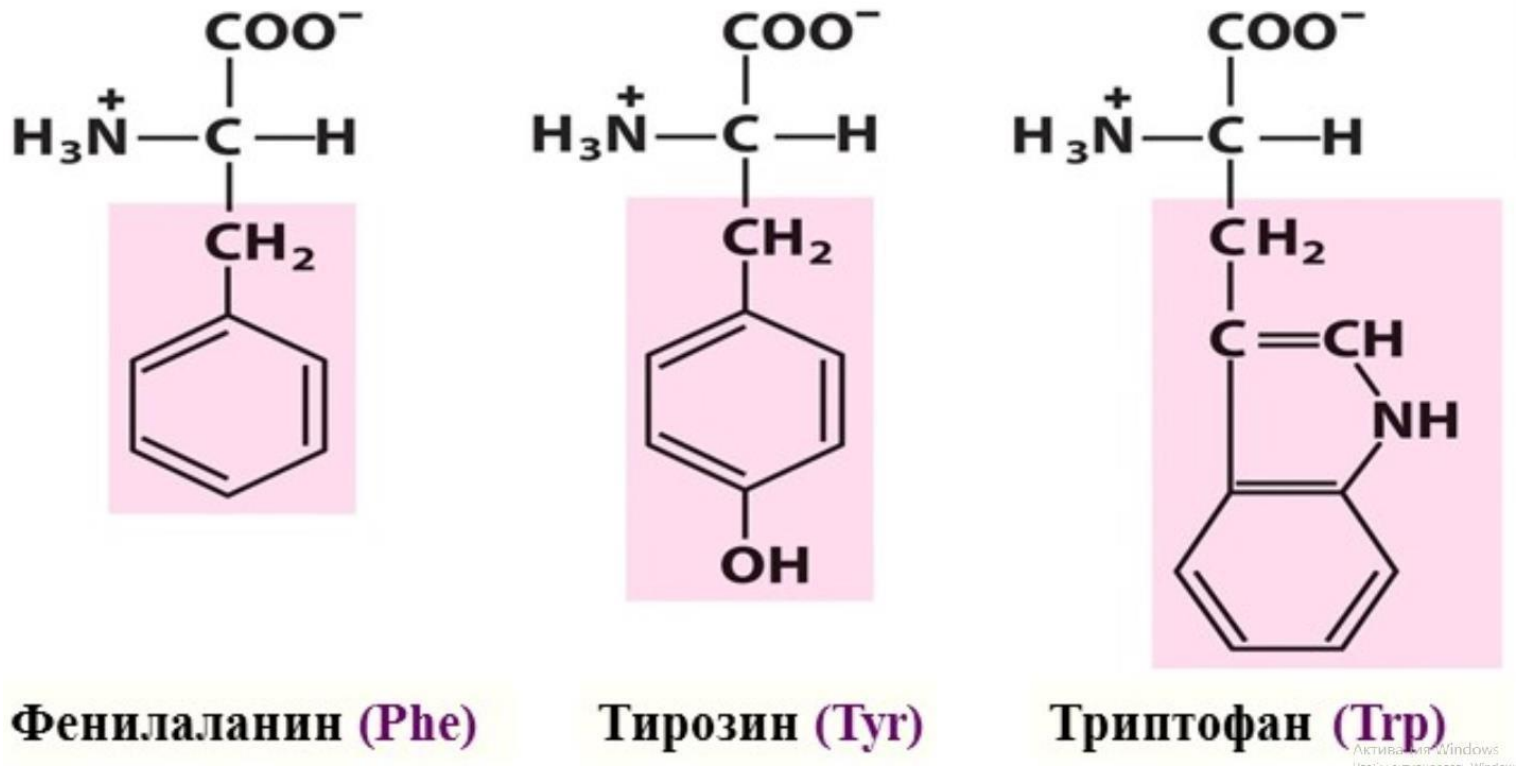
Полярные заряженные положительно при pH=7: *лизин, аргинин, гистидин*

Далее изображены их формулы

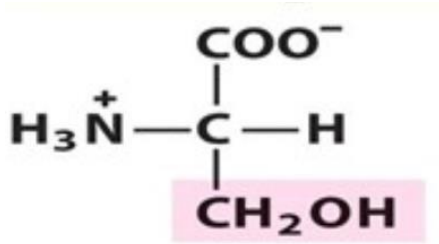
Протеиногенные аминокислоты с *неполярными алифатическими* радикалами



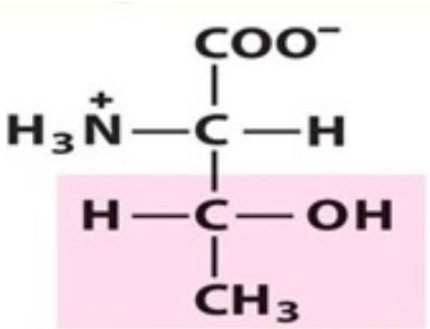
Протеиногенные аминокислоты с *ароматическими* радикалами



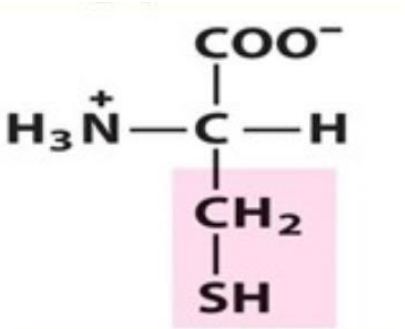
Протеиногенные аминокислоты с *полярными незаряженными* радикалами



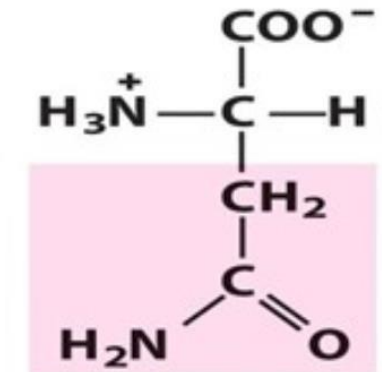
Серин (Ser)



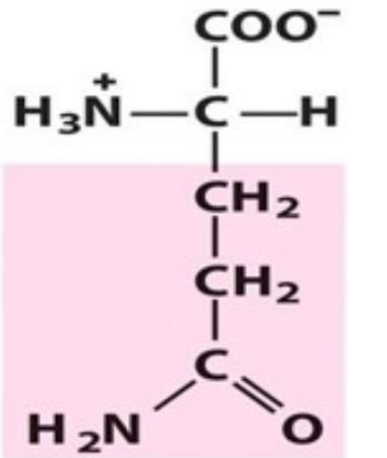
Треонин (Thr)



Цистеин (Cys)

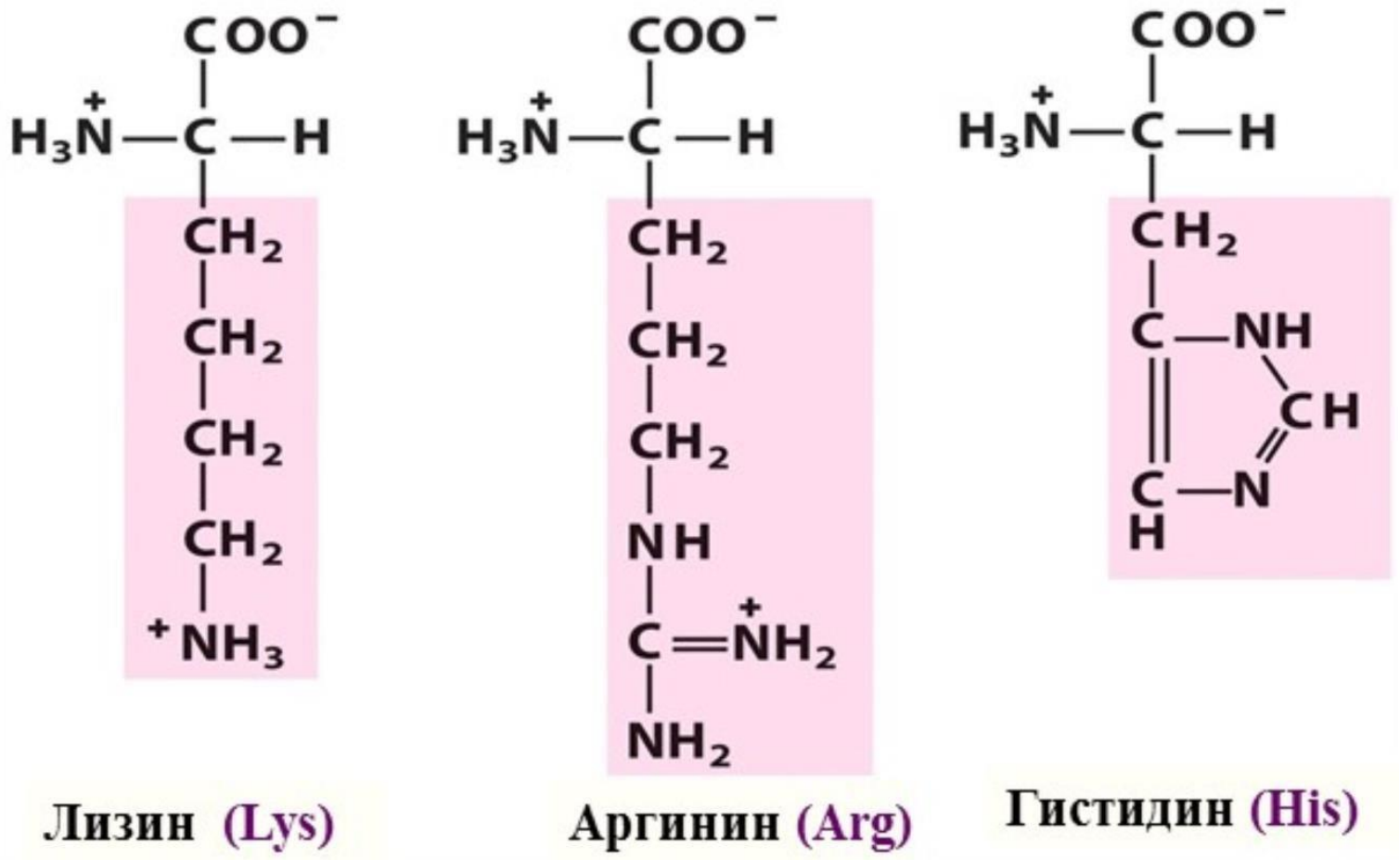


Аспарагин (Asn)

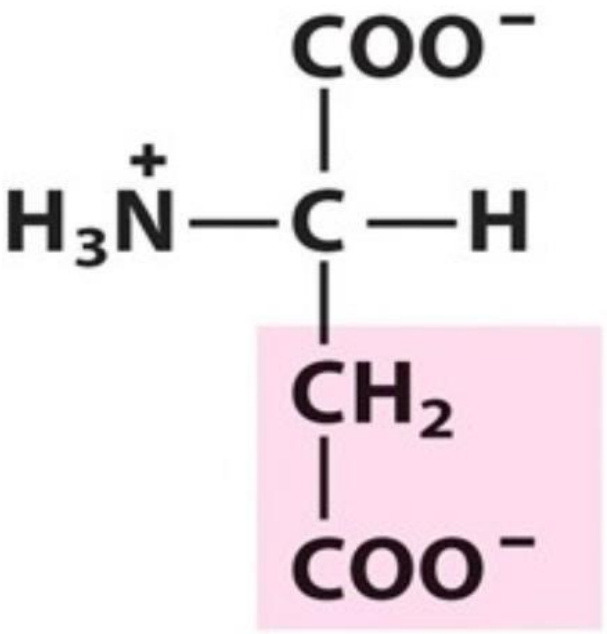


Глутамин (Gln)

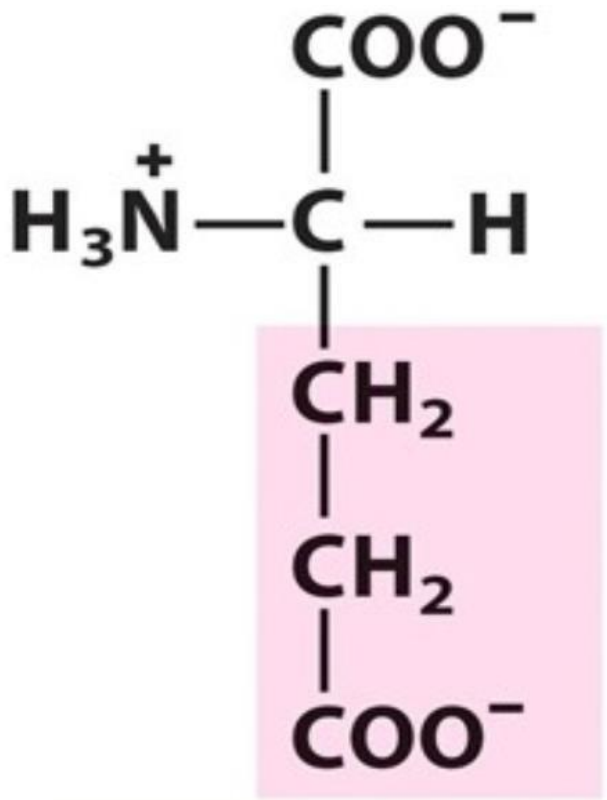
Протеиногенные аминокислоты с *положительно заряженными* радикалами



Протеиногенные аминокислоты с *отрицательно заряженными* радикалами



Аспартат (Asp)

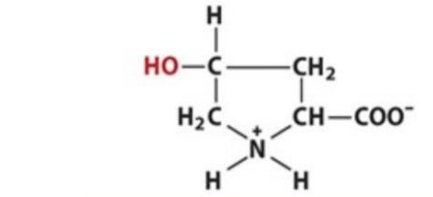


Глутамат (Glu)

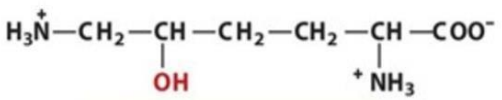
Нестандартные протеиногенные аминокислоты

Не встречаются во всех организмах, но обнаружены в составе некоторых белков

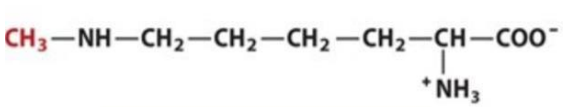
Включаются в состав белков во время биосинтеза (селеноцистеин и пирролизин) и в результате посттрансляционной модификаци (4-гидроксипролин, 5-гидроксилизин, десмозин, N-метиллизин, цитруллин и D-изомеры стандартных аминокислот)



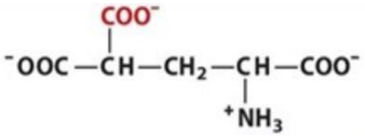
4-Гидроксипролин



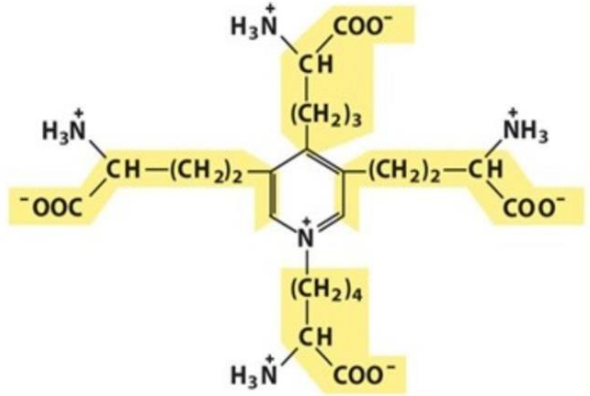
5-Гидроксилизин



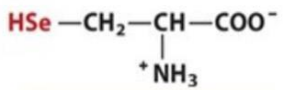
6-N-Метиллизин



γ-Карбоксиглутамат



Десмозин



Селеноцистеин

Рекомбинантные и природные белки

- Преимущества рекомбинантных белков
    1. Могут быть получены в больших количествах
    2. Нет нужды в выделении из опасных/патогенных объектов
    3. Есть контроль над структурой белка -> предсказуемость процесса выделения и очистки
- Преимущества природных белков
    1. Также могут быть получены в больших количествах (напр., BSA – Бычий сывороточный альбумин, его концентрация – 33-55 мг/мл плазмы!)
    2. Сохраняются пост-трансляционные модификации
- Общий подход к получению рекомбинантных белков: изолировать ДНК, кодирующий целевой белок; клонировать кодирующую последовательность; запустить экспрессию данной последовательность в продуценте

Гетерологичная экспрессия в клетках *E. coli*

- Гетерологичная экспрессия – экспрессия кодирующей последовательности, которая (последовательность) не характерна для данного организма, то есть была искусственно добавлена
- В клетках *E. coli* выделяют несколько типов экспрессии белков:
- **Тельца включения** – целевой белок в очень высокой концентрации образует нерастворимые агрегаты в цитоплазме, поскольку «фолдинговые» возможности клетки ограничены.
- Белок из телец включения извлекают по схеме: *центрифугирование* (500-1000g), *солюбилизация* (растворение) белка с использованием детергентов/хаотропных агентов (гуанидин, мочевина)/восстанавливающих агентов (предотвращают окисление – меркаптоэтанол), *удаление компонентов буфера для солюбилизации* – диализ, диафильтрация, гель-хроматография
- **Выделение во внешнюю среду** – низкий выход (!) - < 1г/л, при внутриклеточ. – 5-10 г/л

Гетерологичная экспрессия в клетках *E. coli*

Достоинства	Недостатки
E. coli хорошо охарактеризована, что позволяет легко проводить генетические манипуляции	Низкая способность выделять рекомбинантные белки внеклеточно в больших количествах
Доступны экспрессионные векторы, обеспечивающие простую и высокоуровневую экспрессию рекомбинантных белков	Рекомбинантный белок часто накапливается внутриклеточно в неактивной форме (в виде телец включения)
Быстро растет и достигает высокой плотности клеток на относительно недорогой среде для ферментации	Невозможность проведения посттрансляционной модификации белков
Подходящая технология ферментации хорошо отработана	Наличие дополнительного остатка метионина на N-конце рекомбинантного белка
Хорошо зарекомендована для производства многих коммерческих терапевтических белков	

Гетерологичная экспрессия в клетках дрожжей

Достоинства	Недостатки
Сохраняют многие преимущества, описанные для систем экспрессии *E. coli*, такие как быстрый рост до высокой плотности клеток на недорогих средах	Уровни экспрессии гетерологичных белков часто низкие, обычно они составляют менее 5% от общего количества клеточного белка.
Очень хорошо изучены	Не все рекомбинантные белки успешно секретируются, а высоко экспрессированные белки остаются в эндоплазматическом ретикулуме
Возможно проведение посттрансляционной модификации белков	Гликозилирование не всегда напоминает таковое в нативных белках млекопитающих

Гетерологичная экспрессия в клетках грибов

Достоинства	Недостатки
Могут секретировать белки во внеклеточное пространство	Во внеклеточном пространстве белки могут разрушаться
Очень хорошо изучены	
Возможно проведение посттрансляционной модификации белков	

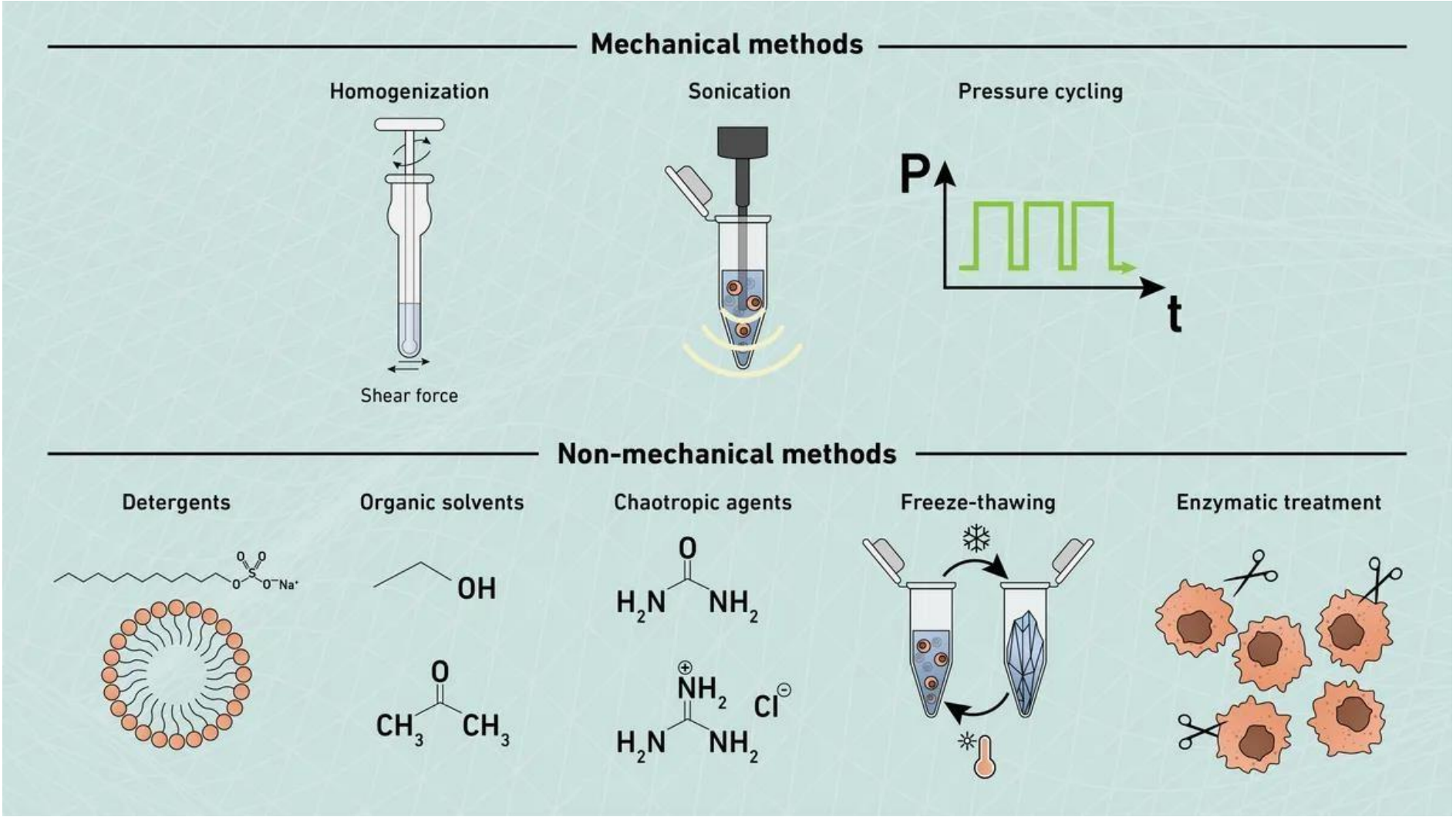
Характеристика белков в различных организмах

Характеристика	Прокариоты	Дрожжи	Клетки млекопит	Клетки растений	Трансген животн
Концентрация	Высокая	Высокая	Высокая	Низкая	Средн-высокая
Мол. вес	Малый	Большой	Большой	Большой	Большой
S-S мостики	Ограничен	Неограничен	Неограничен	Неограничен	Неограничен
Агрегация	Тельца включения	Нет	Нет	Нет	Нет
Фолдинг	Нарушен	Правильный	Правильный	Правильный	Правильный
Гликозилиров.	Ограничено	Ограничено	Возможно	Ограничено	Возможно
Контаминация	Эндотоксин	Низкое	Возможно	Низкое	Возможно
Секреция	Нет	+ / -	Да	+ / -	Да
Цена производ	Низкая	Низкая	Высокая	Высокая	Средне-высокая

Первый этап получения белка - экстракция

- Первое, с чего следует начать анализ – понять, белок расположен внутри или снаружи клетки
- В случае, если белок секретируется в культуральную жидкость, ее отделяют от клеточной массы центрифугированием или фильтрацией
- В случае, если белок расположен в клетке, необходимо ее разрушить (механически (гомогенизация), охлаждение-нагревание, осмотический шок, ультразвук, обработка ферментами и т.д.)
- После разрушения клетки необходимо экстрагировать белок буфером
- В состав буфера входят:
    - Антиоксиданты
    - Ингибиторы ферментов
    - Субстраты и кофакторы ферментов – при выделении самих ферментов
    - ЭДТА
    - Азид натрия
    - Соли (фосфатный буфер, NaCl, Tris)
    - Детергенты (SDS, Triton X-100, CTAB. дезоксихолат)

Первый этап получения белка - экстракция

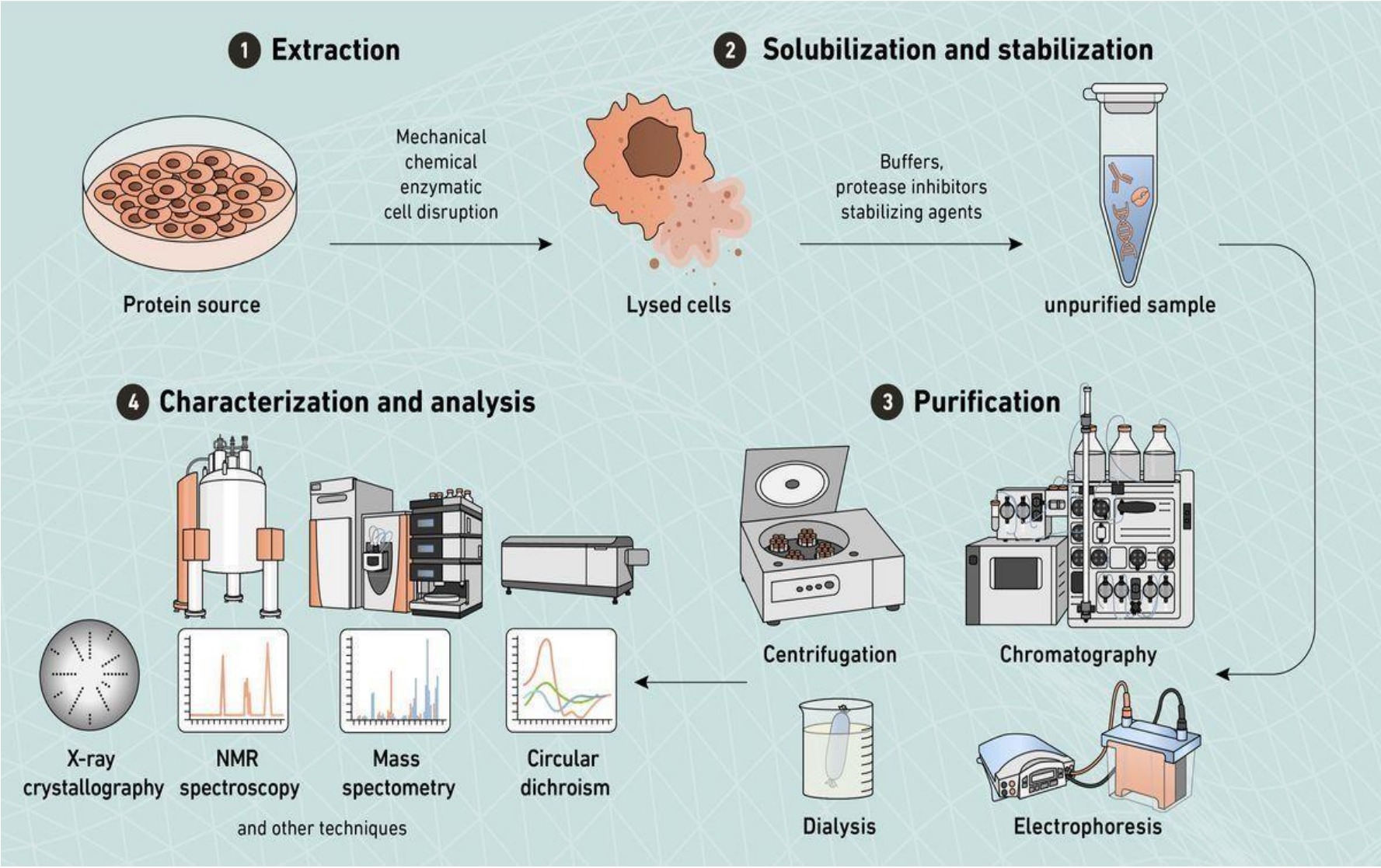


Источник: https://www.technologynetworks.com/analysis/articles/an-introduction-to-protein-purification-methods-technologies-and-applications-388443#D2

Очистка экстракта - обзор

Метод разделения/очистки	Принцип разделения	Разделение основано на…
Фильтрация	Микро, ультра, нано- фильтрация Диализ Глубинная фильтрация Заряженная мембрана	Размер Размер Размер Заряд
Центрифугирование	Градиентное, изопикническое и т.д.	Плотность
Экстракция	Жидкость-жидкостная/твердофазная	Растворимость, распределение
Осаждение	Фракционное высаживание	Растворимость
Хроматография	Ионный обмен Гель-фильтрация Аффинная Гидрофобная Адсорбирующая	Заряд Размер Специфич. взаимодействие Гидрофобность Ковалентн/нековалентн связывание

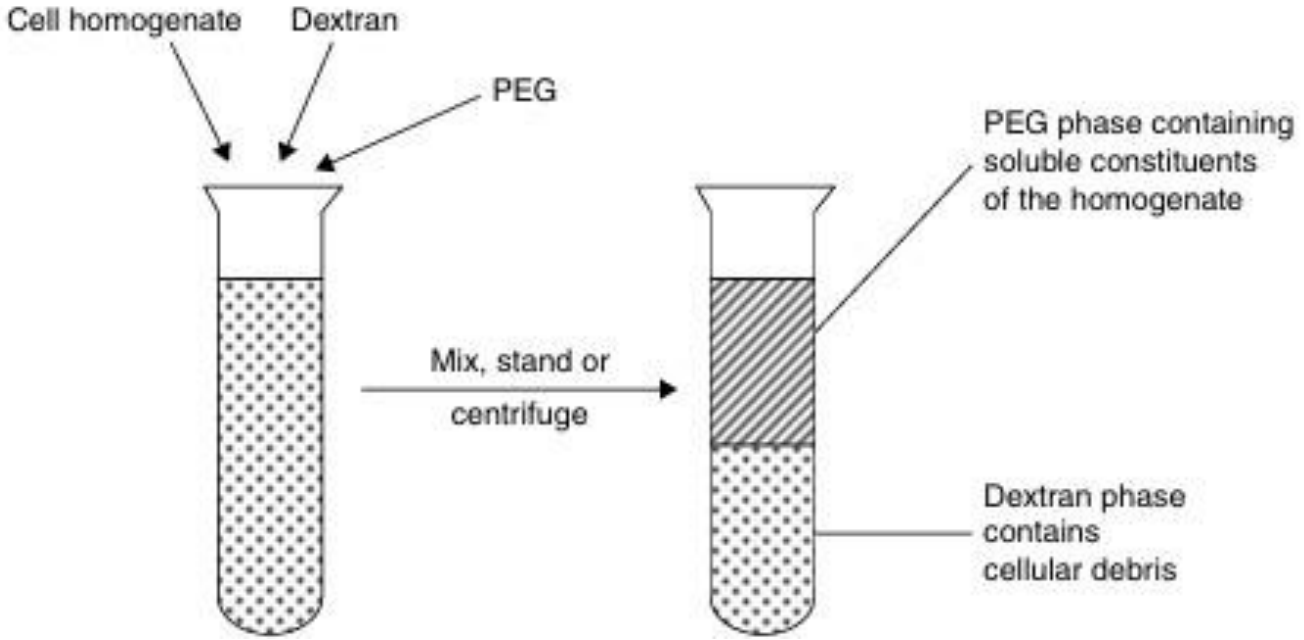
Очистка экстракта - обзор



Источник: https://www.technologynetworks.com/analysis/articles/an-introduction-to-protein-purification-methods-technologies-and-applications-388443#D2

Очистка экстракта

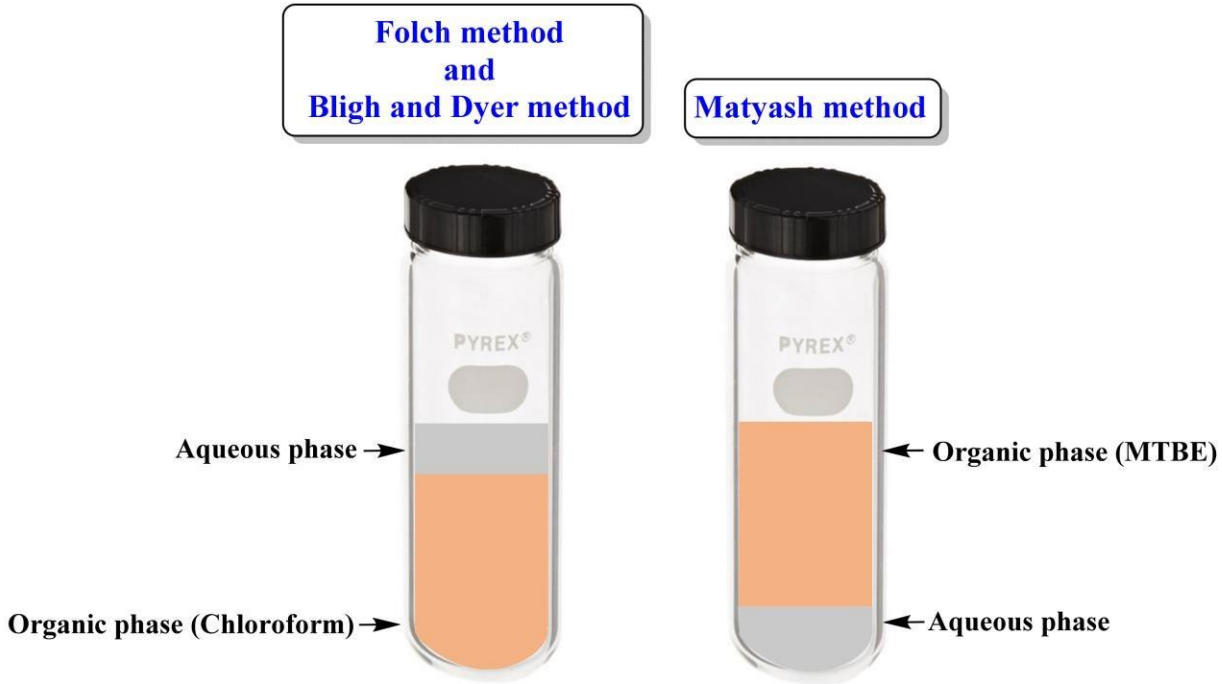
- После разрушения клеток и экстракции белка в буфер необходимо отделить остатки клеток – ЦЕНТРИФУГИРОВАНИЕ или ФИЛЬТРАЦИЯ
- Фильтруют чаще всего через фильтры 0,22 мкм
- Альтернативный метод – водная двухфазная экстракция с использованием водорастворимых полимеров



Источник: Walsh, G 2015, Proteins: Biochemistry and Biotechnology: Second Edition. Wiley-Blackwell. https://doi.org/10.1002/9781119117599

Очистка экстракта

- Удаление нуклеиновых кислот – осаждение полиэтиленимином, протаминсульфатом или обработка нуклеазами (РНКаза, ДНКаза)
- Удаление липидов – метод Фолча, Блая-Дайера (однофазная и двухфазные системы), Матьяша

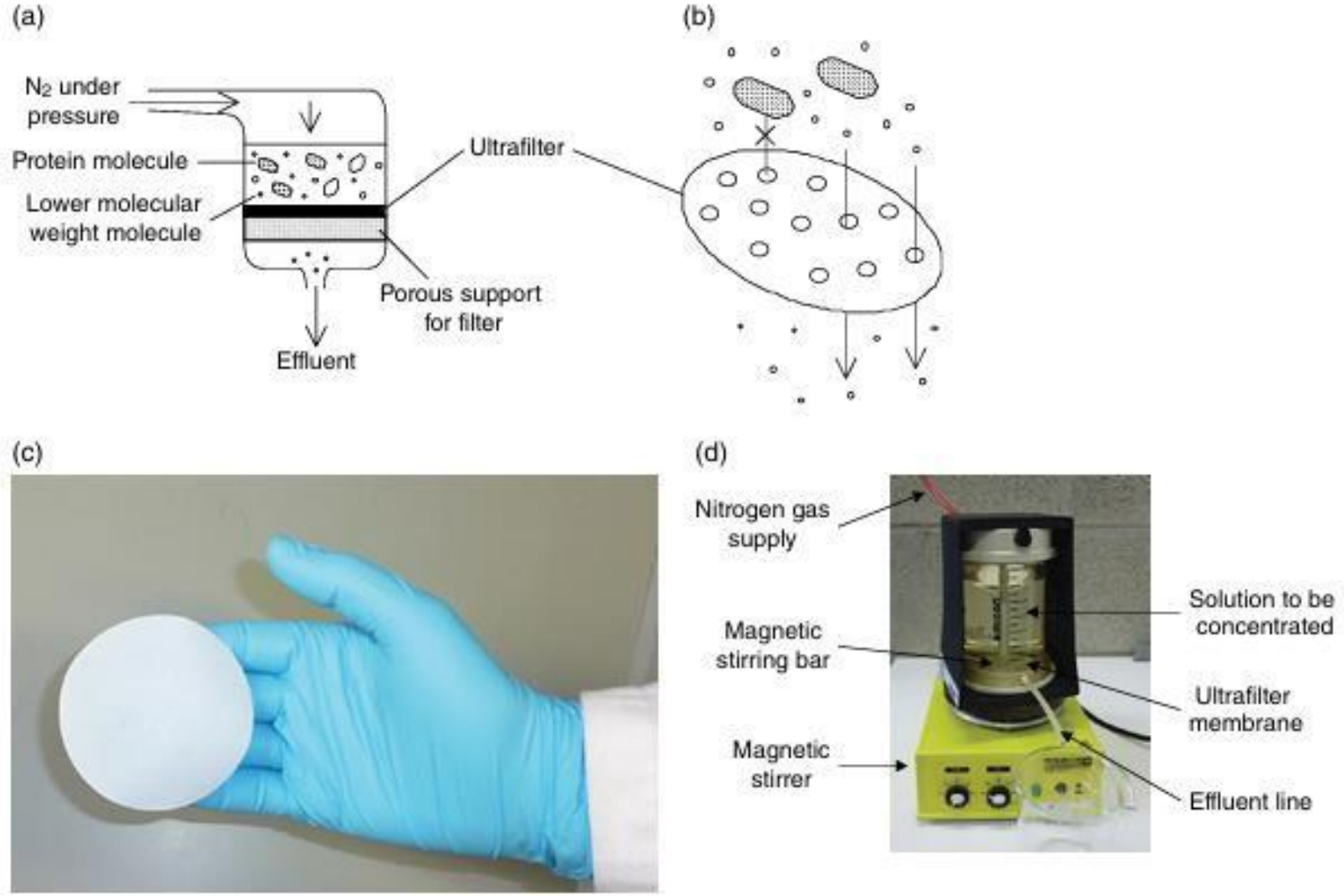


Источник: Saini RK, Prasad P, Shang X, Keum YS. Advances in Lipid Extraction Methods-A Review. Int J Mol Sci. 2021 Dec 20;22(24):13643. doi: 10.3390/ijms222413643

Концентрирование/фракционирование образца

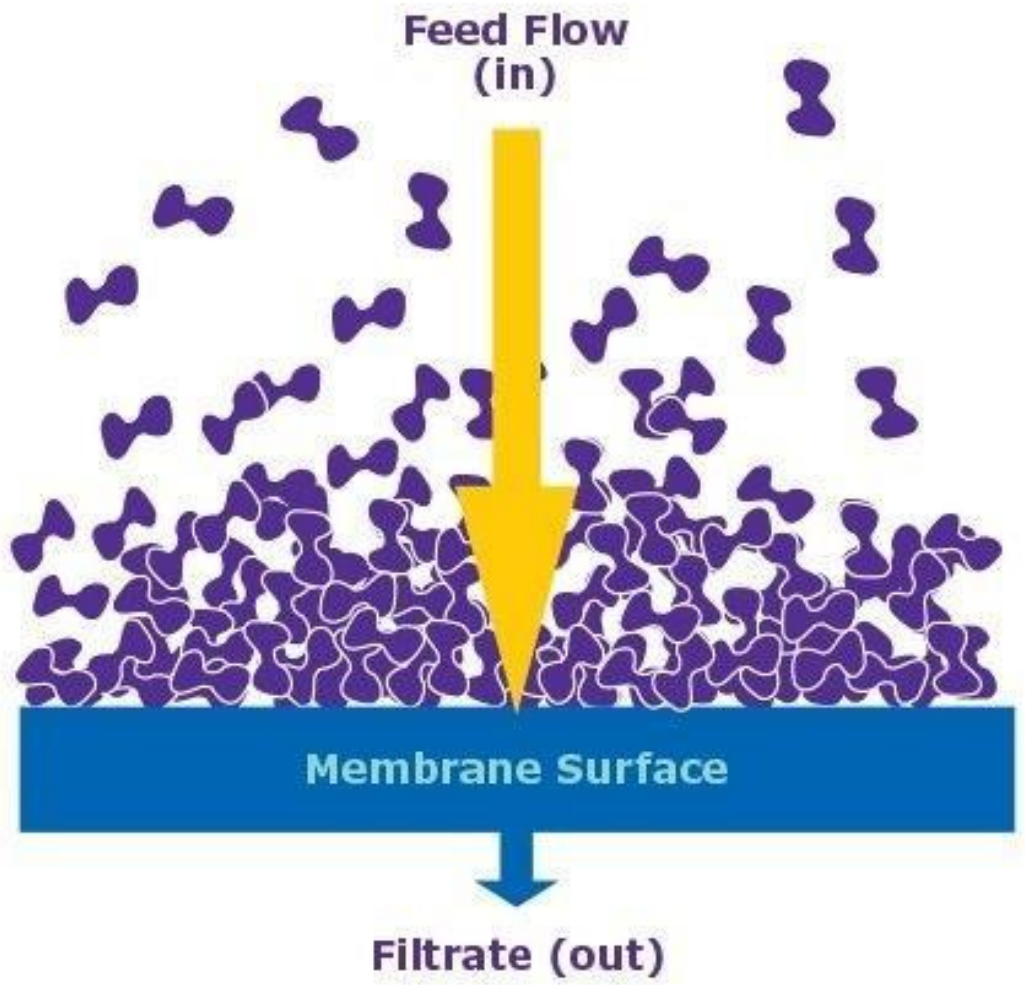
- Всегда удобнее работать с малыми объемами, поэтому стремятся сконцентрировать образец, то есть уменьшить его объем. Помимо концентрирования, возможно также отделить (фракционировать) одни белки от других
- Концентрирование достигается следующими методами:
    1. Высаживание
    2. Ультрафильтрация (диафильтрация)
    3. Хроматография
- Высаживание – достигается добавлением следующих агентов: соли, органические растворители, полимеры, изменение pH (например, подкисление трихлоруксусной кислотой)
- Ультрафильтрация (диафильтрация) – разделение молекул по размеру и форме. Используются мембраны с пределом отсечения («cut-off»), которым обозначается размер поры. Как правило, выбирают фильтрующую мембрану с пределом отсечения на 5 кДа меньше, чем масса целевого белка. ВАЖНО! «Дырка» в мембране в форме круга и предел отсечения относится к глобулярным белкам, т.е. необходимо принимать во внимание пространственную структуру целевой молекулы.

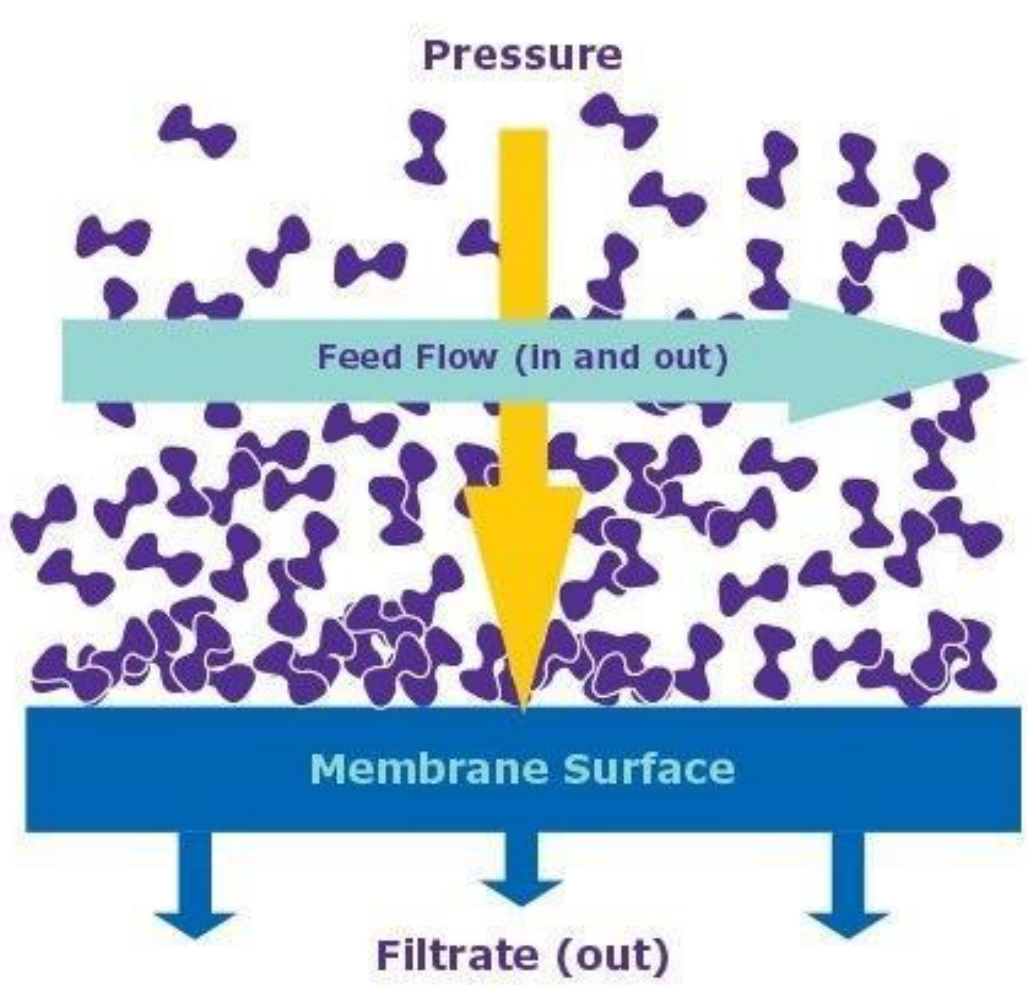
Концентрирование/фракционирование образца



Источник: Walsh, G 2015, Proteins: Biochemistry and Biotechnology: Second Edition. Wiley-Blackwell. https://doi.org/10.1002/9781119117599

Концентрирование/фракционирование образца

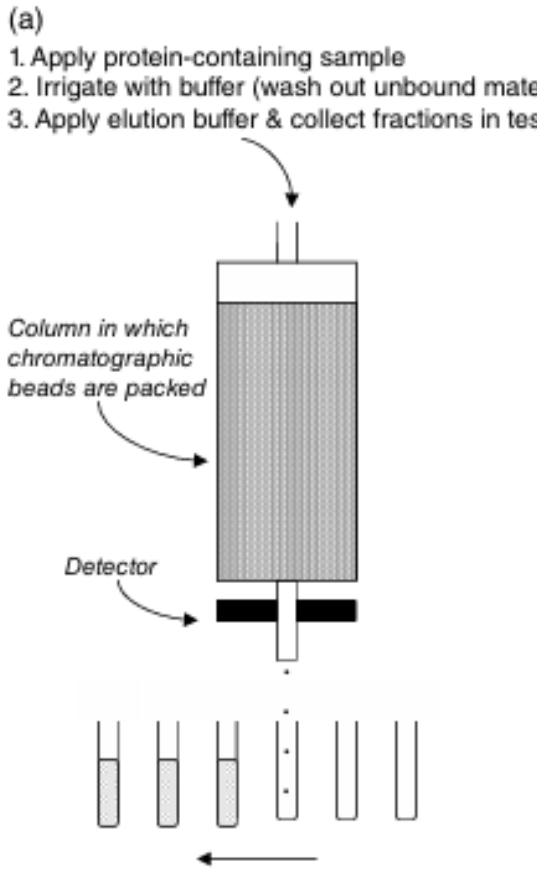


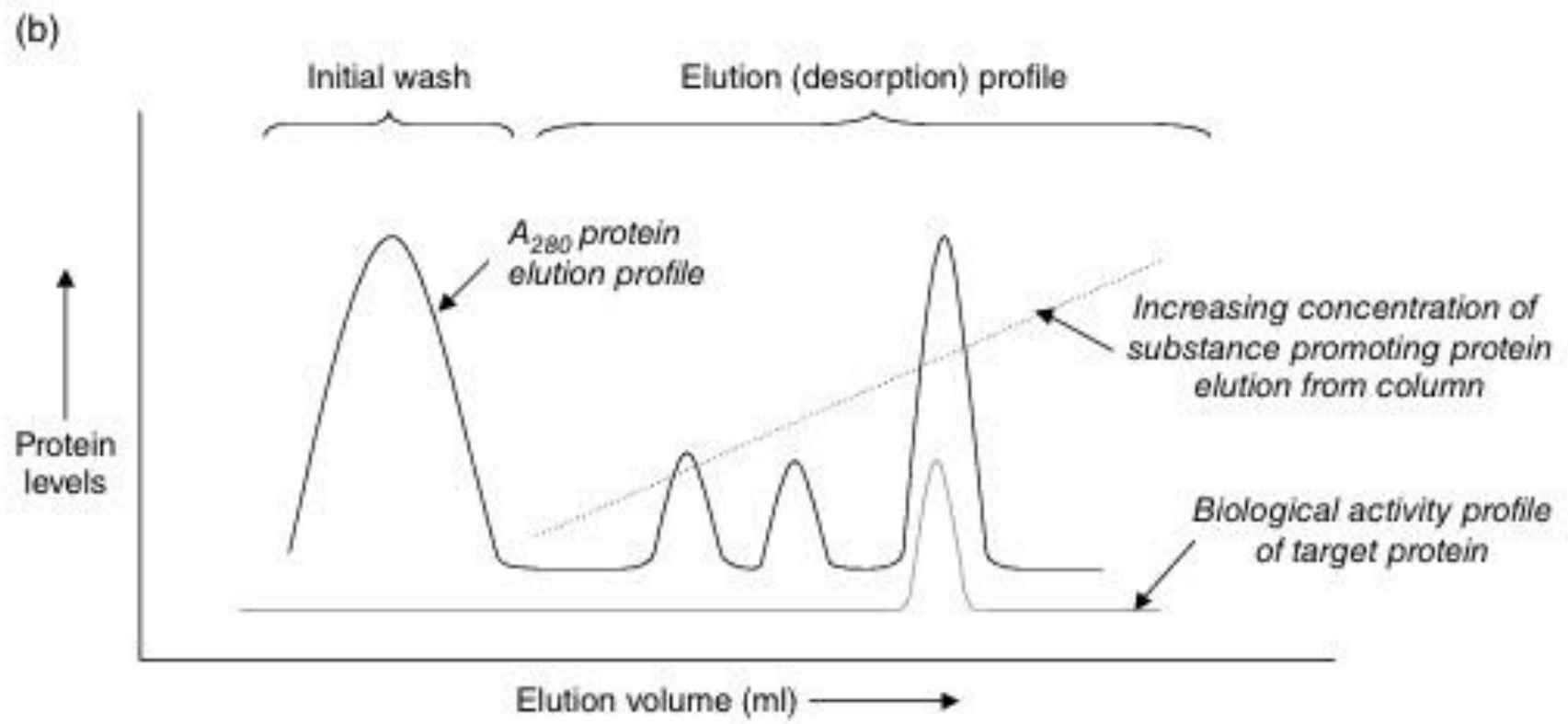


Источник: https://www.sigmaaldrich.com/RU/en/technical-documents/technical-article/pharmaceutical-and-biopharmaceutical-manufacturing/downstream-processing/tangential-flow-filtration-downstream-bioprocessing

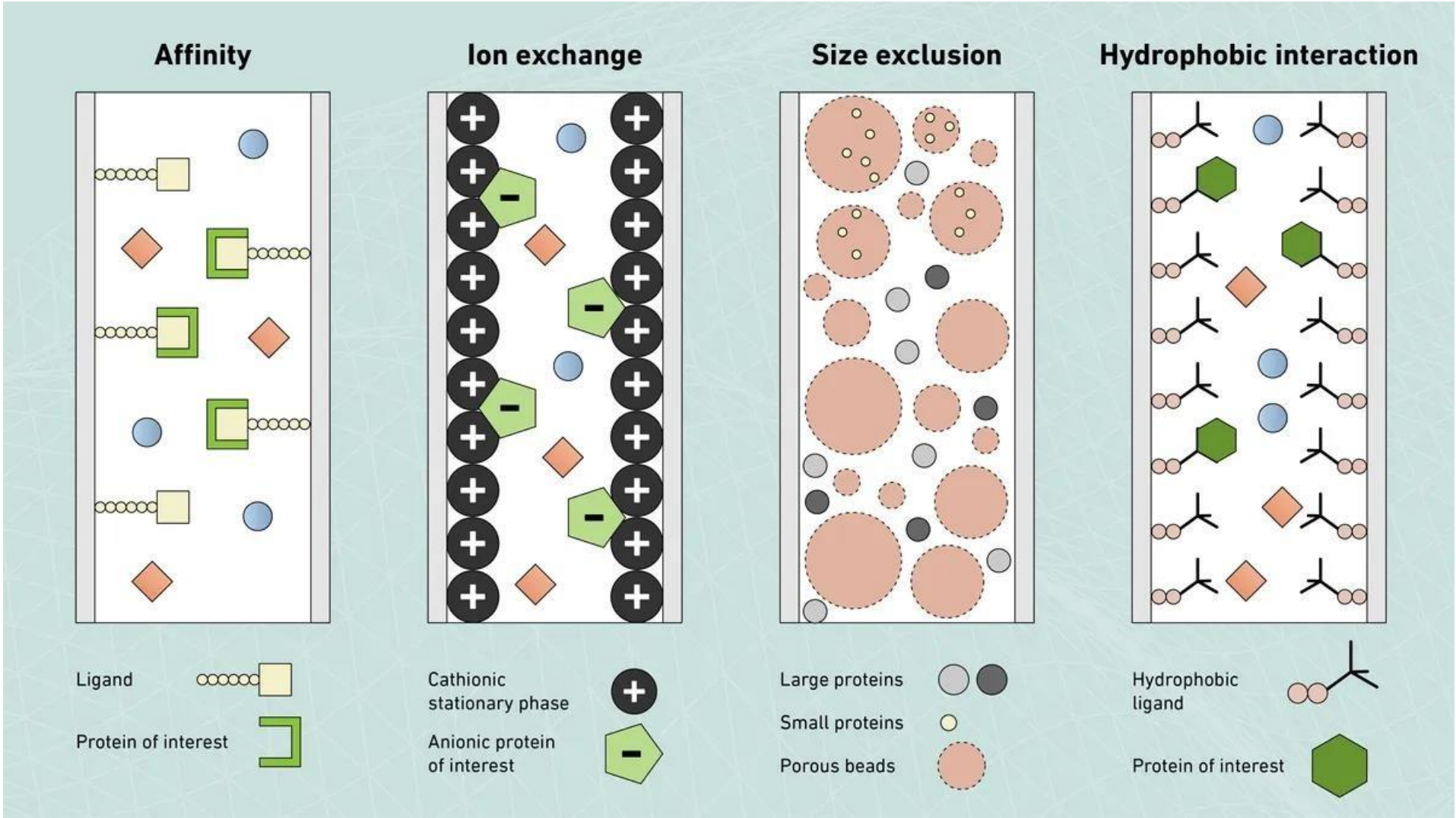
Хроматографические методы при очистке белков

- При очистке белков применяется один или несколько (комбинация) хроматографических методов
- Каждый метод использует различные свойства белков – заряд, гидрофобность, размер и т.д.





Хроматографические методы при очистке белков



Источник: https://www.technologynetworks.com/analysis/articles/an-introduction-to-protein-purification-methods-technologies-and-applications-388443#D2

Хроматографические методы при очистке белков

Метод	Используемое свойство	Емкость	Разрешение	Практические замечания
Гидрофобная хроматография	Гидрофобность	Высокая	Среднее	Можно разделять образцы с высоким содержанием соли, например после осаждения сульфатом аммония. Фракции имеют разные значения pH и/или ионной силы. Выход средний. Обычно применяют на начальных стадиях очистки. Результат непредсказуем.
Ионообменная хроматография	Заряд	Высокая	Среднее	Ионная сила образца должна быть низкой. Фракции имеют разные значения pH и/или ионной силы. Выход средний. Обычно применяют на ранних стадиях очистки.
Аффинная хроматография	Биологическая функция	Средняя (метод ограничен высокой стоимостью)	Высокое	Ограничения накладываются доступностью иммобилизованного лиганда. При элюировании белок может быть денатурирован. Выход от низкого до среднего. Обычно используется на заключительных этапах очистки.
Хроматография с красителями в качестве лигандов	Структура и гидрофобность		Высокое	Необходимо для начального скрининга красителей-лигандов.
Хроматофокусирование	Заряд и p*I*	Высокая–средняя	Высокое–среднее	Ионная сила образца должна быть низкой. Фракции содержат амфолиты.
Ковалентная хроматография	Тиоловые группы	Средняя–низкая	Высокое	Подходит для белков, содержащих тиоловые группы. Ограничения накладываются стоимостью и продолжительностью (3 ч) регенерации.
Металл-хелатная хроматография	Имидазольная, тиоловая и триптофановая группы	Средняя–низкая	Высокое	В литературе описано всего лишь несколько примеров. Дорого.
Эксклюзионная хроматография	Размер молекул	Средняя	Низкое	Обычно используется на заключительных стадиях очистки. Может дать информацию о молекулярной массе белка. Подходит для обессоливания препарата белка.

Источник - Принципы и методы биохимии и молекулярной биологии [Электронный ресурс] / ред. К. Уилсон и Дж. Уолкер ; пер. с англ.—2-е изд. (эл.).—Электрон. текстовые дан. (1 файл pdf : 855 с.).—М. : БИНОМ. Лаборатория знаний, 2015.—(Методы в биологии)

Выделение и очистка нуклеиновых кислот в геномике

Зачем это биоинформатику?

- Пробоподготовка задаёт «границы» данных ещё до анализа
- Ошибки выделения дают систематические смещения (bias), которые выглядят как «биология»
- Правильный выбор метода экономит время/деньги (лучше «переделать» потенциально «плохую» пробу)

От образца к данным

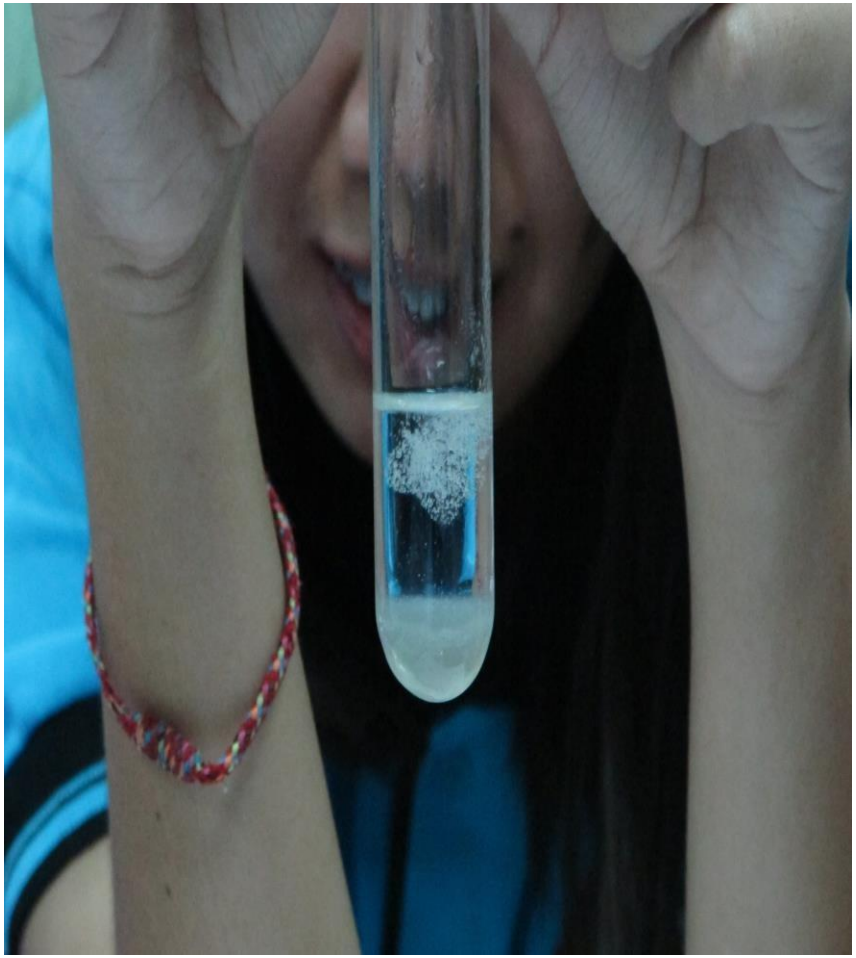
- Преаналитика: сбор, транспорт, стабилизация
- Выделение НК: ингибиторы, потери, фрагментация
- Подготовка библиотеки: потери молекул, перекос длин, дупликации
- Секвенирование: качество чтений, ошибки индексации
- Биоинформатика: фильтрация, маппинг, варианты/экспрессия
- Ключевая идея: «мусор на входе» → систематическая ошибка на выходе

Что именно выделяем в геномике

- Геномная ДНК: WGS (whole-genome sequencing, полно-геномное секвенирование), WES (whole-exome sequencing, секвенирование экзома), таргет-панели
- РНК: RNA-seq (RNA sequencing, секвенирование транскриптома), scRNA-seq (single-cell RNA-seq, одно-клеточное RNA-seq)
- Эпигеномика: bisulfite-seq (бисульфит-секвенирование), ATAC-seq (оценка доступности хроматина), ChIP-seq (иммунопреципитация хроматина)
- Микробиом/среда: метагеномика и eDNA (environmental DNA, ДНК "окружающей среды")

«ДНК можно увидеть»

- Даже простая экстракция — это отделение полимера от «супа» белков/липидов/солей
- В реальной геномике важнее не «красота нити», а отсутствие ингибиторов и воспроизводимость
- *Источник изображения: Wikimedia Commons — File:DNA_Extraction.jpg (https://commons.wikimedia.org/wiki/File:DNA_Extraction.jpg)*



Матрицы и «трудные» образцы

- Кровь/плазма: гемолиз, гепарин (ингибитор), низкий вход cfDNA
- Ткани: РНК быстро деградирует; липиды мешают очистке
- Растения: полисахариды и полифенолы → сильные ингибиторы
- Почва/ил: гуминовые кислоты → ингибирование ПЦР/лигирования
- FFPE: сшивки и химические модификации → фрагментация «на входе»

Требования к НК зависят от задачи

- PCR (polymerase chain reaction, полимеразная цепная реакция): критично убрать ингибиторы; длина вторична
- NGS (next-generation sequencing, секвенирование нового поколения): нужна точная концентрация и чистота
- Long-read: нужна HMW-ДНК (high-molecular-weight DNA, высокомолекулярная ДНК) и минимум сдвиговых усилий
- RNA-seq: нужна целостность РНК и удаление ДНК (DNase-обработка)

Преаналитика: что контролировать до выделения

- Время до стабилизации/заморозки (особенно для РНК)
- Температура и транспортировка
- Тип пробирки (EDTA/гепарин/цитрат; стабилизаторы РНК)
- Гомогенизация: достаточно, но без «перемалывания» HMW-ДНК
- Отслеживаемость: маркировка, журналы, исключение перепутывания

Контаминации: как не «секвенировать лабораторию»

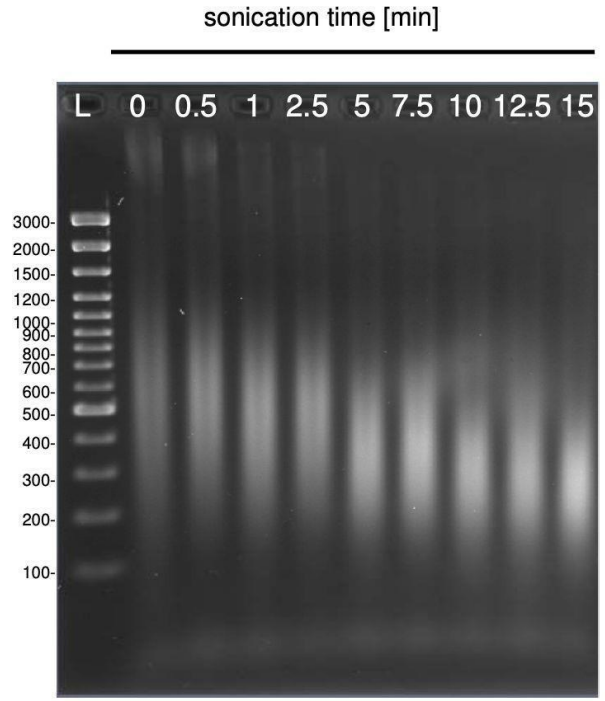
- Разделить зоны: pre-PCR (до амплификации) / post-PCR (после)
- Однонаправленный поток проб и расходников
- Фильтр-наконечники и регулярная деконтаминация
- Blank extraction control (контроль «пустого выделения») особенно важен. В метагеномике контаминации легко становятся «ложными таксонами»

Базовая схема выделения НК

- 1) Лизис + инактивация нуклеаз
- 2) Депротеинизация (протеиназа K и/или хаотропные соли)
- 3) Разделение/связывание НК (органика или твердофазное связывание)
- 4) Промывки (удаление солей, детергентов, ингибиторов)
- 5) Элюция и (при необходимости) концентрирование
- 6) QC и нормализация под следующий этап

Лизис: механический, химический, ферментативный

- Механический: гомогенизация
- Химический: детергенты + соли; EDTA связывает Mg^{2+} и тормозит нуклеазы
- Ферментативный: лизоцим (бактерии), протеиназа K (белки/нуклеазы)
- Для HMW-ДНК: избегать агрессивного измельчения



https://www.hielscher.com/ultrasonic-dna-shearing.htm

Почему РНК «капризнее»

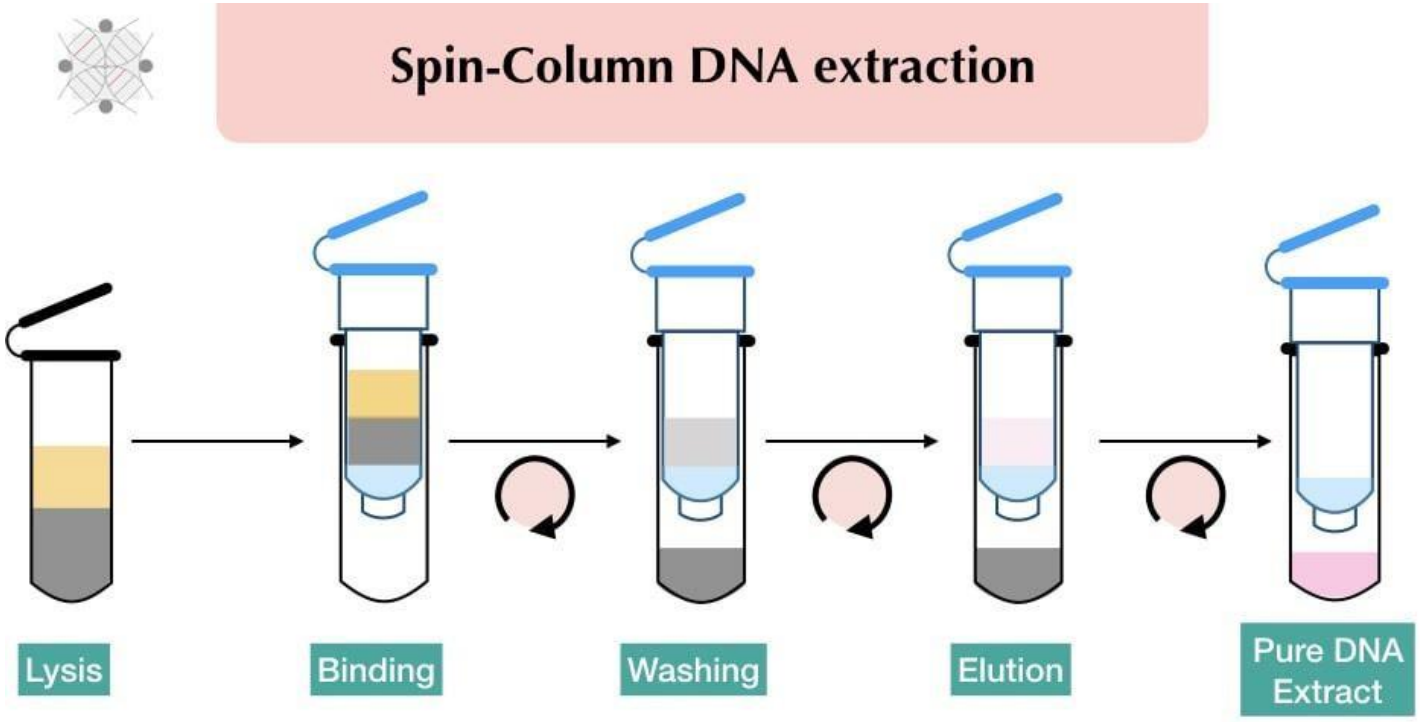
- РНКазы устойчивы и повсеместны (кожа, пыль, поверхности)
- Хаотропные соли денатурируют белки и инактивируют нуклеазы
- Работа с РНК: быстро + холодно + "RNase-free" расходники (свободные от РНК)
- Часто нужна обработка ДНКазами для удаления геномной ДНК

Фенол-хлороформ: когда уместен

- Плюсы: высокая чистота, хорош при «грязных» матрицах, можно получить длинную ДНК
- Минусы: токсичность, ручная работа, риск остаточного фенола
- Остаточный фенол искажает 260/230 и ингибирует ферменты. В рутине NGS чаще выбирают колонки/бусины ради стандартизации

Силика-колонки: принцип

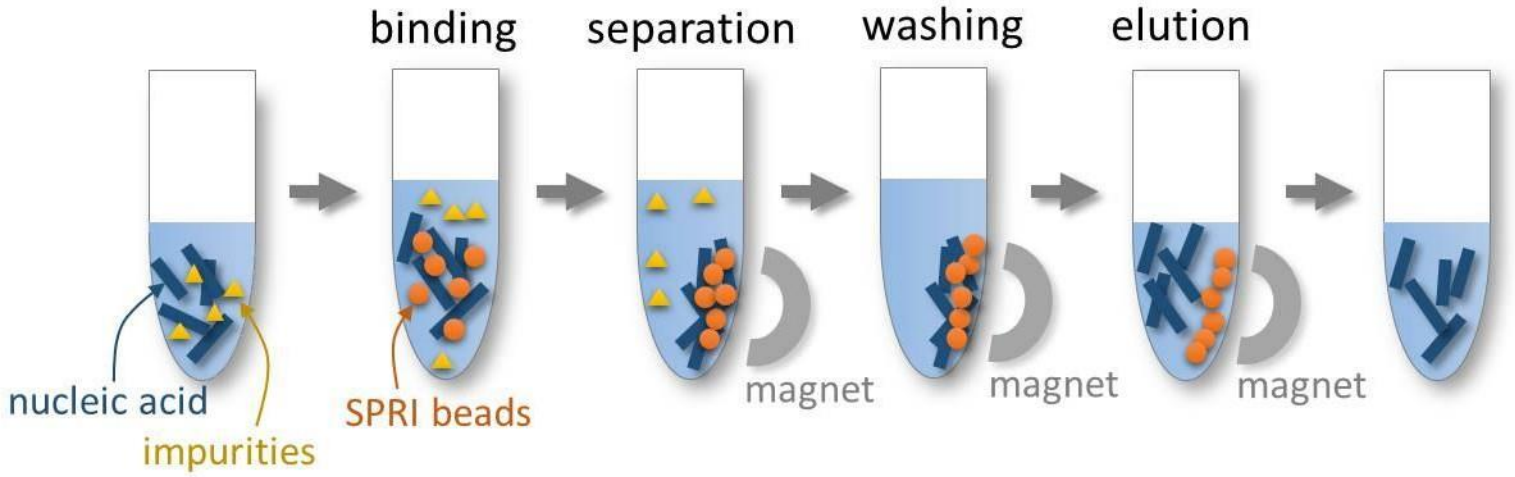
- В хаотропных солях НК связываются с силикой
- Промывки со спиртом убирают соли/белки/детергенты
- Элюция при низкой ионной силе (вода/низкосолевой буфер)
- Плюсы: воспроизводимо; минусы: потери, ограничение по объёму



https://geneticeducation.co.in/advantages-and-limitations-of-spin-column-dna-extraction-technique/

Магнитные частицы SPRI: стандарт NGS

- SPRI (solid-phase reversible immobilization, обратимая иммобилизация на твердой фазе)
- Связывание НК в присутствии PEG (polyethylene glycol, полиэтиленгликоль) и соли
- Удобные промывки на магните → легко автоматизировать
- Размер-селекция (отбор по длине фрагментов) задаётся соотношением бусины/образец



https://geneticeducation.co.in/advantages-and-limitations-of-spin-column-dna-extraction-technique/

Осаждение спиртом: простой способ концентрирования

- Подходит для концентрирования и замены буфера
- Риски: потери при низком входе и перенос солей/ингибиторов
- Нужны корректные промывки и полное удаление спирта
- Для малых концентраций (low input) чаще предпочтительнее bead-cleanup

Растения и CTAB-подход

- CTAB (cetyltrimethylammonium bromide, цетилтриметиламмоний бромид) –катионный детергент - помогает убрать полисахариды
- Полифенолы требуют дополнительных приёмов (сорбенты/антиоксиданты)
- Часто комбинируют: «грубое» выделение → очистка на силике/бусинах
- Цель: получить материал, который не ингибирует ферменты

Метагеномика: баланс лизиса и фрагментации

- Разные таксоны лизируются по-разному → перекос профиля сообщества
- Bead-beating повышает выход, но может «порвать» ДНК
- Ингибиторы (гуминовые кислоты) часто требуют доп. очистки
- Нужны blank-контроли и, по возможности, mock-сообщества

RNA-seq: выбор стратегии влияет на выделение

- Poly(A)-selection (отбор по поли-A): лучше при целой РНК
- rRNA depletion (удаление рРНК): подходит для деградированных/FFPE
- Small RNA-seq: работа с очень короткими фрагментами
- scRNA-seq: ультра-низкий вход → потери на очистках критичны

QC: сколько молекул на самом деле?

- UV-спектрофотометрия (260 нм) «видит всё», что поглощает на 260
- Флуориметрия (специфично к dsDNA/RNA) лучше для расчёта входа
- qPCR (quantitative PCR, количественная ПЦР) — функциональная оценка пригодности входа
- Вход в библиотеку считать по флуориметрии/ qPCR, а UV использовать для чистоты

QC: чистота (260/280 и 260/230) — как читать

- Низкий 260/280: белки или фенол
- Низкий 260/230: соли/гуанидин/углеводы/фенол
- Если UV-концентрация сильно выше флуориметрии — вероятны примеси
- Плохой 260/230 часто предсказывает провал лигирования адаптеров

Когда нужна повторная очистка (cleanup)

- Если 260/230 низкий → вероятны соли/гуанидин → очистка бусинами или колонкой
- Если есть риск остаточного спирта → досушить и повторить элюцию
- При низком входе каждая доп. очистка "съедает" материал

Автоматизация и стандартизация

- В больших проектах важнее воспроизводимость, чем «рекордный выход»
- Bead-based протоколы проще всего роботизировать
- LIMS (laboratory information management system, лабораторная информационная система) помогает трассировать пробы и QC-точки
- Стандартные контроли: blank extraction + контрольный образец

Хранение и биобанкинг

- ДНК: −20/−80 °C; буфер TE (Tris-EDTA, трис-ЭДТА) снижает активность нуклеаз
- РНК: −80 °C; аликвоты (аликвотирование, разлив по аликвотам), избегать заморозки-оттаивания
- Документировать число циклов заморозка-оттаивание и условия
- При долгом хранении — периодический QC

Финальный чек-лист перед запуском секвенирования

- Определена цель и платформа (short-read/long-read; DNA/RNA)
- Преаналитика соблюдена (время/температура/пробирки)
- QC закрыт: количество + чистота + длина/целостность
- Есть контроли (blank extraction)
- Протокол библиотеки выбран под качество образца

Итоги

- Метод выделения выбирают под задачу и матрицу, а не «по привычке»
- QC — это не формальность: он прогнозирует провал/успех библиотек
- Для long-read ключ — HMW-ДНК и мягкая обработка
- Для метагеномики ключ — лизис + контроли контаминации

Секвенирование нуклеиновых кислот

Секвенирование — это «прочитать текст» ДНК или РНК

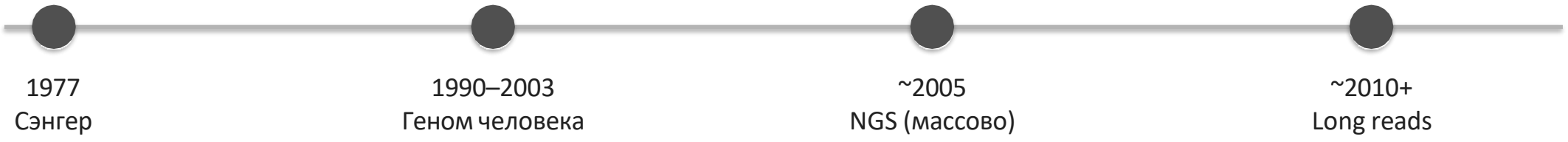
- ДНК/РНК можно представить как длинную строку из букв: A, C, G, T (в РНК вместо T — U)
- Секвенирование — лабораторный метод, который превращает молекулу в цифровую последовательность
- Эта последовательность позволяет искать мутации, сравнивать организмы и измерять активность генов

На какие вопросы отвечает секвенирование

- Кто это? (идентификация микроорганизма, штамма, вирусного варианта)
- Что изменилось? (мутации: замены, вставки/делеции, перестройки)
- Что работает прямо сейчас? (RNA-seq: измерение экспрессии генов)
- Из чего состоит смесь? (метагеномика: сообщества микробов)
- Как устроен геном нового организма? (de novo сборка)

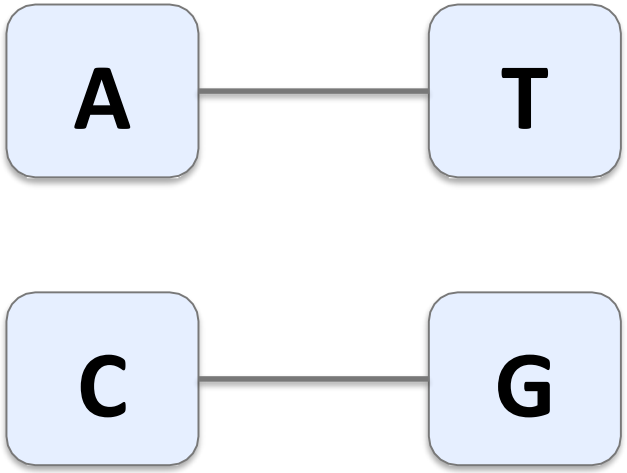
Короткая история: как мы дошли до NGS и long reads

- 1977: метод Сэнгера (цепное прерывание) — «золотой стандарт» на десятилетия
- 1990–2003: проект «Геном человека» — секвенирование стало большим индустриальным процессом
- с ~2005: второе поколение (Next-Generation Sequencing, NGS) — миллионы фрагментов параллельно
- с ~2010: третье поколение — «длинные прочтения» (PacBio, Nanopore): читаем десятки тысяч оснований за раз



Минимум биохимии: почему вообще можно «читать» ДНК

- У ДНК есть комплементарность: A–T, C–G (в РНК: A–U)
- ДНК-полимераза достраивает новую цепь по матрице
- Если мы научимся фиксировать, какая «буква» присоединилась на каждом шаге, мы получим последовательность



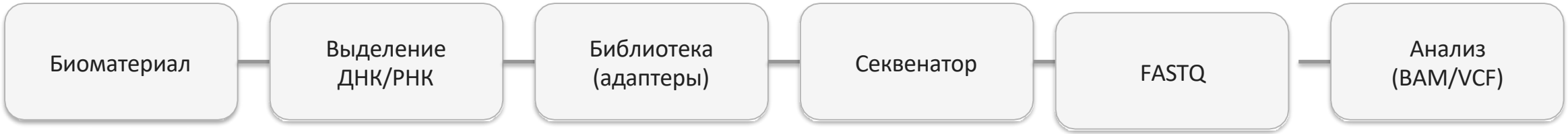
Комплементарность — основа чтения:
полимераза «угадывает» правильную букву по матрице.

Ключевые понятия

Шаблон (template) — исходная ДНК/РНК, которую хотим прочитать
Прочтение (read) — короткий фрагмент последовательности, считанный прибором
Покрытие (coverage) — сколько раз каждая позиция прочитана разными reads
Референс (reference) — «эталонный» геном для выравнивания
Сборка (assembly) — восстановление генома без референса

Общий процесс секвенирования (в любом поколении)

1) Берём биоматериал и выделяем ДНК/РНК
2) Подготавливаем библиотеку: делаем фрагменты и добавляем служебные «хвосты» (адаптеры)
3) Загружаем библиотеку в прибор и запускаем реакцию
4) Прибор регистрирует сигнал → превращает его в буквы (base calling)
5) Получаем файлы и проводим биоинформатический анализ



Типовые форматы проектов (что именно секвенируют)

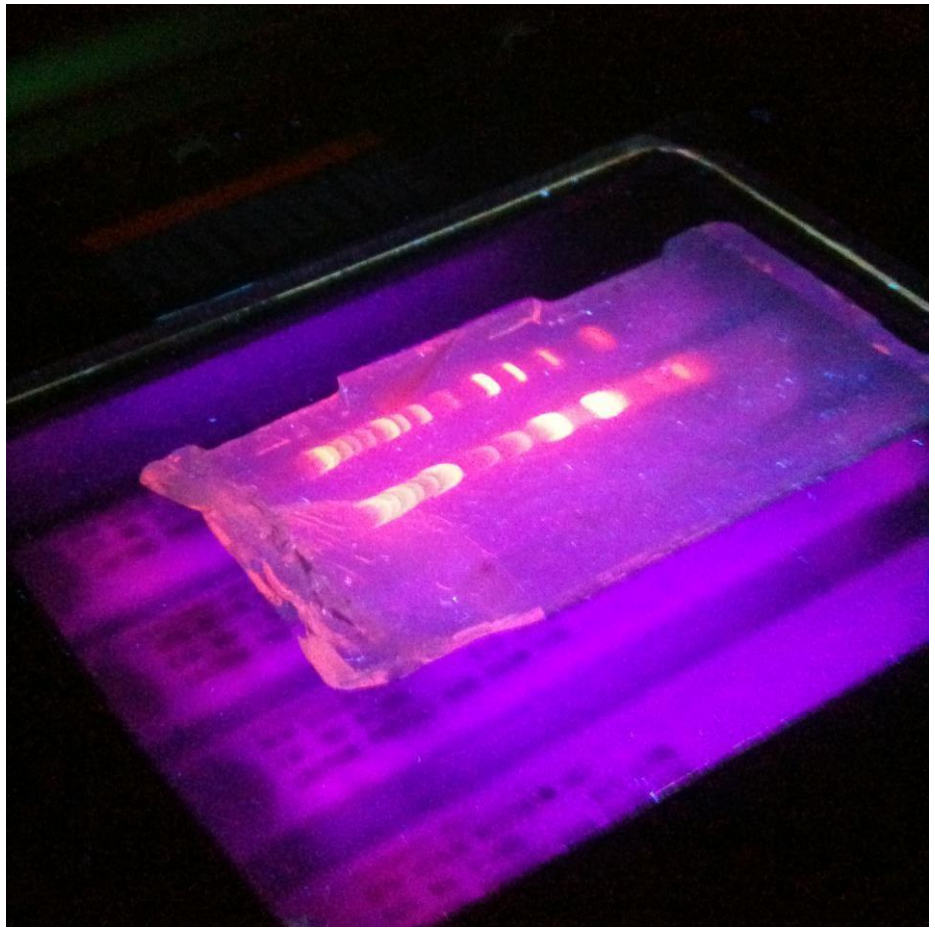
Таргетное секвенирование: один ген/участок (например, «проблемный» экзон)
Панели генов: десятки–сотни генов (диагностика, онкология)
Экзом: только кодирующие участки генома человека
Полногеномное секвенирование: весь геном
16S/ITS-ампликоны: быстрый профиль микробиоты
RNA-seq: транскриптом (набор РНК) — «что экспрессируется»

Что нужно «на столе» для секвенирования

- Качественный образец ДНК/РНК (без деградации и ингибиторов)
- Набор для подготовки библиотеки (ферменты, адаптеры, праймеры)
- Инструменты контроля качества: флуориметр/спектрофотометр, гель/капиллярный анализ
- Секвенатор и расходники (flow cell / чип / реагенты)
- Компьютерный ресурс для анализа и хранения данных

Контроль качества образца: что проверяем до запуска

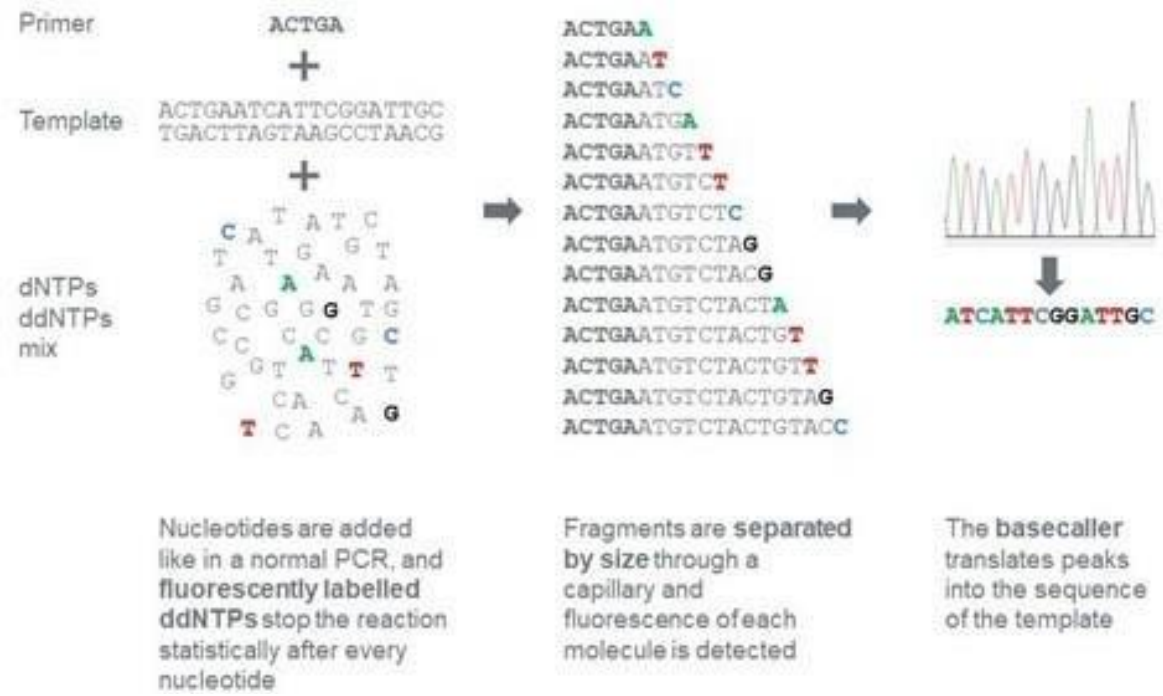
Количество (концентрация): хватает ли ДНК/РНК для протокола
Чистота: нет ли белка, фенола, солей, которые мешают ферментам
Размер/целостность: не порвана ли ДНК (особенно важно для long reads)
Контаминации: чужая ДНК (например, бактерии в клинике или плазмидная ДНК в лаборатории)



Источник изображения: Wikimedia Commons — https://commons.wikimedia.org/wiki/File:Agarose_gel_with_DNA_ladders.jpg (источник фото: Flickr автора Joseph Elsbernd)

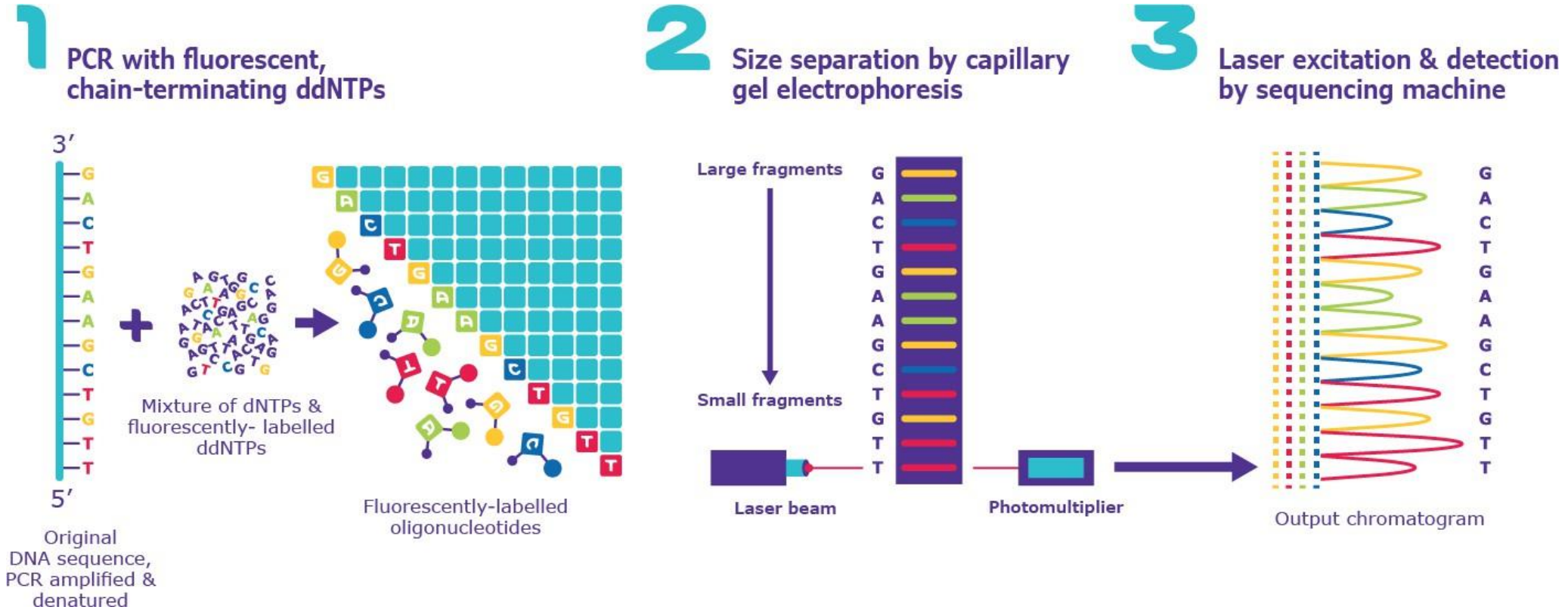
1-е поколение: метод Сэнгера — цепное прерывание

- Делаем синтез ДНК, но добавляем «прерыватели» — дидезоксинуклеотиды (ddNTP)
- Когда ddNTP встраивается, цепь перестаёт удлиняться
- В смеси получается набор фрагментов всех возможных длин
- По длине фрагментов и цветной метке определяем последовательность



Источник изображения: https://eurofinsgenomics.eu/en/eurofins-genomics/material-and-methods/sanger-sequencing/

1-е поколение: метод Сэнгера — цепное прерывание



Источник изображения: https://theory.labster.com/dna_sequencing/

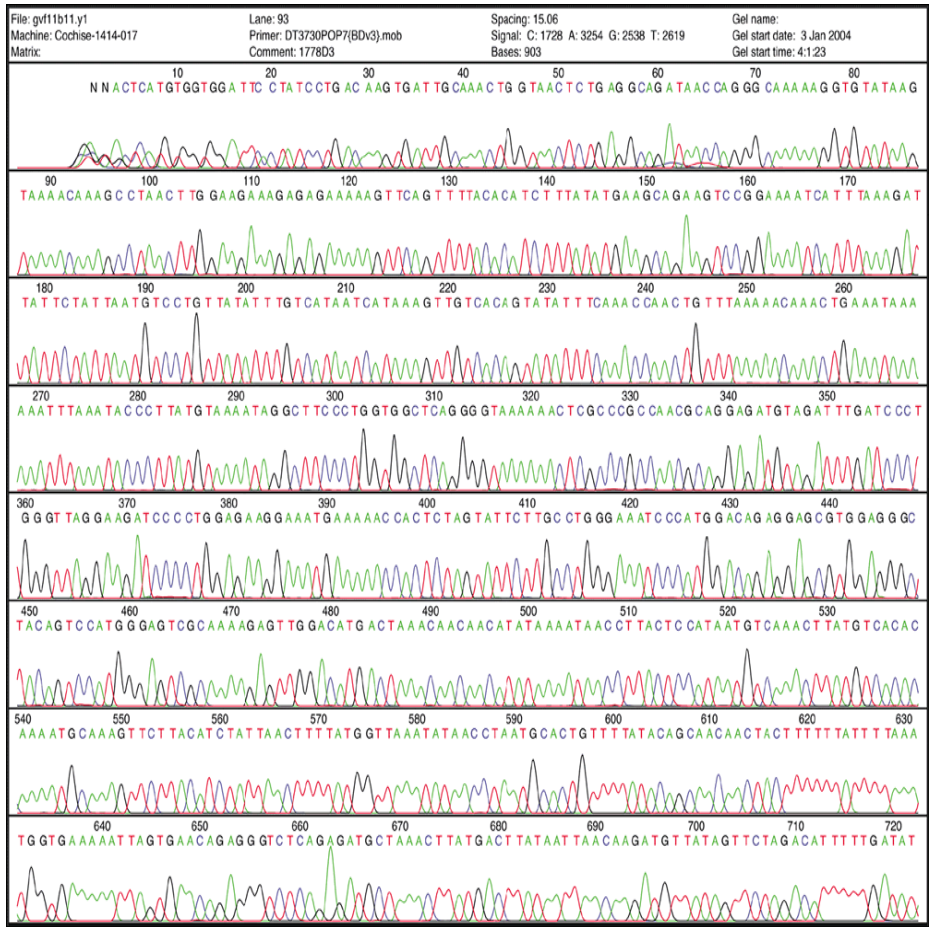
Сэнгер: как выглядит результат (хроматограмма)

Прибор выдаёт график сигналов по позициям: каждый цвет соответствует букве

Идеальная хроматограмма — одиночные пики, без «двойных» сигналов

Двойные пики часто означают смесь (например, гетерозигота или загрязнение)

По качеству пиков можно оценить доверие к каждой букве



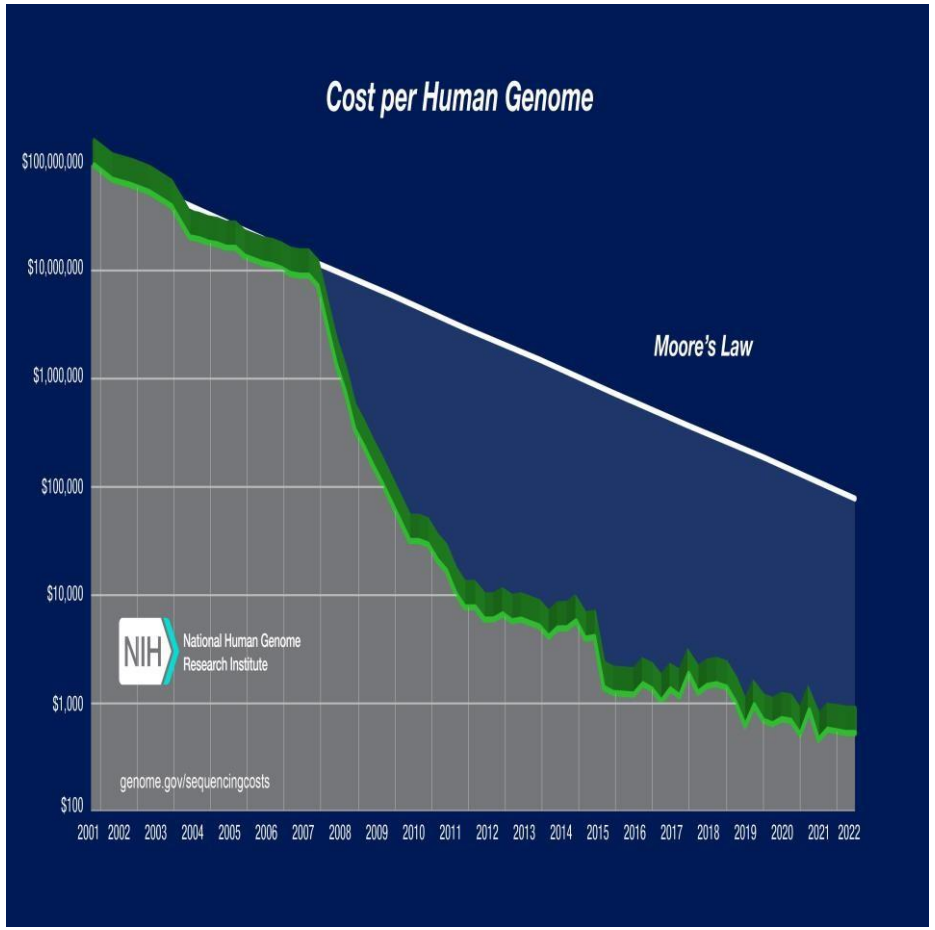
Источник изображения: Wikimedia Commons — https://commons.wikimedia.org/wiki/File:DNA_sequencing_interferogram.png

Где Сэнгер используется сегодня

- Проверка (валидация) мутации, найденной NGS
- Секвенирование плазмид, отдельных ампликонов, клонов
- Подтверждение конструкции (например, CRISPR-редактирование)
- Небольшие проекты, где важна точность, а не масштаб

Почему возникло NGS: масштаб и стоимость

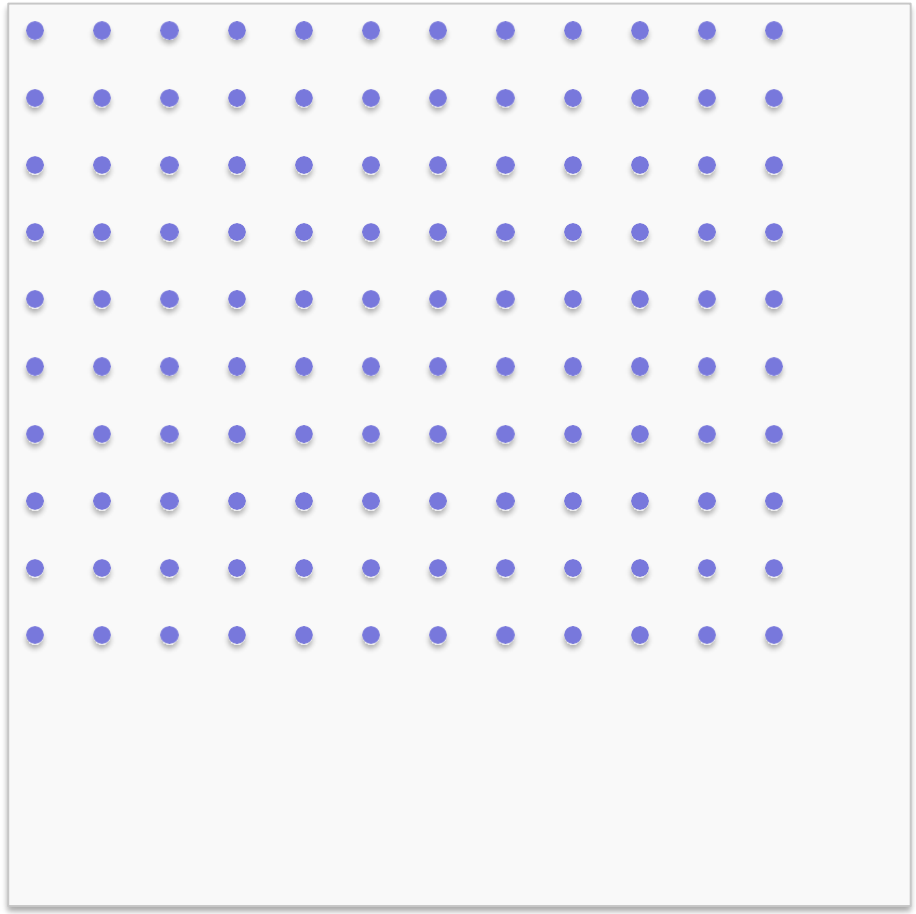
Сэнгер читает «по одной молекуле/ампликону за раз» — это медленно для больших геномов
NGS делает то же самое, но параллельно: миллионы фрагментов одновременно
Это резко снизило стоимость секвенирования и расширило круг задач



Источник изображения: NHGRI / genome.gov — https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (данные) / Wikimedia Commons: https://commons.wikimedia.org/wiki/File:Cost_per_Ge

2-е поколение (NGS): идея параллельного чтения

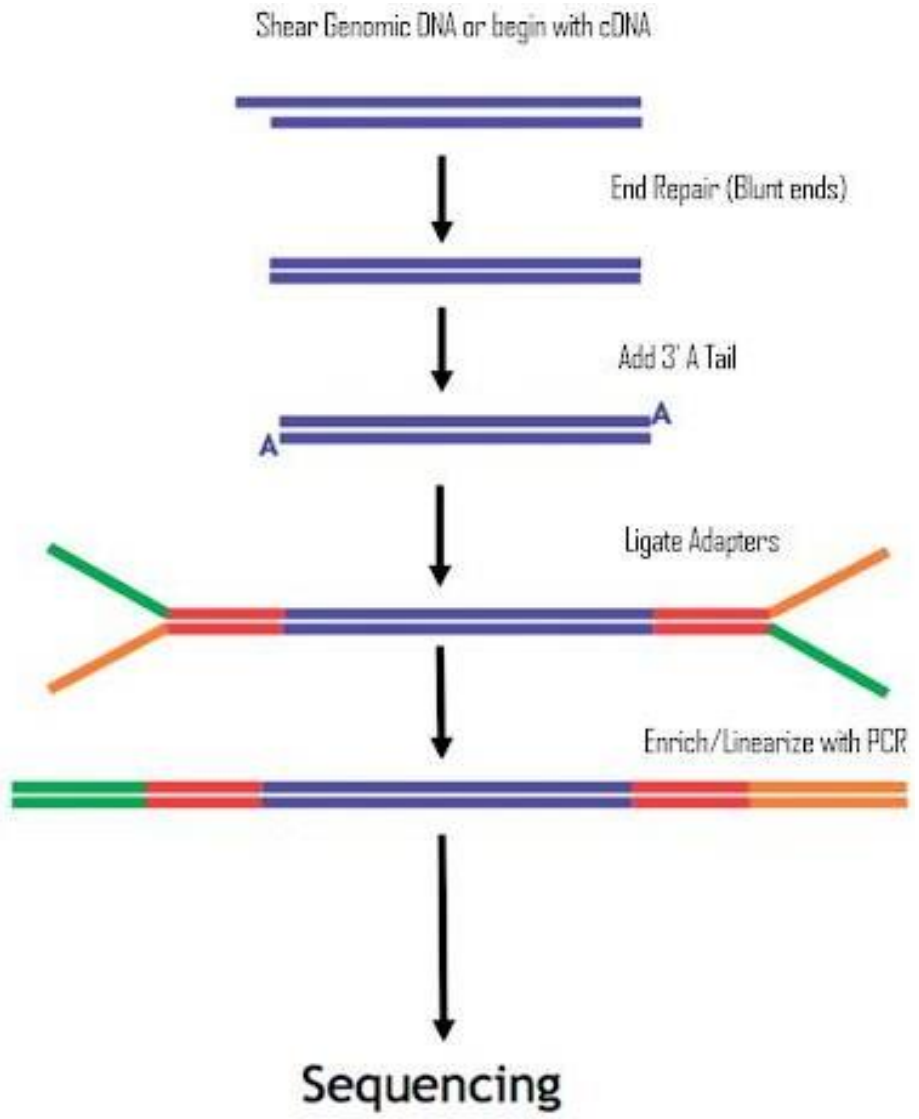
- Мы дробим ДНК на много коротких фрагментов
- Каждый фрагмент закрепляется на поверхности и «размножается» локально
- Камера/сенсор фиксирует сигнал синтеза для миллионов точек параллельно
- Итог — огромный набор коротких reads, который потом собирается вычислительно



Миллионы кластеров читаются одновременно

Подготовка библиотеки: зачем нужны адаптеры

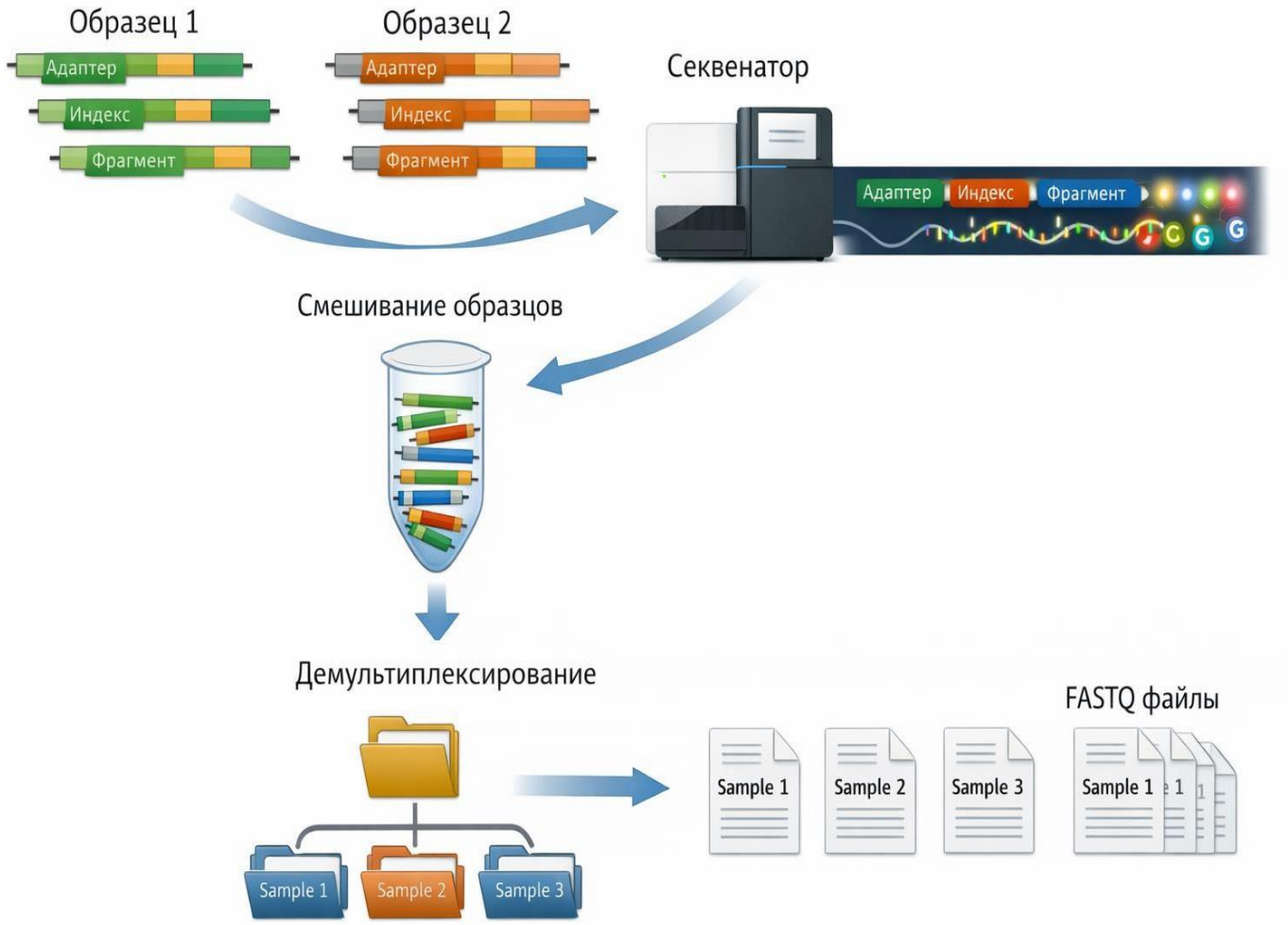
Адаптеры — это короткие «служебные» последовательности, которые добавляют к фрагментам ДНК
Они позволяют: (1) закрепить фрагмент на платформе, (2) запустить синтез праймером, (3) отличать образцы по индексам
Библиотека = множество фрагментов ДНК одинакового «формата», готовых для прибора



Источник изображения: http://tucf-genomics.tufts.edu/home/faq

Индексы (баркоды): как смешивать образцы в одном запуске

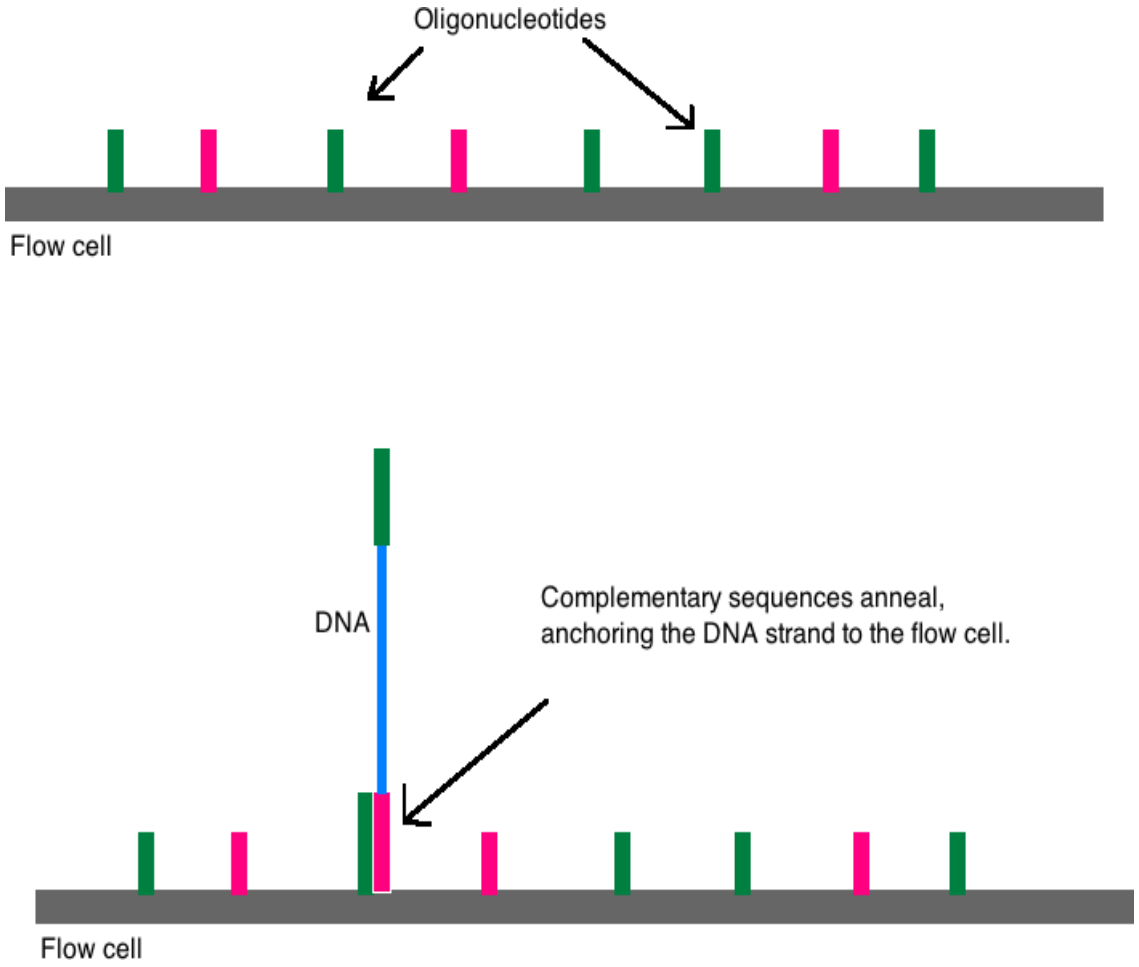
Индекс (barcode) — короткая последовательность, уникальная для образца
Если смешать 10 образцов, после секвенирования можно «разложить» reads обратно по индексам (демультиплексирование)
Это экономит деньги и время: один запуск → много образцов
Важно: индексы требуют аккуратности, иначе возможна «перепутка» reads



Illumina: flow cell — «стекло» с пришитыми олигонуклеотидами

Flow cell — это поверхность, где закрепляются фрагменты ДНК через комплементарные участки
Каждый фрагмент фиксируется в своей точке; дальше начинается локальное размножение
Так прибор создаёт много копий одного фрагмента, чтобы сигнал был заметен камере

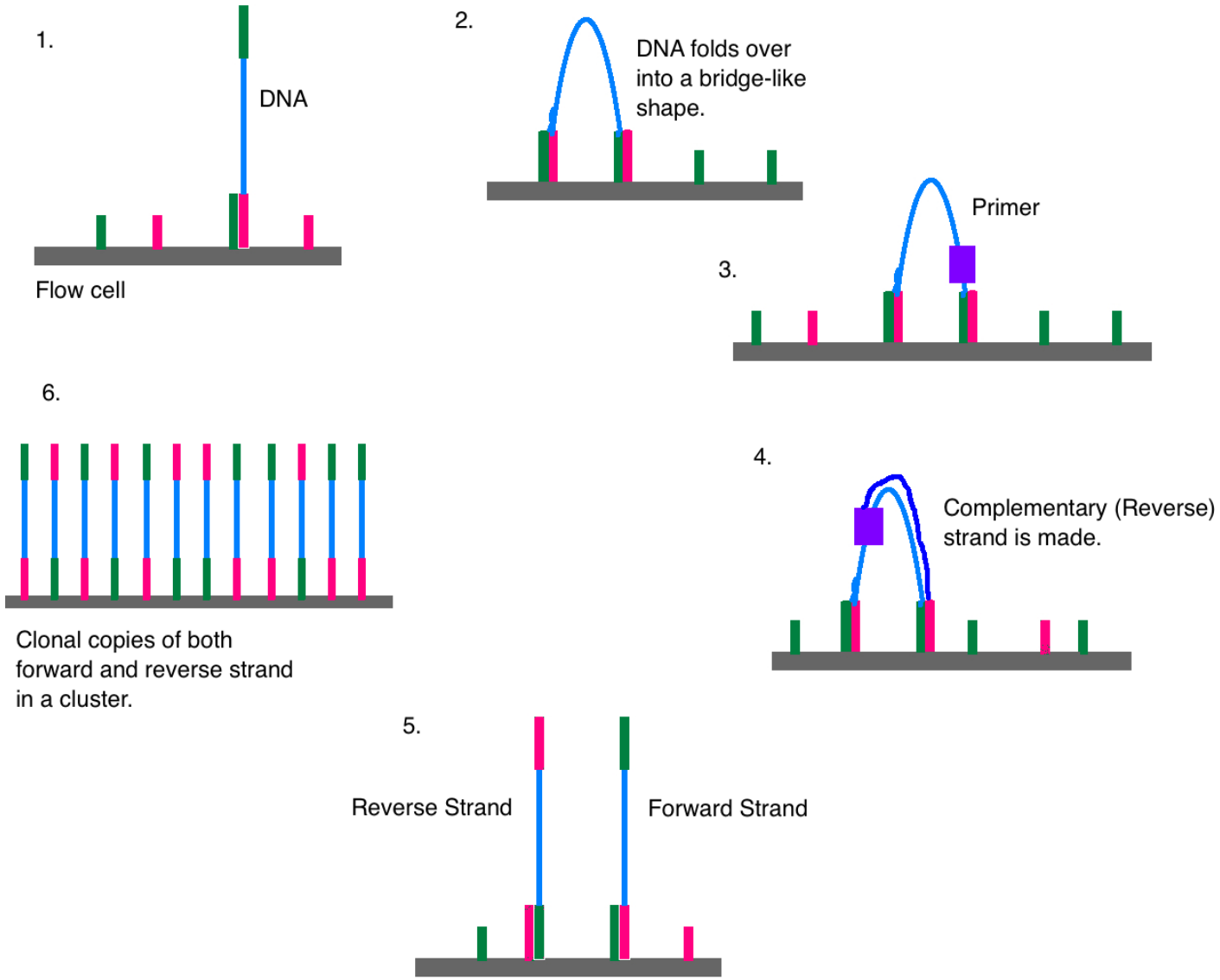
Illumina: flow cell — «стекло» с пришитыми олигонуклеотидами



Illumina: cluster generation (мостиковая амплификация)

Фрагмент «пригибается» и цепляется вторым концом — образуется «мостик»
Полимераза достраивает вторую цепь → получаются копии
Цикл повторяется → в одной точке формируется кластер тысяч одинаковых молекул
Кластер = усиление сигнала для чтения

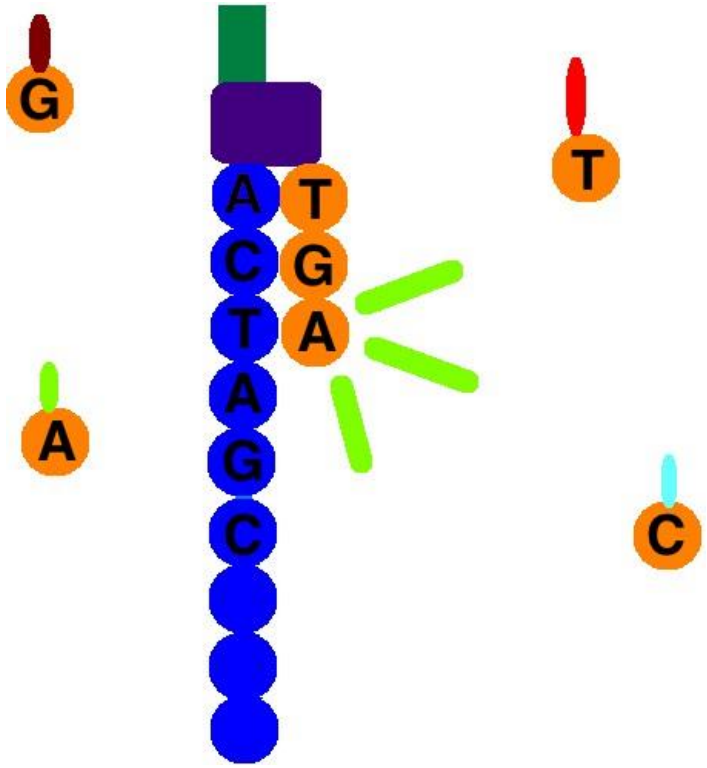
Illumina: cluster generation (мостиковая амплификация)



Источник изображения: Wikimedia Commons — https://commons.wikimedia.org/wiki/File:Cluster_Generation.png

Illumina: sequencing-by-synthesis (чтение при синтезе)

Добавляются 4 нуклеотида с метками и «обратимым стопом»
Встраивается только один нуклеотид за цикл → камера считывает цвет в каждой точке
Метка/стоп снимаются → следующий цикл
По последовательности цветов получаем последовательность bases



Источник изображения: Wikimedia Commons — https://commons.wikimedia.org/wiki/File:Sequence_By_Synthesis.png

Paired-end и глубина чтения: почему это важно

Paired-end: читаем фрагмент с двух концов → легче выравнивать и собирать
Глубина (depth) — сколько reads покрывают участок; выше глубина → выше уверенность
Для поиска редких вариантов (например, опухолевых) нужна высокая глубина
Но «всё зависит от задачи»: для бактериального генома одни числа, для человека другие

Какой сигнал превращается в файлы

Прибор фиксирует физический сигнал (свет/ток) и переводит его в буквы
Этот этап называется base calling (определение основания)
Далее добавляется информация о качестве каждой буквы
Результат почти всегда начинается с FASTQ-файлов

FASTQ: самый частый «сырой» формат данных

FASTQ хранит reads и качество каждой буквы
Каждый read записан 4 строками: идентификатор, последовательность, '+', строка качества
Качество кодируется символами и связано с вероятностью ошибки (Phred score)
Из FASTQ дальше делают выравнивание (BAM) и варианты (VCF)

```
@READ_0001
ACGTTGCA...
+
IIIIHGF...
```

- 4 строки на read
- 4-я строка кодирует качество

FASTQ = последовательность + качество.
Это «сырой» формат, с которого почти всегда начинается анализ.
Дальше: trimming → выравнивание → варианты / экспрессия.

Phred quality score: что означает «Q30»

Phred score Q = −10·log10(P ошибки)
Q20 ≈ 1% ошибка (1 из 100), Q30 ≈ 0,1% (1 из 1000), Q40 ≈ 0,01% (1 из 10 000)
Обычно смотрят долю оснований ≥ Q30
Важно: качество не одинаковое по длине reads — часто падает к концу

Q20 ≈ 1% ошибка (1 из 100)
Q30 ≈ 0,1% ошибка (1 из 1000)
Q40 ≈ 0,01% ошибка (1 из 10 000)
В отчётах часто смотрят долю оснований ≥ Q30.

Phred score: $Q = -10 \cdot \log10(P_{ошибки})$

QC данных: что проверяют после секвенирования

Распределение качества по позициям reads
Содержание адаптеров и необходимость тримминга
Смещение GC-состава (может указывать на контаминацию или смещение библиотеки)
Дубликаты (особенно при ПЦР) и «перекос» по представленности
В практике часто используют FastQC / MultiQC

Выравнивание и сборка: два пути анализа

Если есть референс (например, человек): выравниваем reads на геном → получаем BAM
Если референса нет: собираем reads в контиги (de novo assembly)
После выравнивания можно искать варианты (variant calling)
Для RNA-seq: выравнивание на транскриптом/геном + подсчёт экспрессии

Путь 1: есть референс

- Выравнивание reads на геном
- Получаем BAM/SAM
- Ищем варианты (VCF)
- Часто: клиника, патогены, WGS человека

Путь 2: референса нет

- Сборка (assembly) в контиги
- Скаффолдинг, полировка
- Аннотация генов
- Часто: новые организмы, метагеномика

BAM/SAM и VCF: какие файлы появляются дальше

SAM/BAM: где каждый read выровнен, с координатами и качеством выравнивания
VCF: список вариантов (позиция, референсная буква, альтернативная, качество, частоты)
Эти форматы позволяют хранить большие проекты компактно и стандартно
Для обмена данными важно уметь «читать шапку» и понимать поля

```
    SAM (упрощённо):
read1   chr112345   ...
read2   chr112380   ...

   BAM = бинарный SAM
```

```
      VCF (упрощённо):
#CHROM POS REF ALT QUAL ...
chr1  12345  A    G   99  ...
```

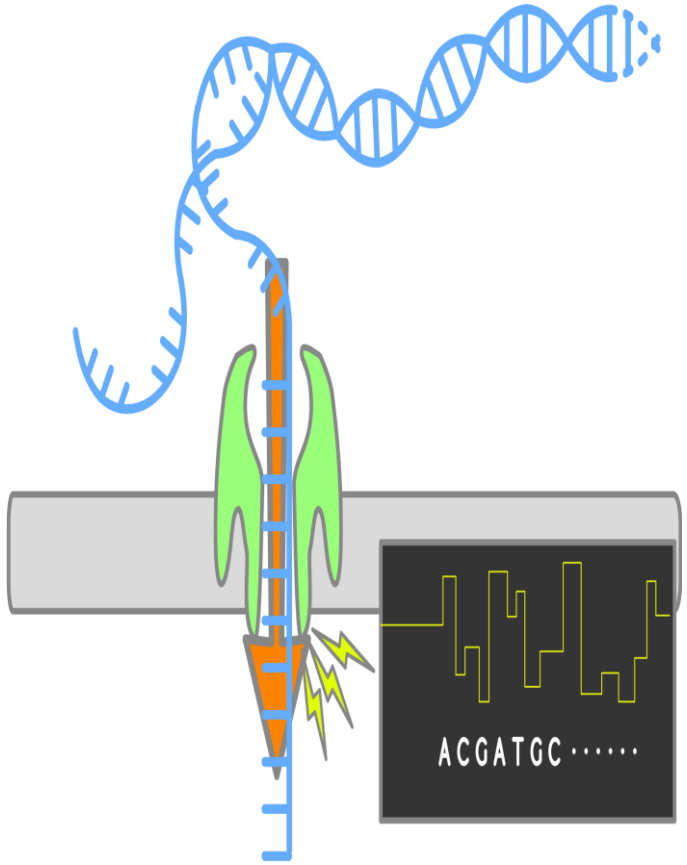
Эти форматы — «мост» между экспериментом и выводами: их нужно уметь читать хотя бы на базовом уровне.

3-е поколение: длинные прочтения (long reads)

Идея: читать одну молекулу ДНК «длинным куском», а не сотнями коротких
Плюсы: проще собирать геномы, видеть повторяющиеся участки и перестройки
Минусы: исторически выше ошибка на одном read (сейчас сильно улучшилось)
Типичные платформы: PacBio (SMRT) и Oxford Nanopore

Nanopore: принцип (пора + электрический ток)

Есть мембрана с нанопорой и приложенное напряжение → течёт ионный ток
Когда ДНК проходит через пору, ток меняется (сигнал зависит от последовательности)
Компьютер по сигналу восстанавливает буквы (base calling)
Плюс: чтение в реальном времени и очень длинные фрагменты



Источник изображения: Wikimedia Commons — https://commons.wikimedia.org/wiki/File:202001_nanopore_sequencing.svg

Nanopore: как выглядит прибор и что в нём происходит

Миниатюрные устройства (например, MinION) + расходник (flow cell) с тысячами пор
Секвенирование идёт в реальном времени: можно остановить, когда данных достаточно
Важный шаг — base calling (часто на видеокарте) и фильтрация по качеству
Nanopore удобно для полевых/быстрых задач и сборок геномов

Источник изображения: Wikimedia Commons — https://commons.wikimedia.org/wiki/File:Oxford_Nanopore_MinION_top.jpg

Фото: секвенаторы (пример форм-фактора)

Illumina MiSeq — типичный «настольный» NGS для небольших/средних задач
Oxford Nanopore MinION — переносной секвенатор, работает от ноутбука
Важно понимать: платформы отличаются не только химией, но и логикой данных и анализа

Источник изображения: Wikimedia Commons — https://commons.wikimedia.org/wiki/File:Illumina_MiSeq_sequencer_2.jpg | Wikimedia Commons — https://commons.wikimedia.org/wiki/File:Oxford_Nanopore_MinION_t

Сравнение платформ (очень упрощённо)

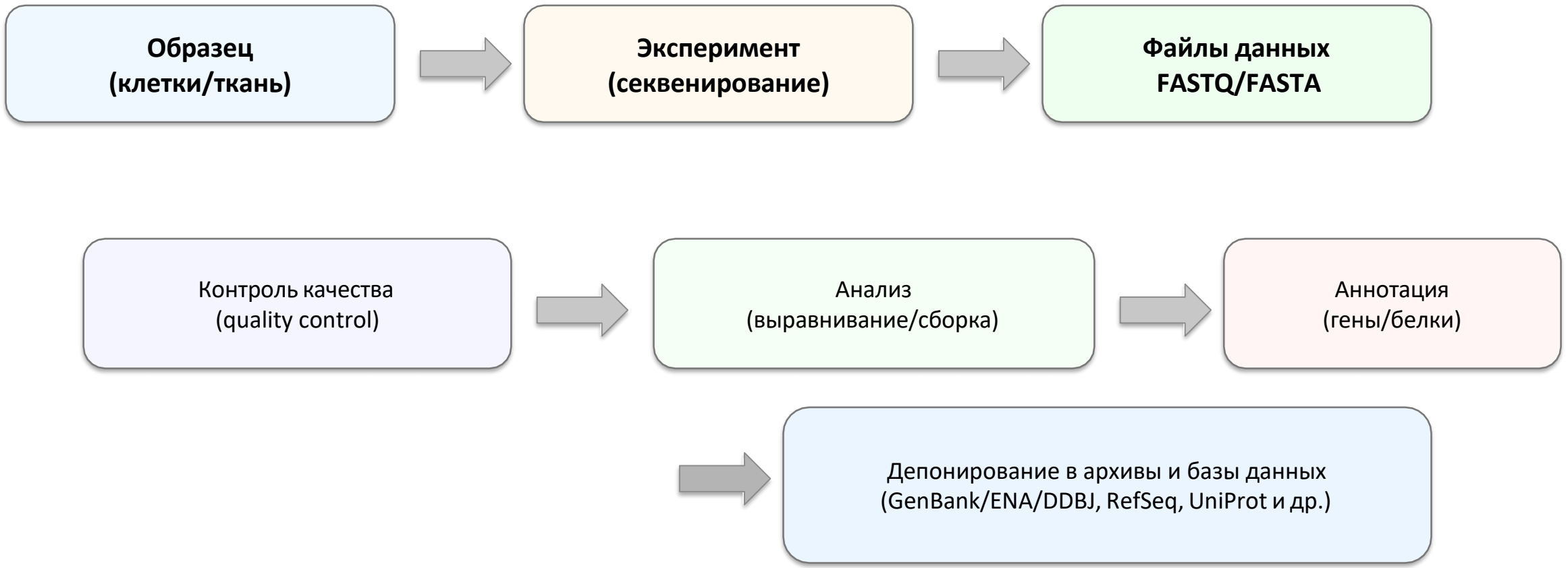
	Сэнгер	NGS (Illumina)	Long reads (PacBio/ONT)
Длина read	~700–1000	~50–300	~10 000–1 000 000+
Точность (1 read)	очень высокая	высокая	от средней до высокой*
Сильные стороны	валидация, плазмиды	масштаб, точность	сборка, перестройки
Слабые стороны	мало данных	короткие reads	другой анализ, сырой сигнал

Итоги

Секвенирование превращает молекулу в данные; дальше всё решает качество и анализ
Ключевые шаги: образец → библиотека → запуск → FASTQ → QC → анализ → интерпретация
Для старта: Illumina basics, вводные по FASTQ/QC

Источники данных в современной геномике

От биологического образца до записи в базе данных



Ключевая мысль: «данные» — это не только файлы, но и записи в базах с аннотацией и ссылками между объектами.

Три уровня: архив → курируемая база → специализированные р

1) Архив
(сырой «склад»)

Принимает все депонирования
Главное — сохранить
Аннотация может быть разной

2) Курируемая база
(«проверенные» записи)

Редакторы/алгоритмы улучшают качество
Единые правила именования
Меньше дубликатов

3) Специализированные
(под задачу)

Экспрессия (RNA-seq)
Структуры (PDB)
Протеомика (PRIDE)
Варианты (SNP/VCF)

На практике вы ходите «вверх-вниз»: нашли последовательность в архиве → уточнили белок в UniProt → вернулись к гену в NCBI → проверили, есть ли протеомные подтверждения.

Международные архивы последовательностей (INSDC)

INSDC = International Nucleotide Sequence Database Collaboration

GenBank
(NCBI, США)

ENA
(EMBL-EBI, Европа)

DDBJ
(Япония)

Смысл: вы можете загрузить последовательность в один архив, а через некоторое время она появится и в двух других.
Это один «мировой склад» данных.

Запись GenBank: что вы видите на экране

Условная «шапка» записи:

- LOCUS / DEFINITION — что это
- ACCESSION — постоянный ID
- VERSION — версия последовательности
- ORGANISM — организм
- REFERENCE — публикации
- FEATURES — разметка (gene, CDS...)
- ORIGIN — сама последовательность

Главные вопросы при чтении:

1) Что это за объект? (ген? плазмида? участок генома?)
2) Насколько запись надёжна? (RefSeq/курирование/публикация)
3) Где кодирующая часть? (CDS = coding sequence)
4) Есть ли белок? (перевод в FEATURES)
5) Какой идентификатор использовать в отчёте/домашке?

Accession и VERSION: «паспорт» записи

Accession (идентификатор) — это «номер паспорта» записи, который обычно не меняется.

Пример (условно):
ACCESSION: AB123456
VERSION: AB123456.1

Что значит «.1»?
Это версия последовательности.
Если автор исправит ошибку (букву/участок), версия станет .2 и т.д.

Практическое правило: в отчётах удобно указывать accession + версию (например, AB123456.1), чтобы было ясно, с какой именно последовательностью вы работали.

FEATURES: где «живет смысл»

FEATURES — таблица разметки: какие элементы есть на этой ДНК и где они расположены.

gene
(границы гена)

mRNA/exon
(для эукариот)

у прокариот часто нет экзонов

CDS
(кодирующая часть)
+ /translation
(аминокислотная последовательность)

FASTA и «аннотированная запись» — в чём разница

FASTA:
>заголовок
ATGCGT...

Плюсы:
- простой формат
- удобно для выравнивания

Минусы:
- почти нет контекста
- непонятно, где ген/белок

GenBank/ENA-запись:
- описание объекта
- ссылки на статьи
- FEATURES (разметка)
- перевод белка
- перекрёстные ссылки на другие базы

Это «файл + смысл вокруг файла».

Белковые базы данных: зачем они отдельно

Белок — это не просто перевод из ДНК: у него может быть несколько изоформ, домены, сигнальные пептиды, пост-трансляционные модификации.
Белковые базы собирают сведения о функции, доменах, локализации, участии в путях, известных вариантах.
Очень часто удобнее начинать с белка (функция понятнее), а потом возвращаться к гену и последовательности.

UniProtKB:
Swiss-Prot (курируемые записи)
TrEMBL (автоматически аннотированные)

NCBI Protein:
белковые последовательности и ссылки
на соответствующие нуклеотидные записи

Связь ДНК → РНК → белок: как не запутаться

Что обычно встречается в базах:

- у гена есть имя/идентификатор (GeneID и т.п.)
- у транскриптов и белков — свои accession/версии
- в GenBank запись может содержать и ДНК, и перевод белка (в CDS /translation)
- UniProt чаще даёт «лучшее» описание функции белка, а NCBI — удобные ссылки на геном

Трансляция in silico: перевод триплетов в аминокислоты

in silico = «на компьютере». Мы переводим кодоны (триплеты) ДНК/РНК в аминокислоты по генетическому коду.

ATG GAA TTT … → Met Glu Phe …

Почему это важно?
- помогает проверить, что CDS задан правильно
- даёт белок для поиска мотивов/доменных баз
- используется в протеомике для сопоставления пептидов

Типичные ошибки новичков:
- перепутали цепь (+/−)
- не учли рамку считывания
- переводят участок, где есть интроны
- забыли про стоп-кодон

«Выравнивание» последовательностей: идея простыми словами

Выравнивание — это попытка сопоставить две (или больше) последовательности так, чтобы увидеть сходство и различия.

```
Seq1: ATGCTGACCT---GAT
      |||| ||||   ||
Seq2: ATGCT-A CCTTTGAT
```

Зачем: (1) найти «похожий» ген/белок, (2) оценить консервативность, (3) найти мутации/инделы, (4) построить филогению.

Как читать результат BLAST (без математики)

Identity (процент совпадений): насколько похожи буквы/аминокислоты в выровненном участке.
Coverage (покрытие): какую долю вашей последовательности удалось сопоставить.
E-value: насколько вероятно получить такое совпадение «случайно» (меньше — лучше).
Важно: высокая identity на очень коротком участке может ничего не значить — всегда смотрите coverage.
Если выравнивание белковое, оно обычно «переносимее» между видами, чем нуклеотидное.

Мини-кейс: по одному скриншоту BLAST решить — это «тот же белок» или просто общий домен?

Белковые мотивы и домены: простая интуиция

Мотив — короткий повторяющийся «узор» (паттерн) в белке. Домен — более крупный «модуль», который часто соответствует функции.

Белок можно представить как «швейцарский нож» из модулей: [Домен A]—[Домен B]—[Домен C].

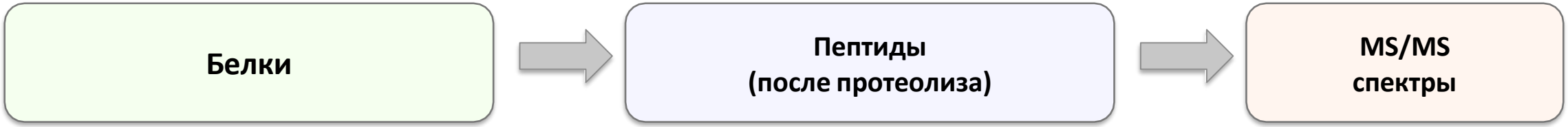
Зачем это нужно?
- помогает догадаться о функции
- объясняет, почему похожие белки есть у разных организмов
- помогает выбрать участки для экспериментов

Где смотреть:
- UniProt (раздел Features)
- InterPro/Pfam (доменная аннотация)

Сегодня: общий принцип, без углубления.

Где в этой истории протеомика и масс-спектрометрия

Протеомика — это про набор белков в клетке/условии. Часто основной метод — масс-спектрометрия.



Как это связывается с геномикой:

- чтобы понять, какой белок дал спектр, нужен «словарь» возможных белков
- этот «словарь» строится из геномных данных (аннотации CDS → белки)
- поэтому базы ДНК и белков — основа протеомики

Мини-кейс 1 «прочитать» запись GenBank

Задача на 7–10 минут (в парах/мини-группах).

Откройте выбранную преподавателем запись GenBank (ссылка/ID).
Найдите: ORGANISM, ACCESSION, VERSION.
Найдите в FEATURES элемент CDS и его координаты.
Скопируйте /translation (аминокислотную последовательность) или хотя бы первые 30 аминокислот.
Ответьте: это полный белок или фрагмент? (подсказка: смотрите длину/комментарии).

Мини-кейс 2: найти белковую запись и связать с геномом

Задача на 7–10 минут.

По /translation или Protein ID (если указан) найдите запись в NCBI Protein или UniProt.
Найдите, как эта белковая запись ссылается на исходную нуклеотидную (геномную) запись.
Найдите домены/Features (если есть) и сделайте очень короткий вывод «что это за белок».
Сформулируйте 1–2 предложения: что вы поняли о функции и на чём основан вывод.

Мини-кейс 3: BLAST «по-человечески»

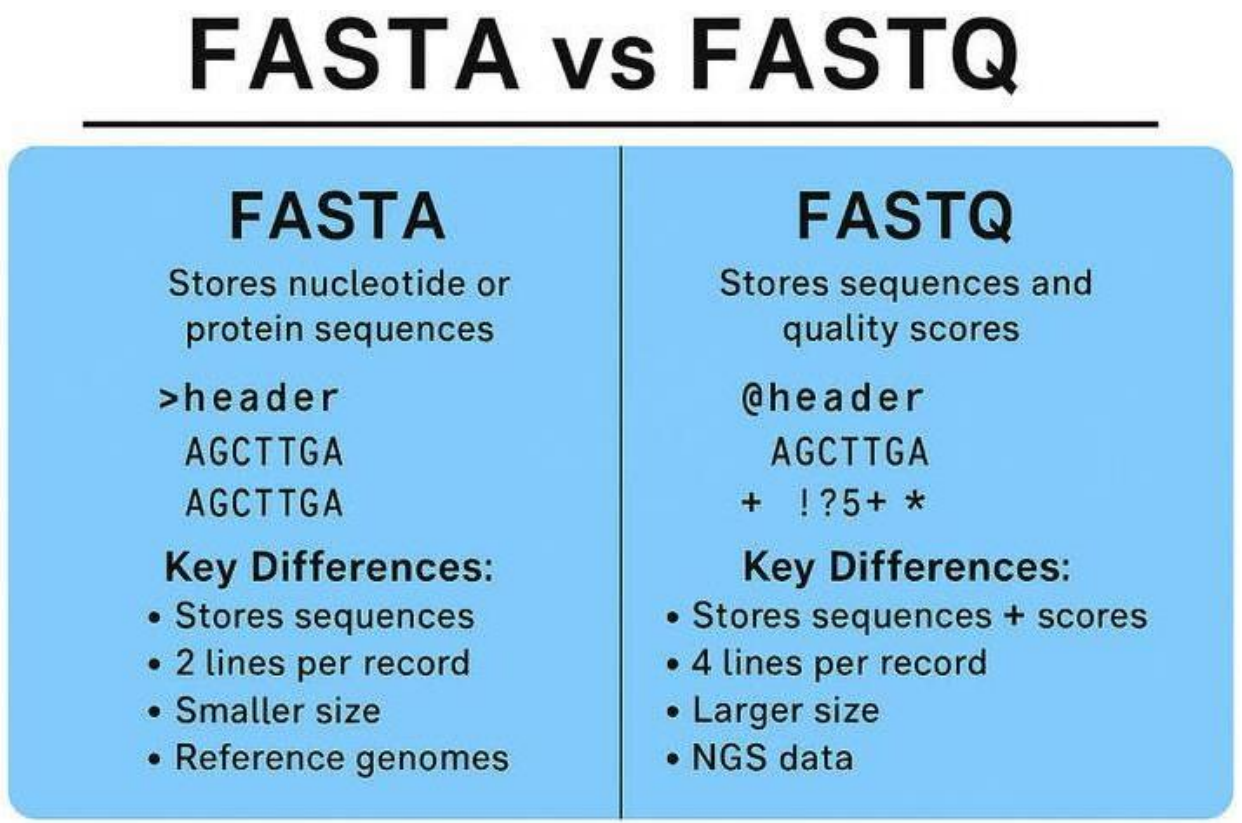
Задача на 10–12 минут.

Запустите BLAST (nucleotide или protein — как даст преподаватель).
Выберите 1 лучший hit и 1 «подозрительный» hit (похожий, но не до конца).
Сравните: identity, coverage, E-value (на уровне интуиции).
Ответьте: почему лучший hit — лучший? что не так со вторым?
Сделайте вывод: это гомологичный белок/ген или «совпадение домена/участка»?

Контроль качества данных секвенирования

Зачем нужен QC при секвенировании

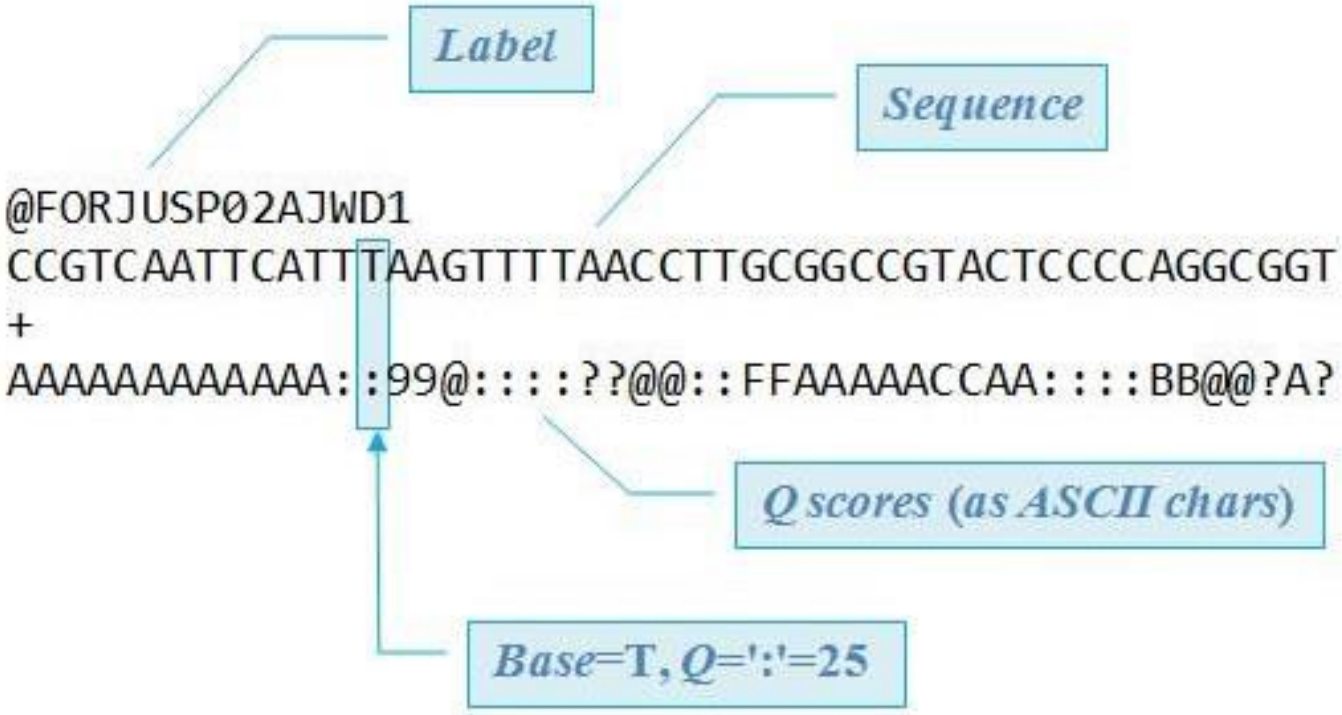
- QC = проверка пригодности данных к анализу
- Ошибки на входе = артефакты на выходе
- QC делают до выравнивания, сборки и поиска вариантов



https://www.linkedin.com/posts/bulut-hamali_understanding-fasta-and-fastq-core-file-activity-7338194481399123968-oYpB

Где QC стоит в пайплайне и что такое FASTQ?

- FASTQ → QC → trimming/фильтрация → выравнивание или сборка → BAM/VCF → анализ
- FASTQ хранит:
- Последовательность
- качество чтения
- 1 read = 4 строки



https://www.drive5.com/usearch/manual/fastq_files.html

Что такое quality score

- Quality score = оценка вероятности ошибки
- Чем выше Q, тем надёжнее base call
- Используется шкала Phred

Формула Phred

- Q = −10 log10(P)
- P = вероятность ошибочного base call

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

https://learn.gencore.bio.nyu.edu/ngs-file-formats/quality-scores/

Что значат Q10, Q20, Q30

- Q10 → ошибка 1/10
- Q20 → ошибка 1/100
- Q30 → ошибка 1/1000
- Q40 → ошибка 1/10000

Как качество записано в FASTQ

- Quality хранится как ASCII-символы
- Чаще всего используется Phred+33
- Длина quality-строки = длина read
- Phred+33 — это способ записывать числовой quality score в FASTQ через ASCII-символ. ASCII code = Phred score + 33. Пусть quality score = 30. ASCII-код 63 — это символ ?

Что такое FastQC

- FastQC = инструмент первичной диагностики FASTQ
- Показывает PASS / WARN / FAIL
- Даёт быстрый обзор проблемных мест

Как читать отчёт FastQC

- Сначала Summary
- Потом ключевые графики
- Затем решение: что делать

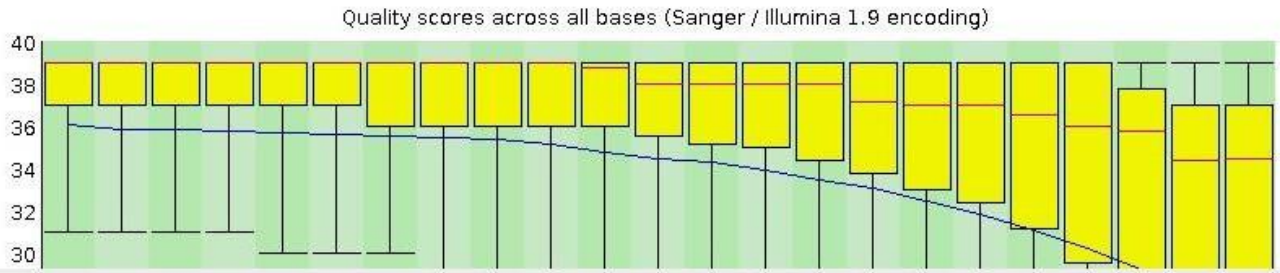
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

Measure	Value
Filename	WES_human_Illumina.pe_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4942814
Filtered Sequences	0
Sequence length	76
%GC	47

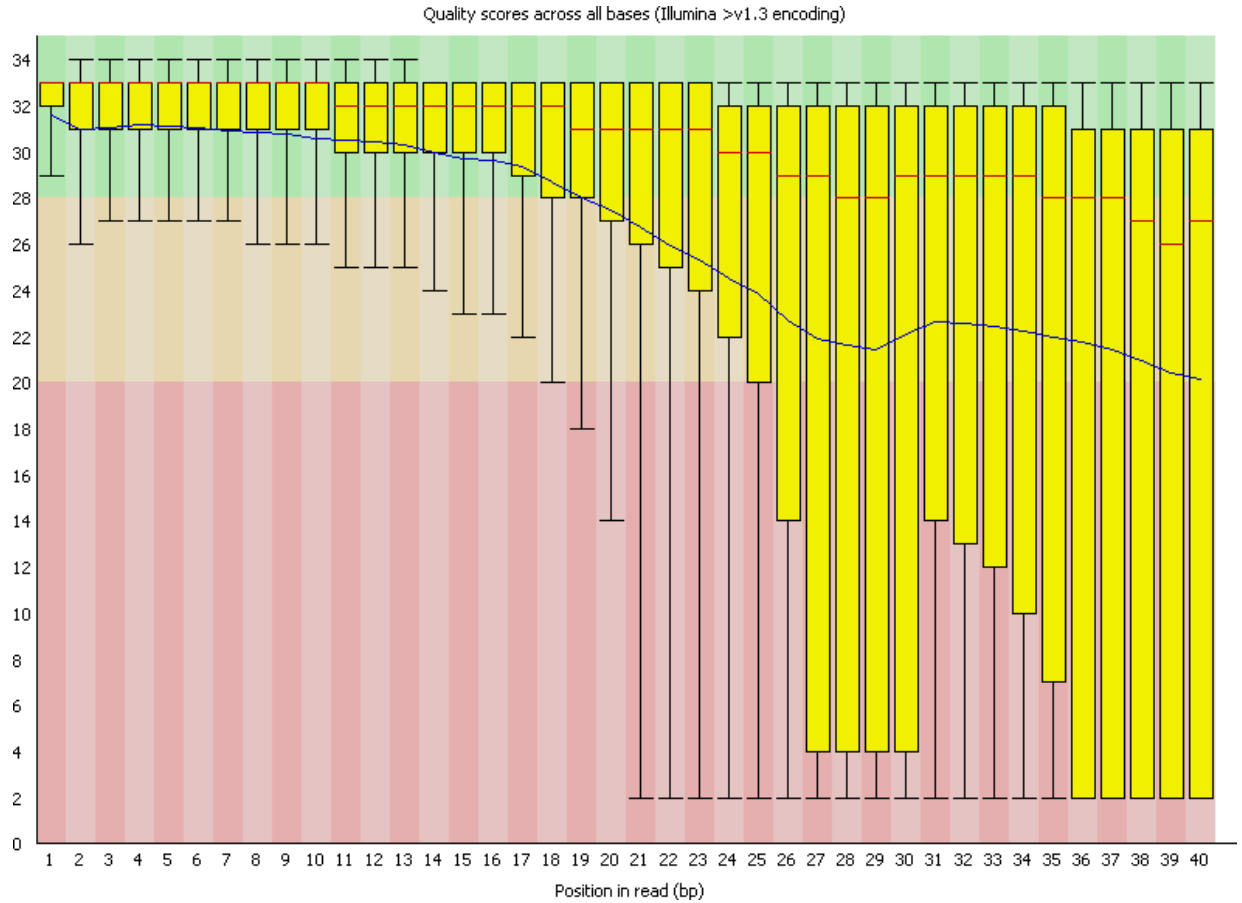
Per base sequence quality



Produced by FastQC (version 0.10.1)

Per base sequence quality

- Качество по каждой позиции в read
- По оси X — позиция
- По оси Y — Q-score
- Для каждой позиции строится boxplot



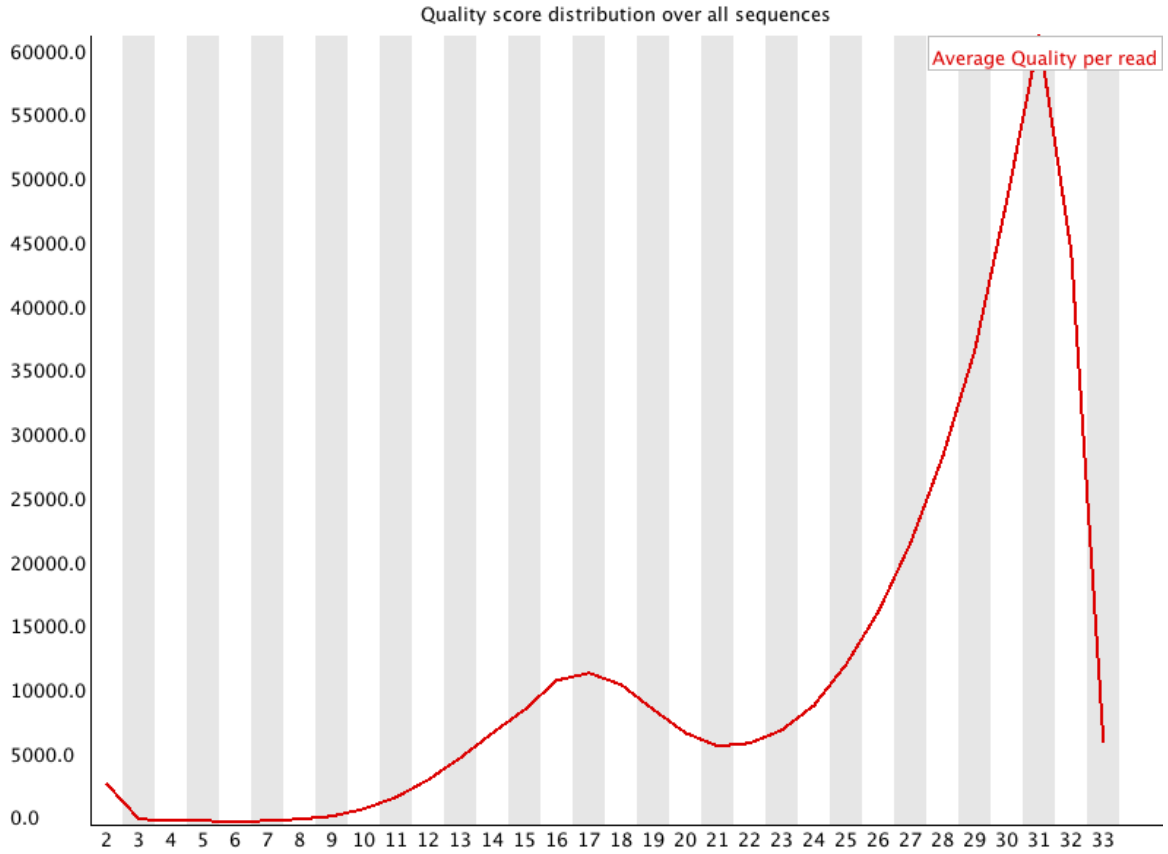
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality.html

Как интерпретировать падение качества

- Небольшое падение к 3’-концу — обычно нормально
- Резкий обвал качества — проблема
- Хвост часто убирают trimming’ом

Per sequence quality scores

- Среднее качество по каждому read
- Один пик на высоком качестве — хорошо
- Хвост низкокачественных reads — повод фильтровать



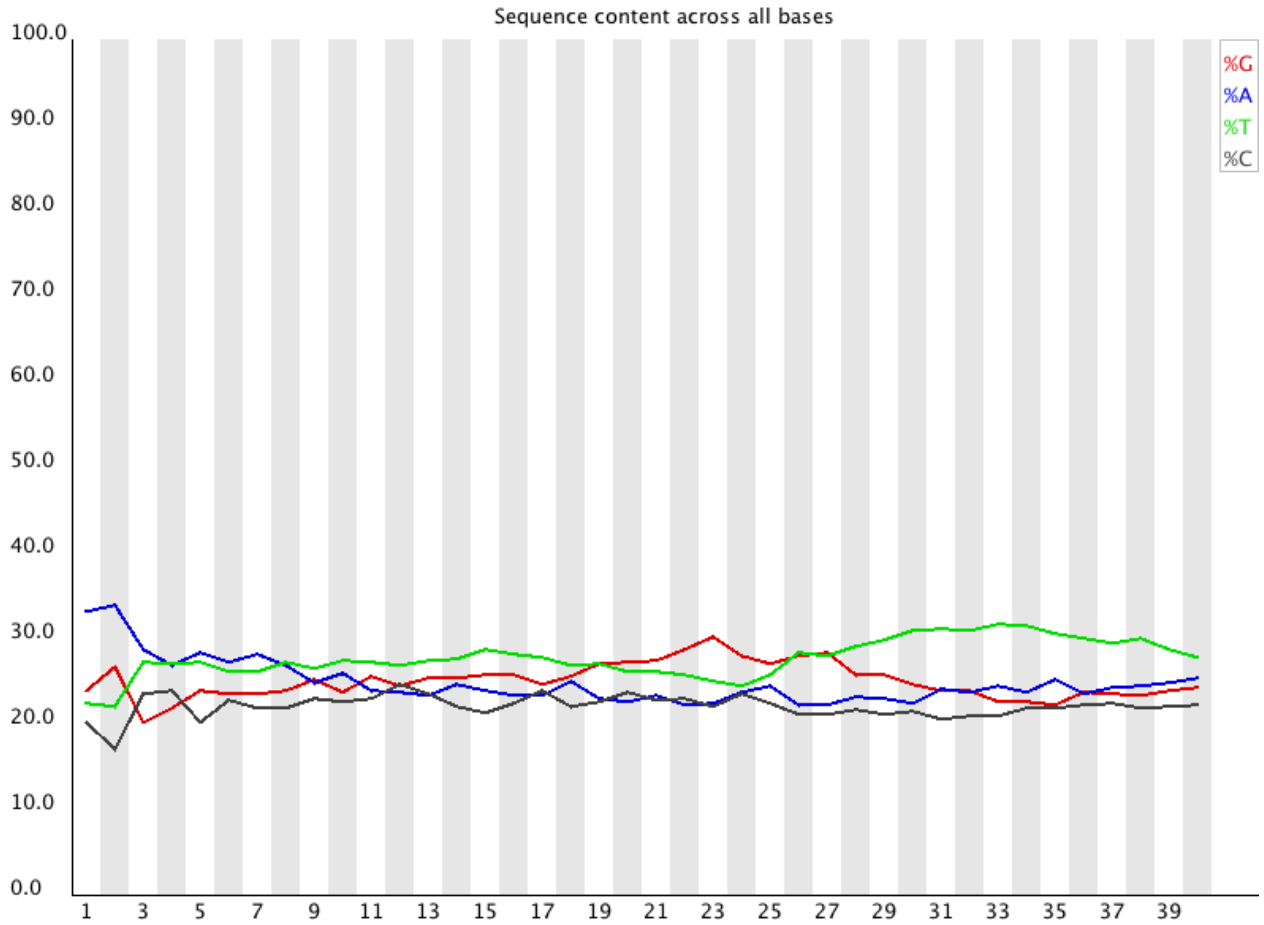
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/3%20Per%20Sequence%20Quality

%20Scores.html

Per base sequence content

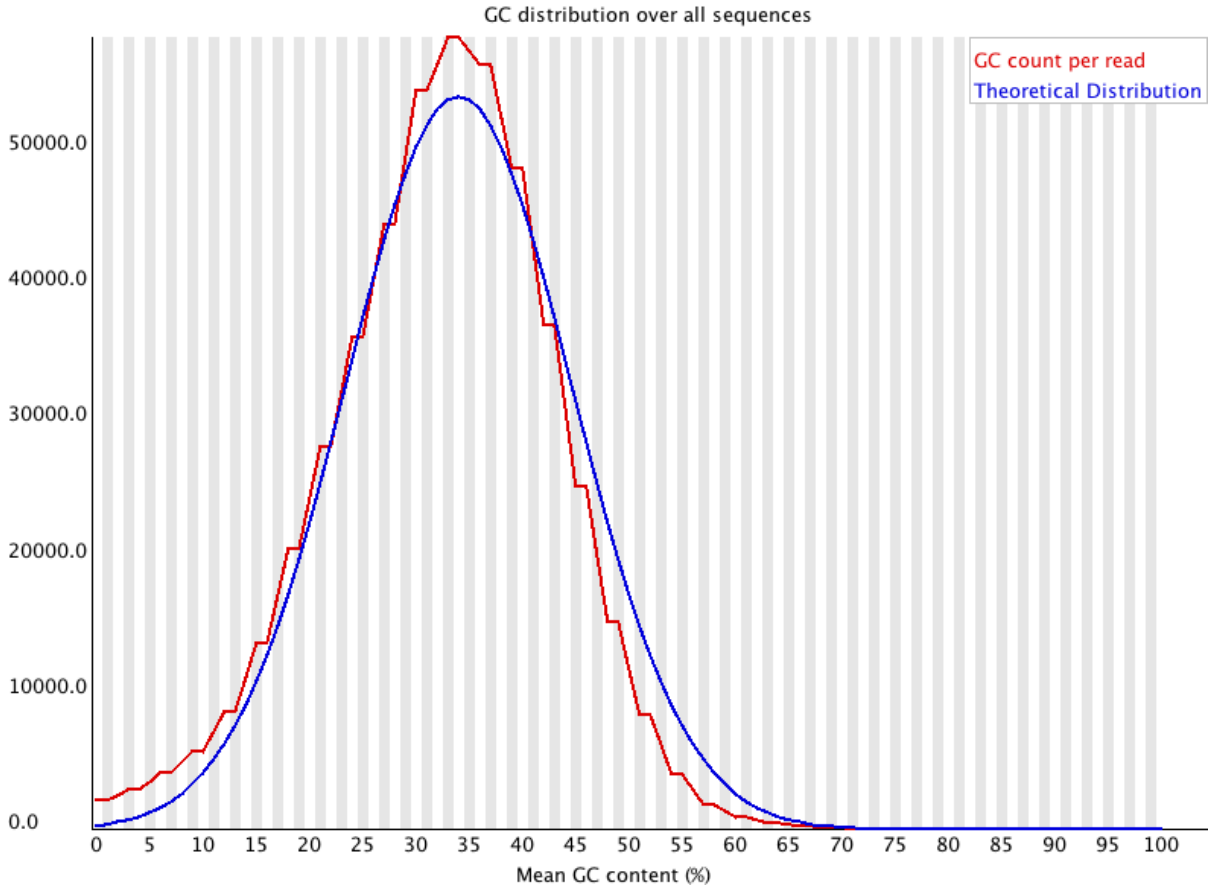
- Доли A, T, G, C по позициям
- Ровный профиль — не всегда обязателен
- Перекосы нужно интерпретировать в контексте эксперимента

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/3%20Per%20Sequence%20Quality%20Scores.html



GC состав и длина reads

- GC состав: общий профиль библиотеки
- Распределение длин последовательностей: распределение длин reads
- Необычный GC-профиль или слишком короткие reads требуют внимания



Overrepresented sequences, adapters, k-mers

- Часто встречающиеся последовательности
- Возможные причины:
- Адаптеры
- Праймеры
- РНК-контаминация
- Adapter content часто указывает на read-through

Что делать с проблемными данными

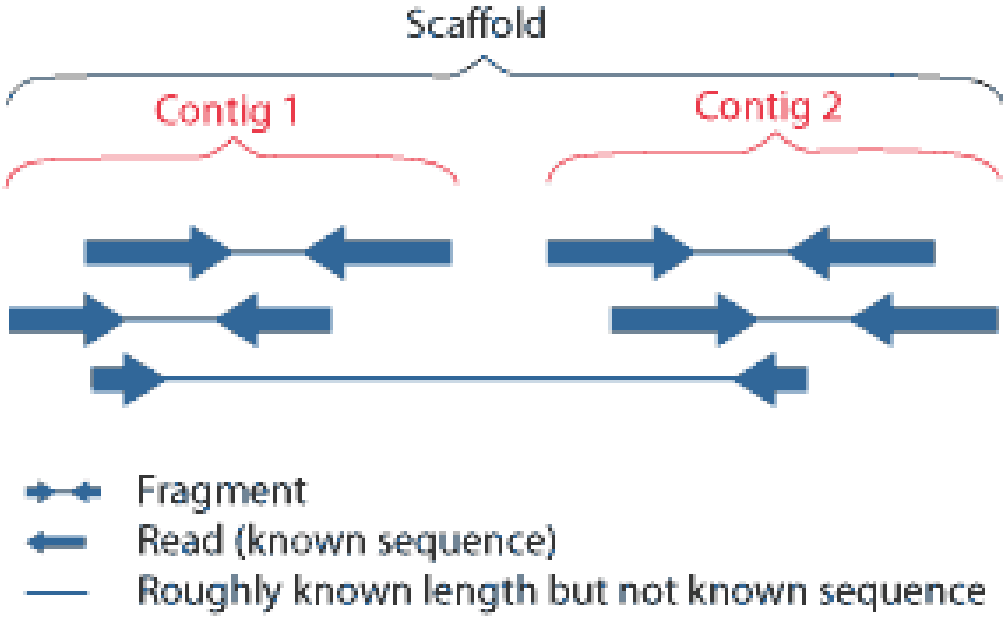
- Удалить адаптеры
- Подрезать низкокачественные концы
- Удалить слишком короткие reads
- Повторить QC
- Trimming, filtering, deduplication

Инструменты и итоговый вывод

- Инструменты: cutadapt, Trimmomatic, fastp
- QC влияет на:
- Mapping
- BAM/SAM
- variant calling
- количественный анализ
- Итог: FastQC — инструмент принятия решений

Почему после секвенирования геном не получается сразу “цельным”

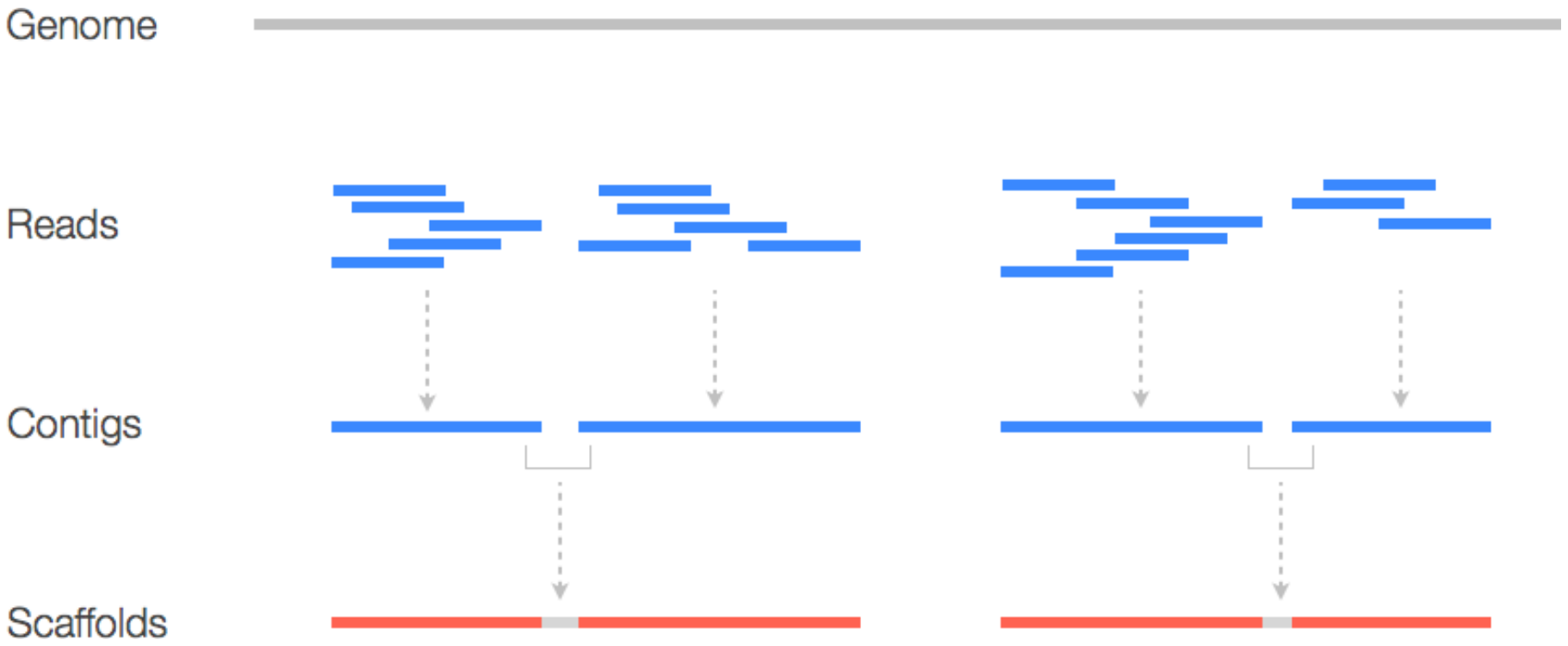
•**Сборка генома** — это восстановление более длинной последовательности из множества коротких reads. На выходе сначала обычно получают не целый геном, а набор **контигов и скаффолдов.**



https://mycocosm.jgi.doe.gov/help/scaffolds.jsf

Что такое контиг

•Контиг (contig) — это непрерывная собранная последовательность, полученная из перекрывающихся reads без неизвестных участков внутри.



https://bioinformaticamente.com/2020/12/16/assembly-of-reads/

Почему сборка распадается на несколько контигов

- Сборка часто прерывается из-за:
- повторяющихся последовательностей;
- недостаточного покрытия;
- ошибок секвенирования;
- сложных по структуре участков генома.

Что такое скаффолд

•**Скаффолд (scaffold)** — это набор контигов,расположенных в правильном порядке и ориентации,с оценкой расстояний между ними.

Чем контиг отличается от скаффолда

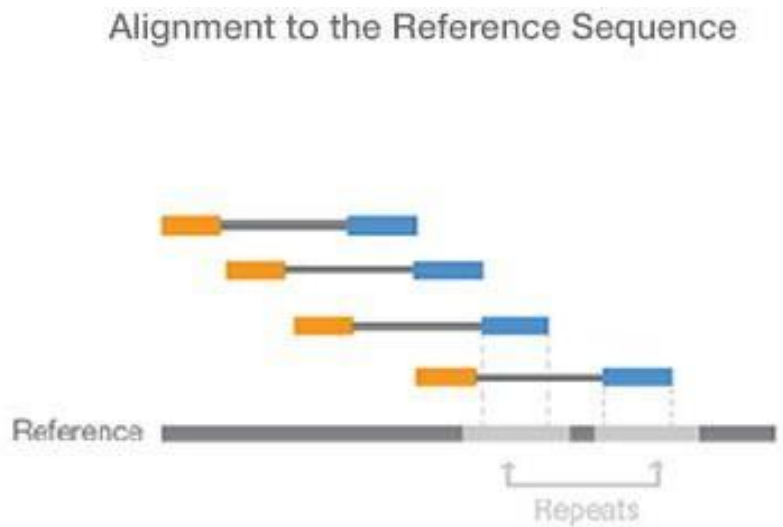
- **Контиг** = непрерывная известная последовательность
- **Скаффолд** = несколько контигов + порядок + ориентация + расстояния между ними

Как в скаффолде обозначают неизвестные промежутки

• Промежутки между контигами в скаффолде
обычно обозначают символами **N**.
Это значит: участок здесь есть,
но его точная последовательность пока неизвестна.

За счёт чего контиги объединяют в скаффолды

- Контиги объединяют в скаффолды по информации о связи между ними:
- paired-end reads;
- mate-pair / long-insert libraries;
- длинные риды;
- картирование на референс или дополнительные методы.



Библиотеки с протяжёнными клонированными фрагментами ДНК

•Библиотеки с длинными клонированными фрагментами ДНК помогают упорядочивать контиги в скаффолды, потому что:
•связывают удалённые участки генома;
•задают порядок и ориентацию контигов;
•позволяют оценить расстояние между ними.

Логика сборки: от reads к контигам и скаффолдам

Reads → contigs → scaffolds → более полная сборка генома

- reads дают локальную информацию;
- контиги собирают непрерывные участки;
- скаффолды связывают эти участки в более крупную структуру.

Главная идея, которую нужно запомнить

Reads → contigs → scaffolds → более полная сборка генома

- **Контиг** показывает, что удалось собрать непрерывно.

Скаффолд показывает, как несколько контигов расположены друг относительно друга.
Библиотеки с длинными фрагментами ДНК помогают связать контиги в более крупные структуры.

Сравнительная геномика

Функциональная аннотация генов, геномные сравнения и основные инструменты

Сравнительная геномика: какие вопросы она решает

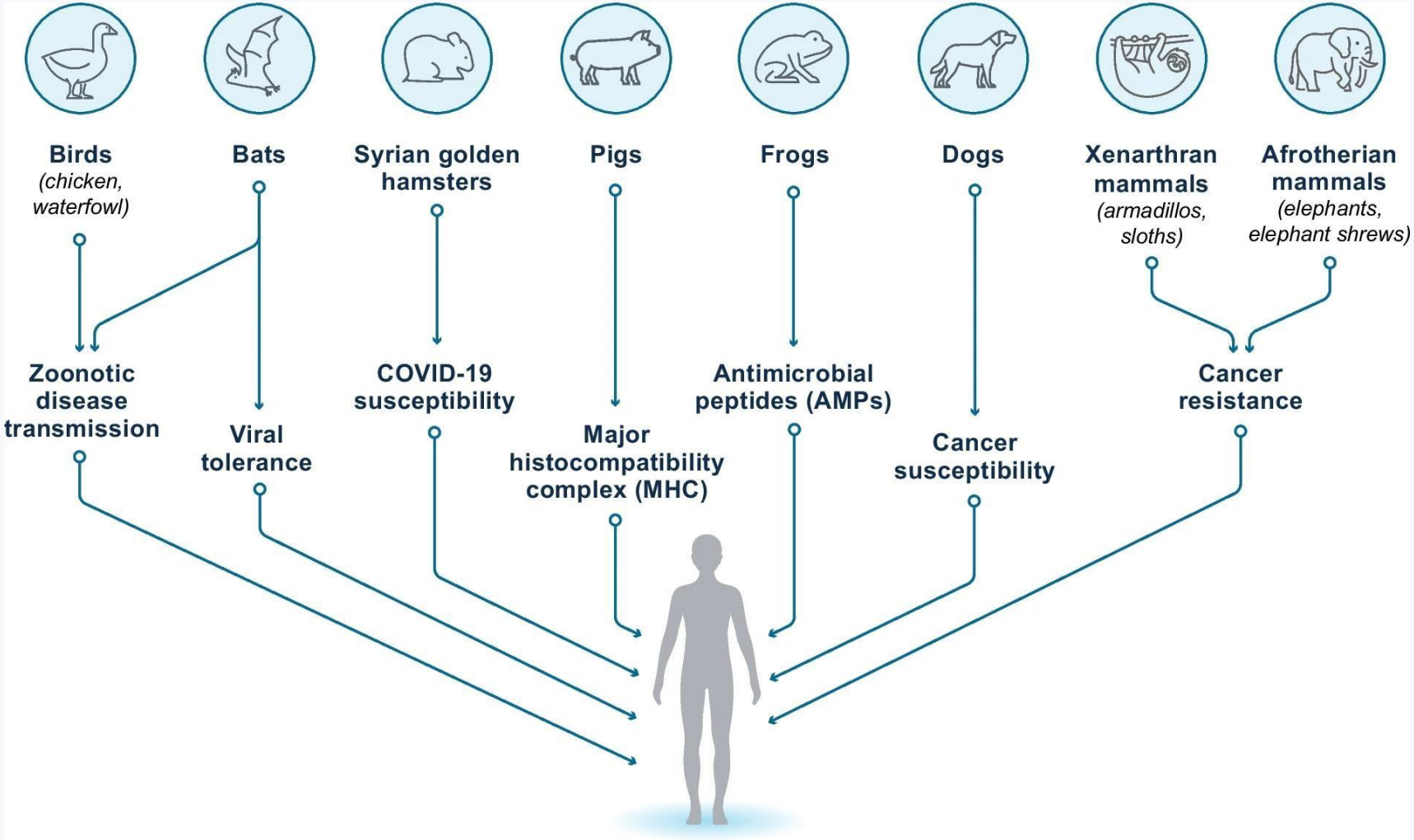
- Что общего и что уникально в наборах генов у разных организмов?
- Насколько похожи геномы по последовательности, составу и архитектуре?
- Какие события лежат в основе различий: дупликации, потери, перестройки, горизонтальный перенос?

Главная идея

Сравнение само по себе является источником биологической информации.

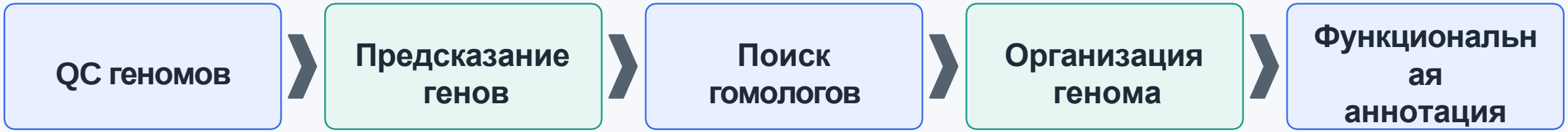
Один геном даёт описание. Набор геномов даёт контекст: эволюционный, функциональный и структурный.

Сравнительная геномика: какие вопросы она решает



Bornstein K. et al. The NIH Comparative Genomics Resource: addressing the promises and challenges of comparative genomics on human health //BMC genomics. – 2023. – Т. 24. – №. 1. – С. 575.

Типовой путь сравнительно-геномного анализа



- На входе нужны сопоставимые по качеству данные: полнота сборки, контаминация, единообразие аннотации.
- На выходе исследователь получает не просто список генов, а модель различий между геномами и гипотезы о функции.

«Паспорт» генома: с чего начинается сравнение

- длина генома в парах оснований и число молекул (хромосомы/плазмиды)
- оценка молекулярной массы генома
- число белок-кодирующих генов и кодирующая плотность
- GC-состав, повторный контент, кодонное использование, k-мерный проф иль
- качество данных: полнота сборки, ф рагментация, контаминация, устойчивость аннотации

Почему это важно

Грубые метрики быстро показывают аномалии и задают рамку для интерпретации.

Например, неожиданно большой размер генома может означать не только новые ф ункции, но и повторы, плазмиды или деф екты сборки.

Размер генома, молекулярная масса и число генов

Количественные ориентиры

- для дц ДНК часто используют оценку ≈ 660 г/моль на 1 п.о.
- молекулярная масса помогает связать длину генома с молекулярными количествами
- но в сравнительной геномике важнее не сама масса, а её связь с длиной и числом генов

Что нельзя упрощать

- одинаковый размер генома не означает одинаковое число генов
- прокариоты обычно компактнее и кодирующая плотность у них выше
- у эукариот значительный вклад в размер генома дают интроны, повторы и межгенные области

GC%, кодонное использование и количественная оценка сходства

- GC-состав можно рассматривать как композиционную «подпись» генома; локальные отклонения часто заставляют думать о горизонтальном переносе генов или геномных островах.
- Кодонное использование (codon bias) связан и с эволюционной историей, и с уровнем экспрессии.
- Для общей близости геномов у прокариот широко используют ANI; ориентир около 95% часто соответствует границе вида.
- AAI и k-мерные/MinHash-подходы полезны как дополнительные и быстрые меры сходства.

Практический вывод

Сравнивать нужно и содержание, и состав, и организацию.

Одна метрика почти никогда не даёт полной картины.

Организация генома: синтения и перестройки

Что такое синтения

- сохранение порядка генов или крупных блоков между геномами
- особенно полезна для анализа эволюции и для интерпретации функции в прокариотах
- сохранённые кластеры часто указывают на совместную работу генов

Что нарушает синтению

- инверсии и транслокации
- вставки и делеции
- острова патогенности, плазмиды, профаги, горизонтальный перенос генов
- дупликации и крупные перестройки

Гомология: ортологи, паралоги, ксенологи

Гомологи

общий эволюционный предок

Ортологи

расхождение через видообразование

Паралоги

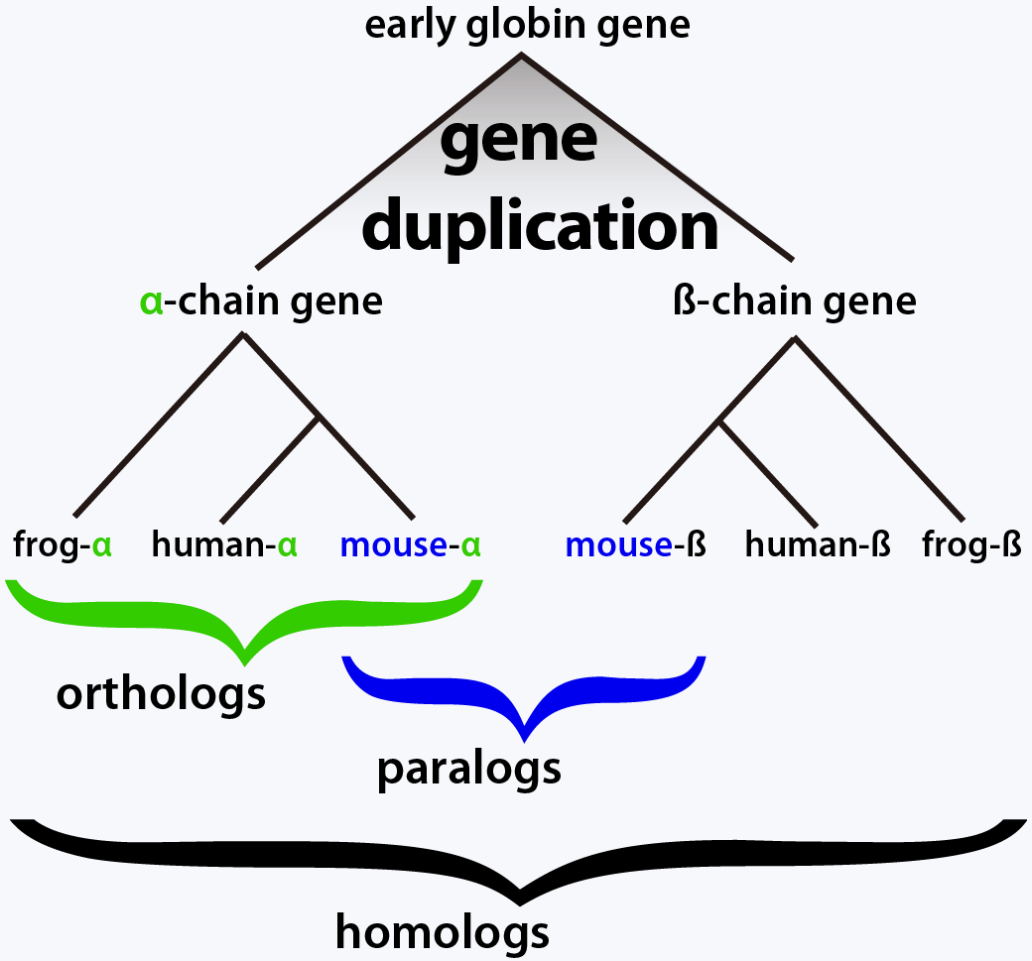
расхождение после дупликации

Ксенологи

связь через горизонтальный перенос

Ключевой вывод: похожий белок — ещё не обязательно функционально эквивалентный белок.

Гомология: ортологи, паралоги, ксенологи



https://commons.wikimedia.org/wiki/File:Homology.png

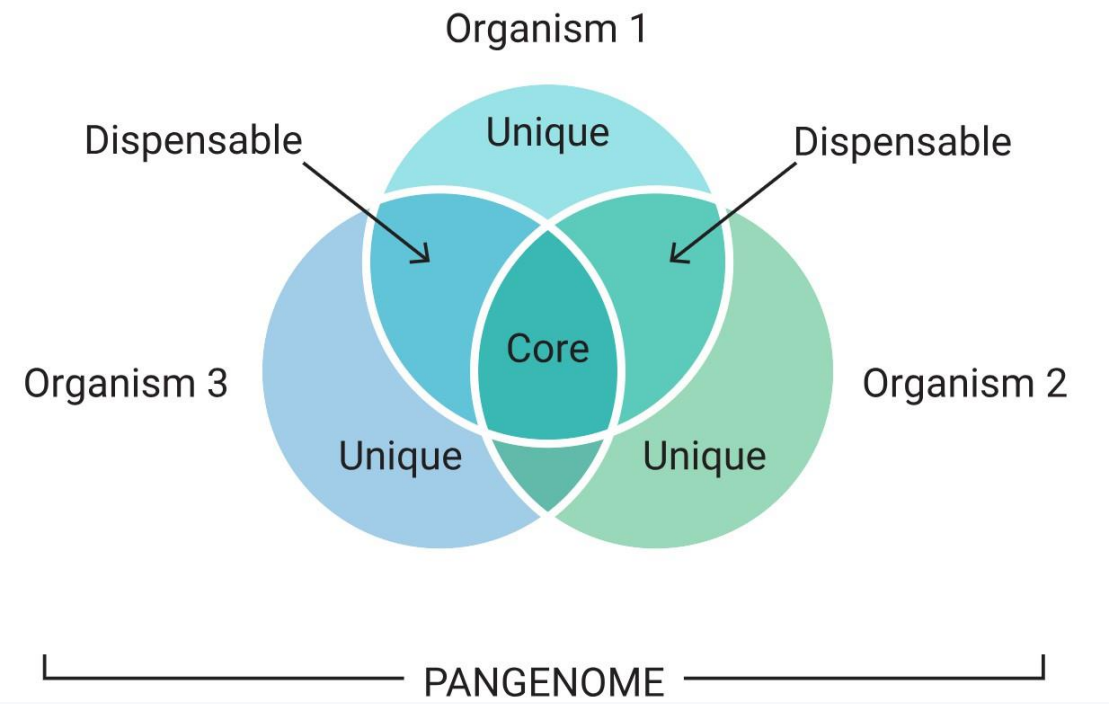
Core genes, accessory genes и пангеном

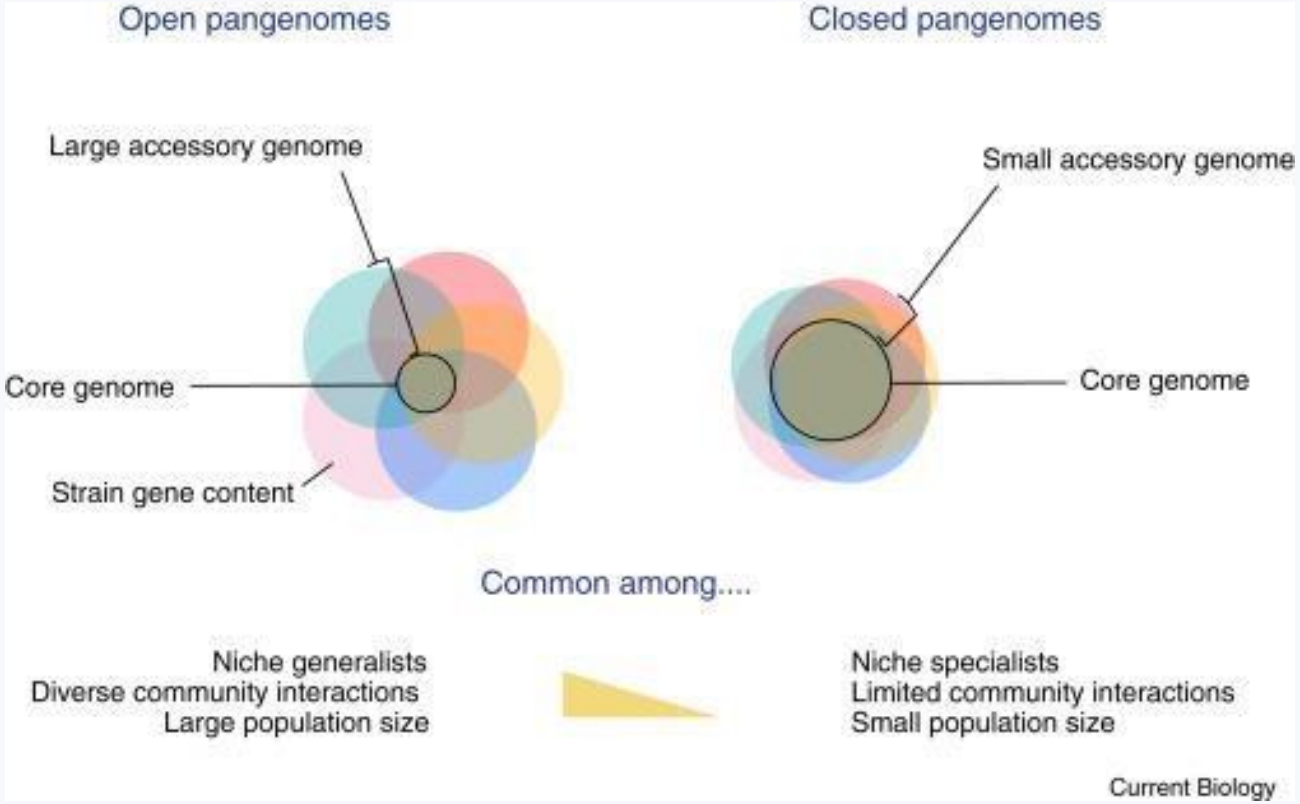
- Core genome — гены, встречающиеся у всех или почти всех представителей рассматриваемой группы.
- Accessory genome — вариабельная часть, часто связанная с адаптацией к нише, устойчивостью, патогенностью и пластичностью метаболизма.
- Unique genes — гены, обнаруженные только у одного генома или узкой подгруппы.

«Уникальный ген» часто требует перепроверки.

Иногда это новая функция, а иногда — пробел сборки, ошибка аннотации или слишком жёсткий критерий кластеризации.

Core genes, accessory genes и пангеном





https://www.pacb.com/blog/sequencing-101-looking-beyond-the-single-reference-genome-to-a-pangenome-for-every-species/

Brockhurst M. A. et al. The ecology and evolution of pangenomes //Current Biology. – 2019. – Т. 29. – №. 20. – С. R1094-R1103.

Почему функциональная аннотация — нетривиальная задача

- Ошибки аннотаций легко переносятся дальше по цепочке «похожих» белков.
- Мультидоменные белки затрудняют прямой перенос функции по одному совпавшему домену.
- Паралоги могут сохранять общую архитектуру, но расходиться по функции.
- Нужна не одна улика, а несколько независимых линий доказательств.

Четыре стратегии функциональной аннотации

A) По сходству

последовательность
домены
мотивы

B) По ко-локализации

соседство
генов
и кластеры

C) Филогенетический паттерн

совместное
Присутствие/Отсутствие

D) Ко-регуляция

ко-экспрессия
и сети

A) Аннотация по сходству последовательностей

Что усиливает вывод

- высокое покрытие выравнивания, а не только высокая высокая степень сходства последовательностей
- сохранённые домены и ключевые мотивы/активные центры
- сходная длина и доменная архитектура белка
- экспериментально подтверждённые reference-аннотации

Типичные ловушки

- совпал только один домен в мультидоменном белке
- Лучшее совпадение оказалось паралогом, а не ортологом
- функция слишком детально перенесена без подтверждения
- геномное окружение гена не согласуется с данной аннотацией

B) Аннотация по ко-локализации (gene neighborhood)

- У прокариот гены одного пути или комплекса часто расположены рядом и образуют опероны/кластеры.
- Если неизвестный ген стабильно находится рядом с генами одного и того же процесса, это сильный аргумент в пользу функциональной связи.
- Особенно ценна не разовая близость, а консервация соседства в нескольких геномах.
- Анализируют расстояние между генами, ориентацию и устойчивость порядка в эволюции.

Что даёт этот подход

Позволяет распознавать модули:

ферменты пути
транспортёр
регулятор
белок сборки комплекса

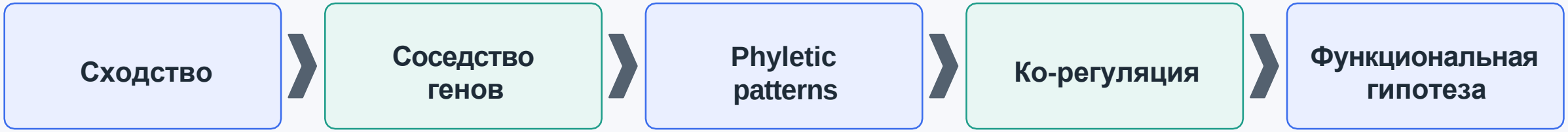
C) Филетические профили (phyletic patterns)

- Для каждого гена строят вектор присутствия/отсутствия по множеству геномов.
- Если профили двух генов совпадают, это может означать, что они функционально связаны и сохраняются/теряются совместно.
- Подход особенно полезен для белковых комплексов, транспортных систем и путей, где компоненты эволюционно «ходят пакетом».
- Корректная интерпретация требует учёта филогении: сходный профиль не всегда означает общую функцию.

D) Ко-регуляция и ко-экспрессия

- Гены одного процесса часто регулируются согласованно и реагируют на одни и те же условия.
- У бактерий это может быть общий регулон или даже оперон; у эукариот — ко-экспрессируемый модуль или регуляторная сеть.
- Источники данных: RNA-seq, транскриптомные панели, иногда ChIP-seq или данные о связывании регуляторов.
- Сильный сигнал ко-регуляции особенно полезен там, где сходство последовательности уже не даёт однозначного вывода.

Лучший результат даёт интеграция признаков



Одна улика даёт предположение. Несколько независимых улик дают убедительную аннотацию.

- На практике результаты часто сводят в функциональные сети и «confidence score».
- Именно эту интеграционную логику наглядно показывает STRING.

Инструменты сравнительной геномики: карта темы

COG	KOG	STRING	SEED/RAST
ортогруппы для прокариот	эукариотический аналог COG	сети функциональных ассоциаций	подсистемный подход к аннотации

Плюс: NCBI Orthologs для межвидового поиска ортологов и сопоставления генов.

COGs и KOGs: филогенетическая классификация белков

COG

- Clusters of Orthologous Groups of proteins
- исторически и практически особенно важен для прокариот
- даёт ортологическую группу и функциональную категорию
- удобен для функционального профиля целого генома

KOG

- Eukaryotic Orthologous Groups
- аналогичный подход для эукариот
- полезен для грубой функциональной классификации протеома
- требует большей осторожности из-за расширенных семейств и паралогов

NCBI Orthologs и HomoloGene: что важно знать сегодня

- HomoloGene долго был учебным и практическим ресурсом для межвидового сравнения генов.
- Сейчас NCBI перенаправляет пользователей к современным страницам Datasets/Gene и NCBI Orthologs.
- Legacy-данные HomoloGene сохраняют историческую ценность, но в актуальной работе удобнее ориентироваться на NCBI Orthologs.
- Методический вывод: важна не столько конкретная кнопка на сайте, сколько логика поиска ортологов между видами.

Для чего использовать

поиск межвидовых соответствий

сравнение гена в нескольких таксонах

перенос информации о функции и семействах

STRING: функциональные ассоциации белков

- STRING показывает не только прямые белок-белковые контакты, а более широкий контекст функциональной связи.
- Сеть интегрирует несколько каналов доказательств: эксперименты, курируемые базы данных, ко-экспрессию, контекст генома, литературу и вычислительные предсказания.
- Пользователь получает как саму сеть, так и оценку уверенности (confidence score) и функциональное обогащение (enrichment).
- Это особенно удобно для аннотации неизвестного белка по окружению известных белков.

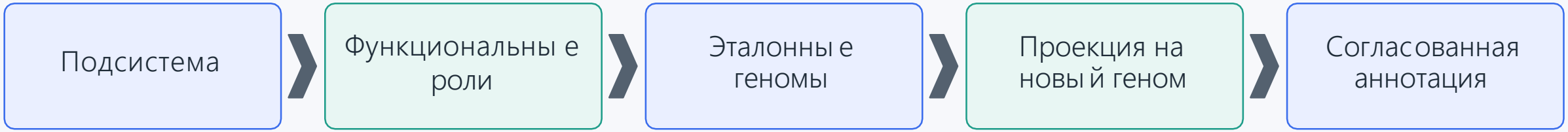
Как читать STRING

важно смотреть не только на граф, но и на источники связей

какой канал дал основной вклад?

есть ли enrichment по пути или процессу?

SEED/RAST: подсистемный подход к аннотации



Мини-кейс: как аннотировать hypothetical protein

Дано

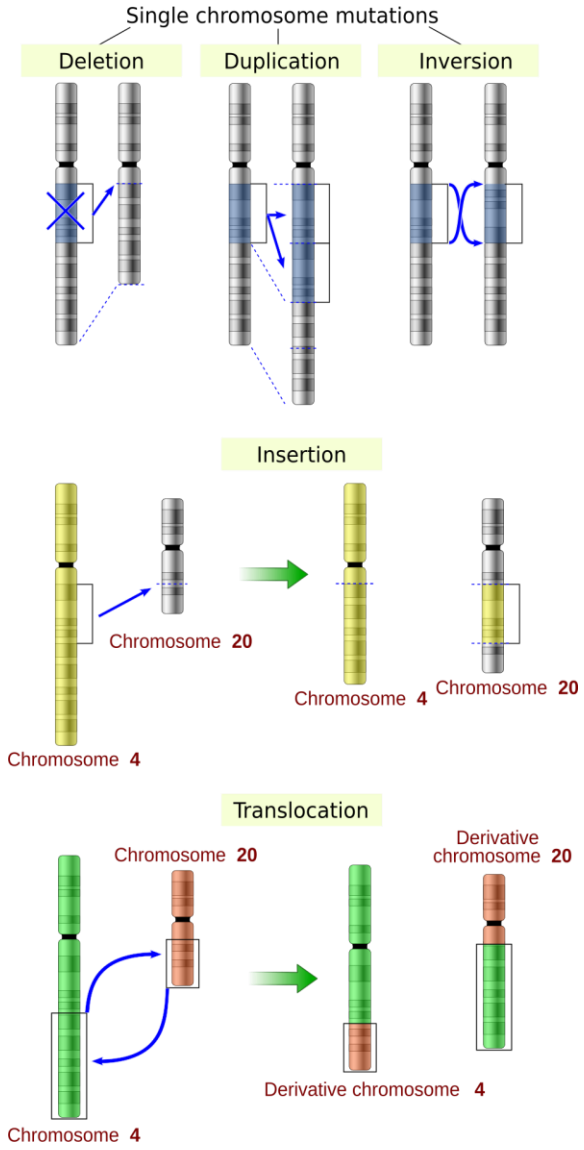
- неизвестный белок в кластере из 6 генов
- по соседству — ферменты одного пути и регулятор
- похожий профиль присутствия у нескольких компонентов системы
- в S TR ING белок попадает в сеть того же процесса

Логика вывода

- сходство даёт семейство/домен
- ко-локализация указывает на участие в том же модуле
- Филетические паттерны усиливают функциональную связь
- ко-регуляция/S TR ING повышают уверенность
- итог: формулируем гипотезу + уровень уверенности + план проверки

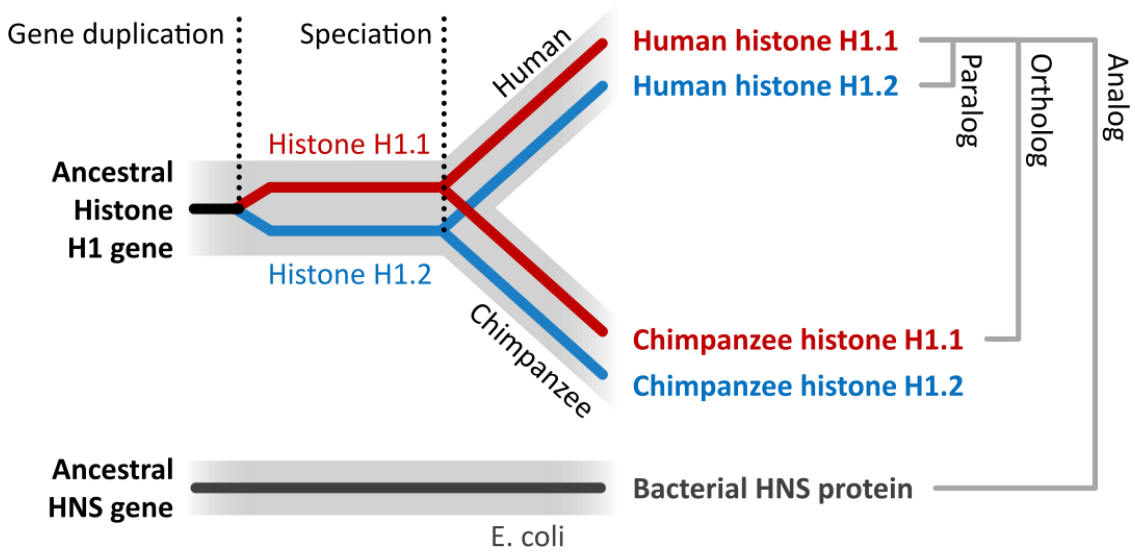
- Сравнительная геномика сравнивает не только последовательности, но и состав, организацию и эволюционный контекст геномов.
- Надёжная функциональная аннотация строится на сумме независимых признаков: сходство, соседство, phyletic patterns, ко-регуляция.
- COG/KOG помогают классифицировать белки по ортологическим группам и функциональным категориям.
- NCBI Orthologs, STRING и SEED/RAST позволяют перейти от отдельных генов к связям, подсистемам и функциональным гипотезам.

Пути эволюции геномов: от точек до хромосом



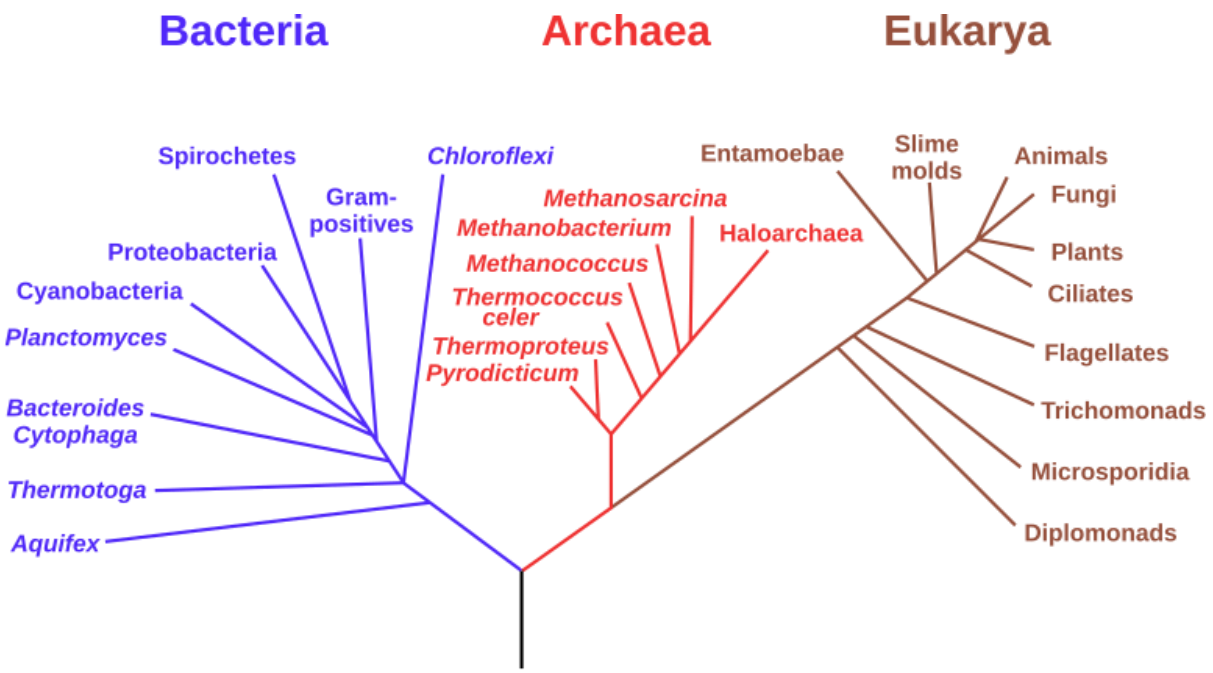
- Точечные замены и небольшие вставки/делеции
- Крупные структурные варианты: инверсии, транслокации
- Дупликации фрагментов и генов → сырьё для новизны
- Стабильные блоки генома и «горячие точки» перестроек

Гомология: кто кому «родственник» в генах



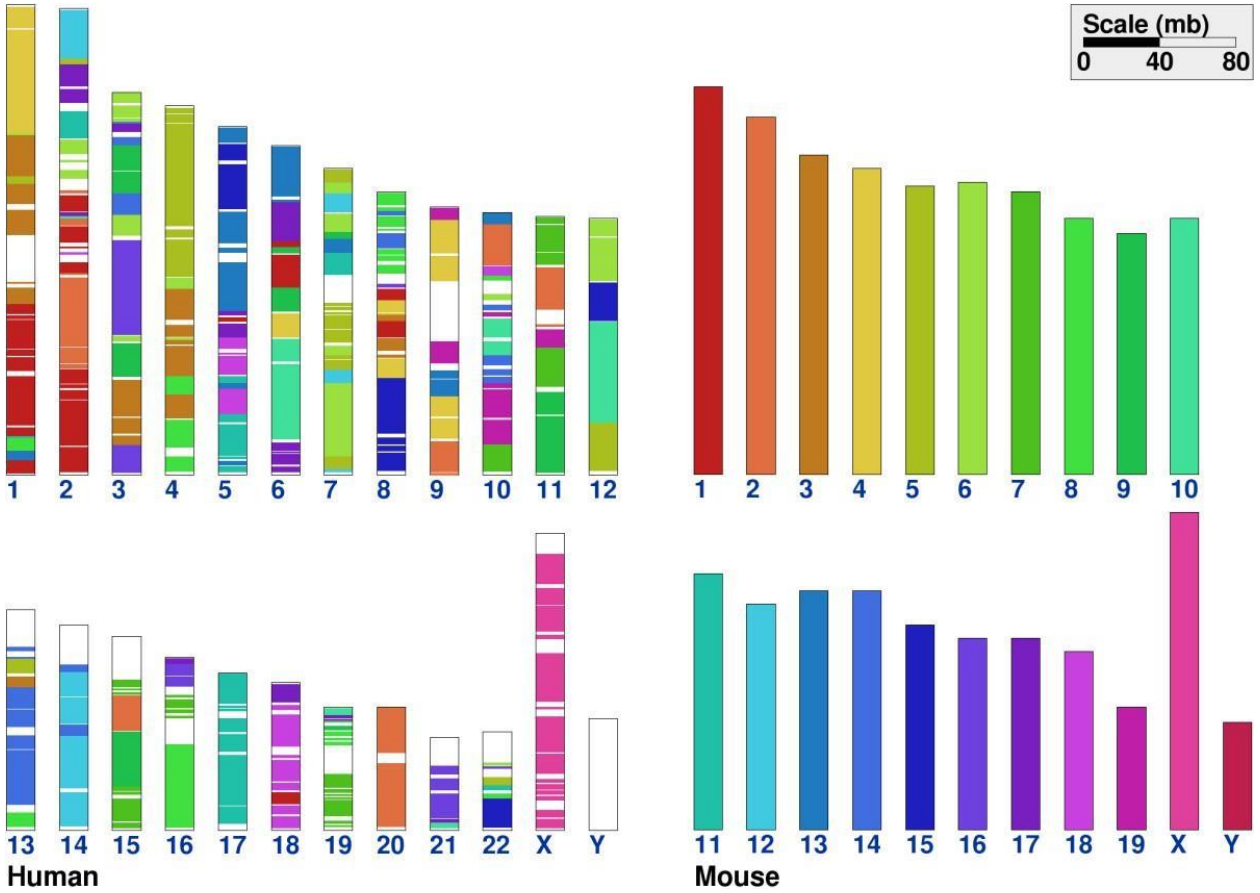
- Ортологи (orthologs, ортологи): разделение по видам
- Паралоги (paralogs, паралоги): дупликация внутри линии
- Аналоги (analogs, аналоги): похожая функция без общего происхождения
- Ошибки: путать ортологию и паралогию при переносе функций

Филогенетические деревья: что они показывают



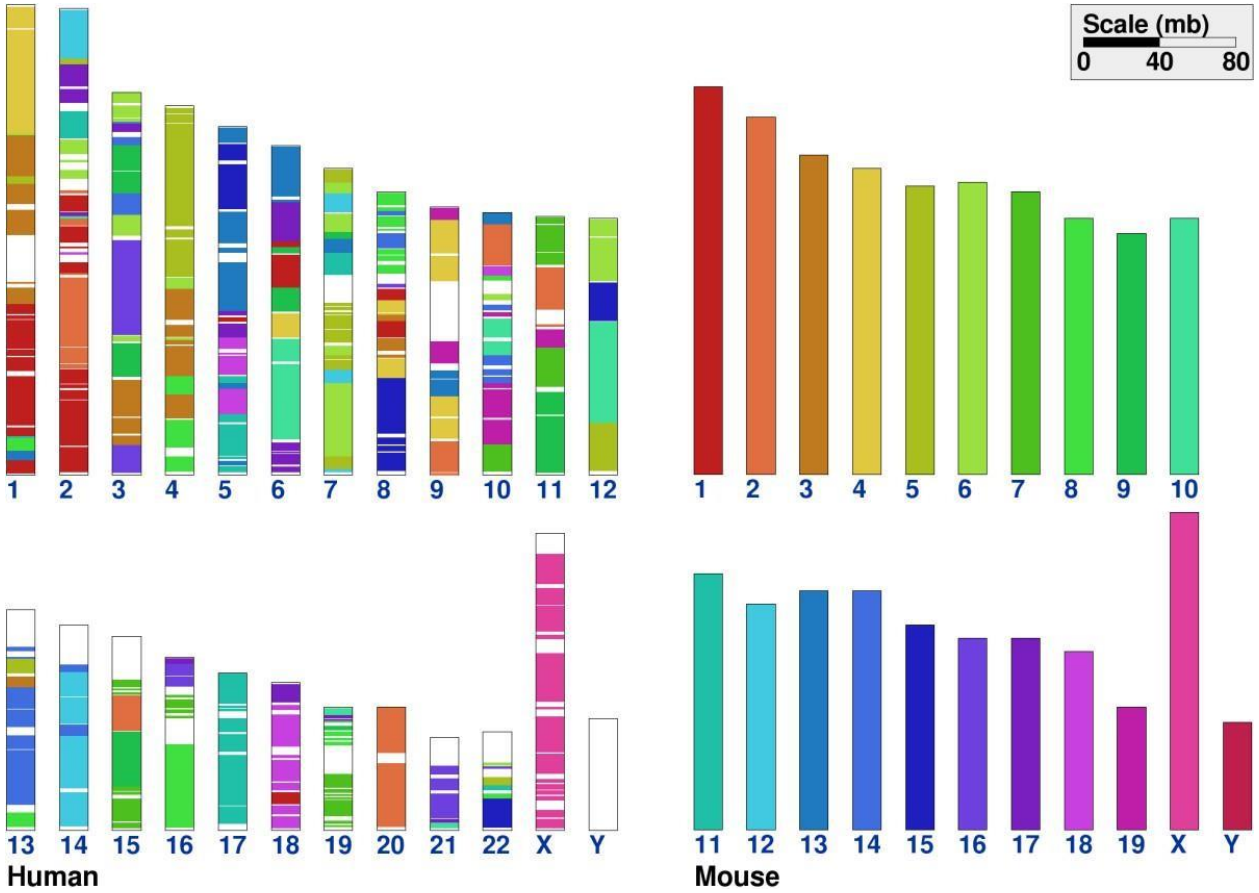
- Вершины: виды/штаммы/гены; рёбра: «ветви» родства
- Длина ветви часто отражает количество изменений
- Дерево зависит от модели и типа данных (ДНК, белки, перестройки)
- Горизонтальный перенос генов (HGT) усложняет «деревья»

Синтения: порядок генов как историческая «память»



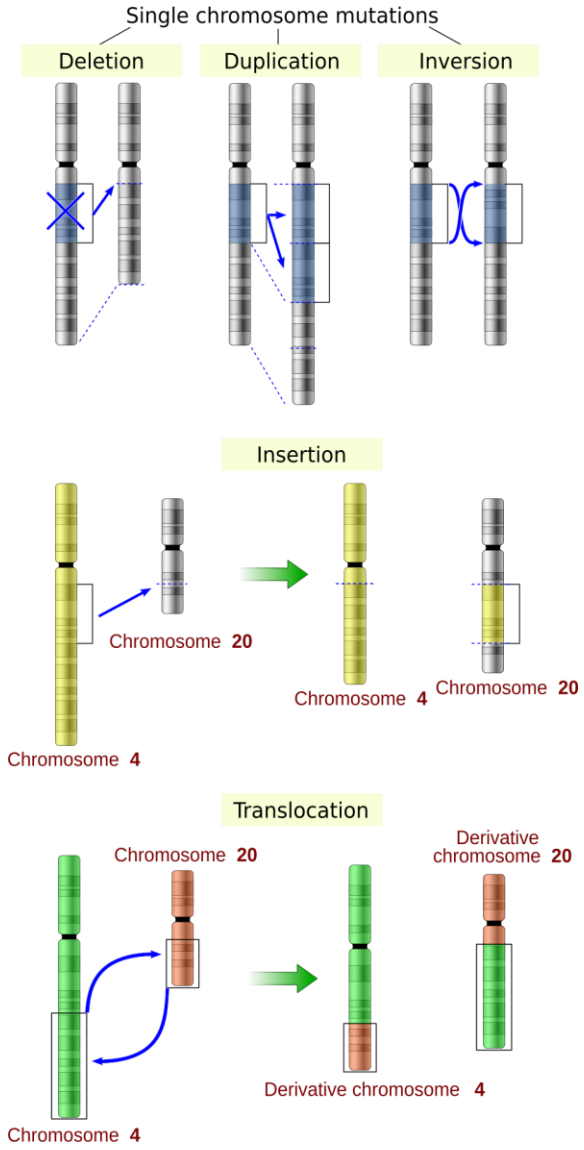
- Синтения (synteny, синтения): сохранённые блоки генов
- Перестройки «перемешивают» блоки, но часть сохраняется
- По синтении ищут ортологи и следы древних дупликаций
- Синтения помогает изучать стабильность участков генома

Стабильность участков генома и комплексы генов



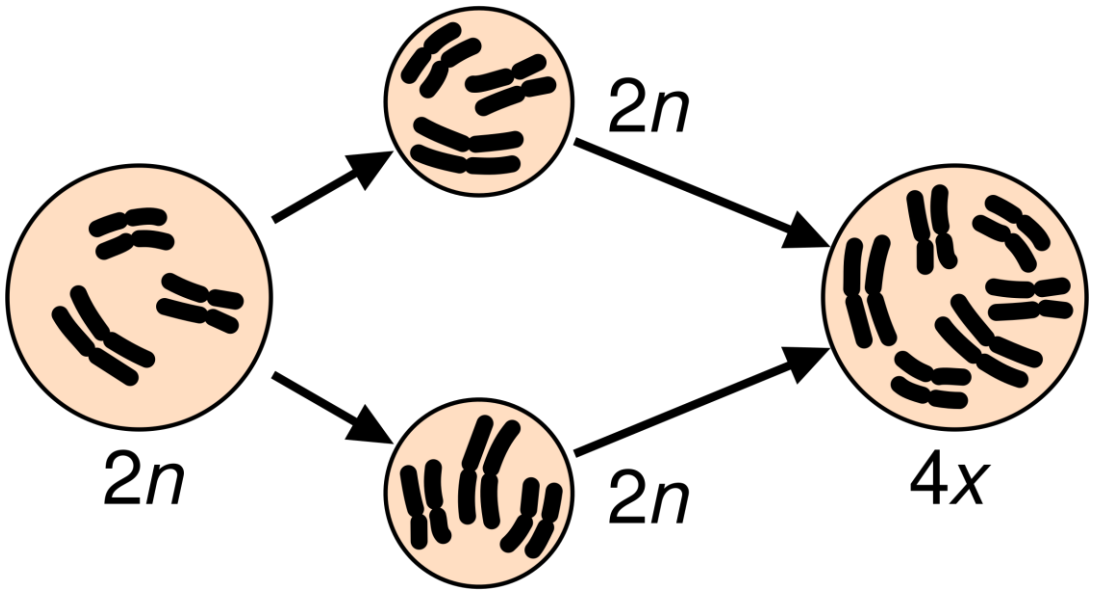
- Комплексы генов: кластеры, опероны, локусы развития
- Консервативные «островки» часто связаны с важными функциями
- Перестройки могут разрывать ко-регуляцию и менять фенотип
- Эволюционный анализ помогает отличить «шум» от отбора

Сортировка перестановками: sorting by reversals



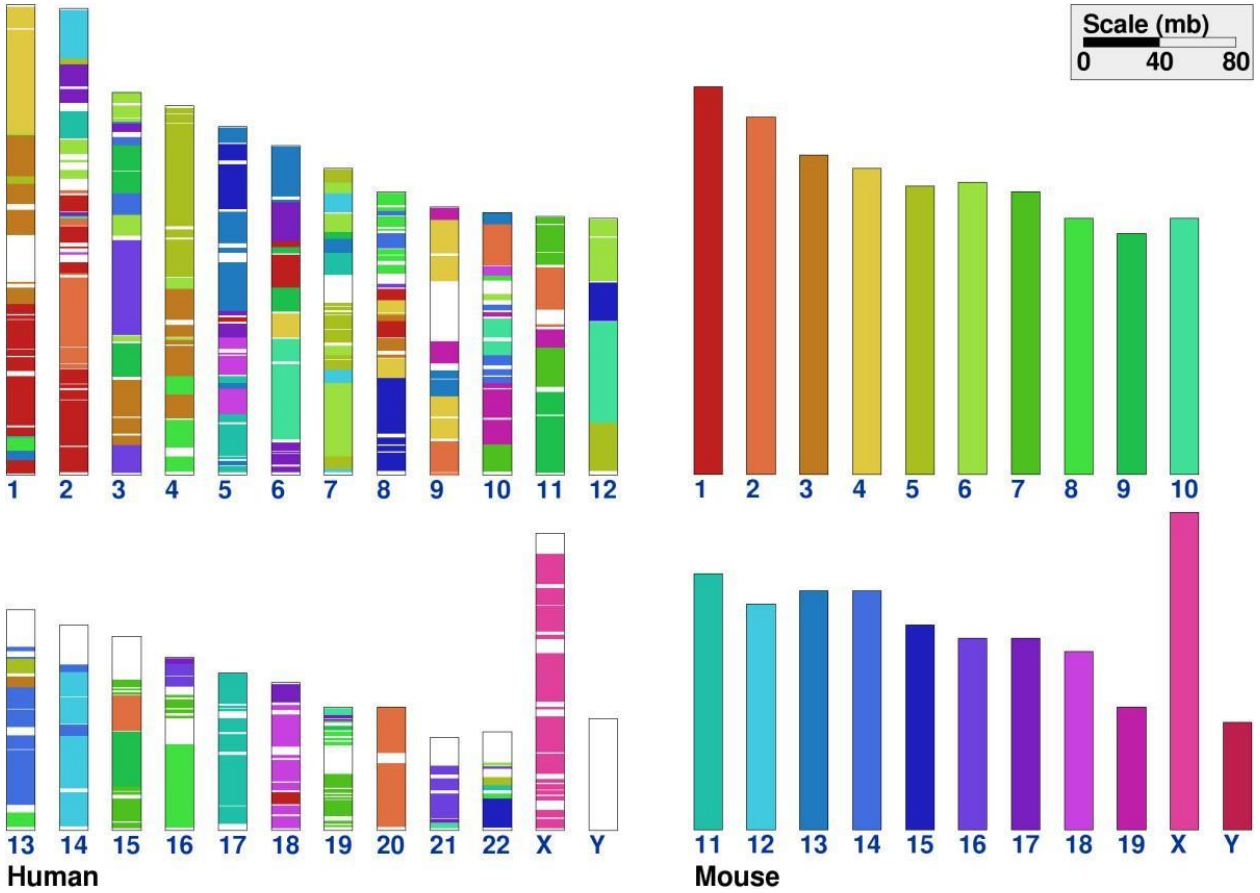
- Reversal (reversal, разворот/инверсия): переворот участка
- Задача: минимальным числом разворотов привести геном A к B
- Результат: «расстояние» между геномами по архитектуре
- Применение: сравнение хромосом и реконструкция предков

Полногеномные дупликации: WGD



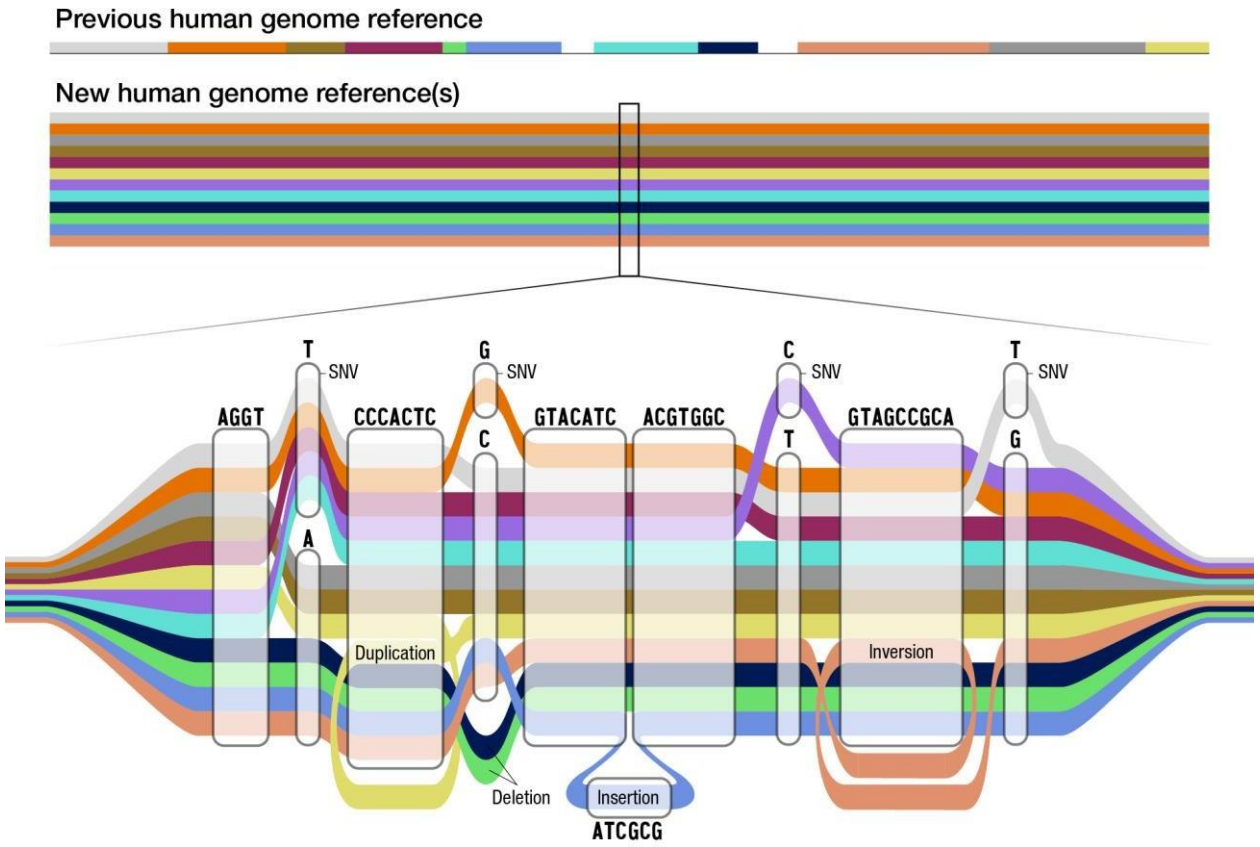
- WGD (whole-genome duplication, полногеномная дупликация)
- Внезапное удвоение набора генов (часто у растений/рыб)
- Дальше: потеря части копий + расхождение функций
- Эффект: новые пути регуляции, «сырьё» для инноваций

Как находят следы WGD в современных геномах



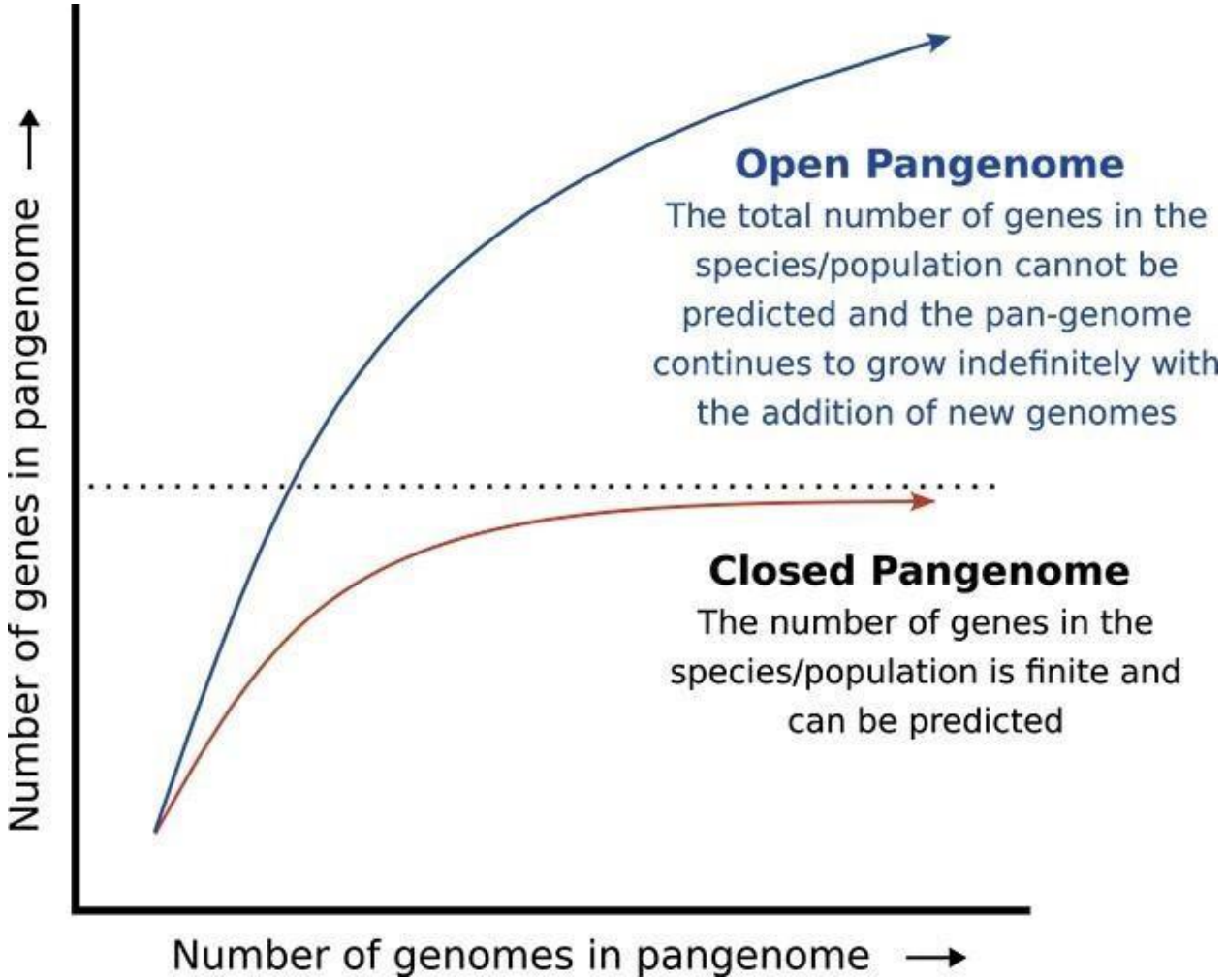
- Дублированные синтенные блоки («параллельные» участки)
- Сходные наборы генов в разных хромосомах/локусах
- Следы «дробления» после удвоения (fractionation, фракционирование)
- Сопоставление с филогенетикой: когда могло случиться WGD

Пан-геном: «все гены вида», а не один референс



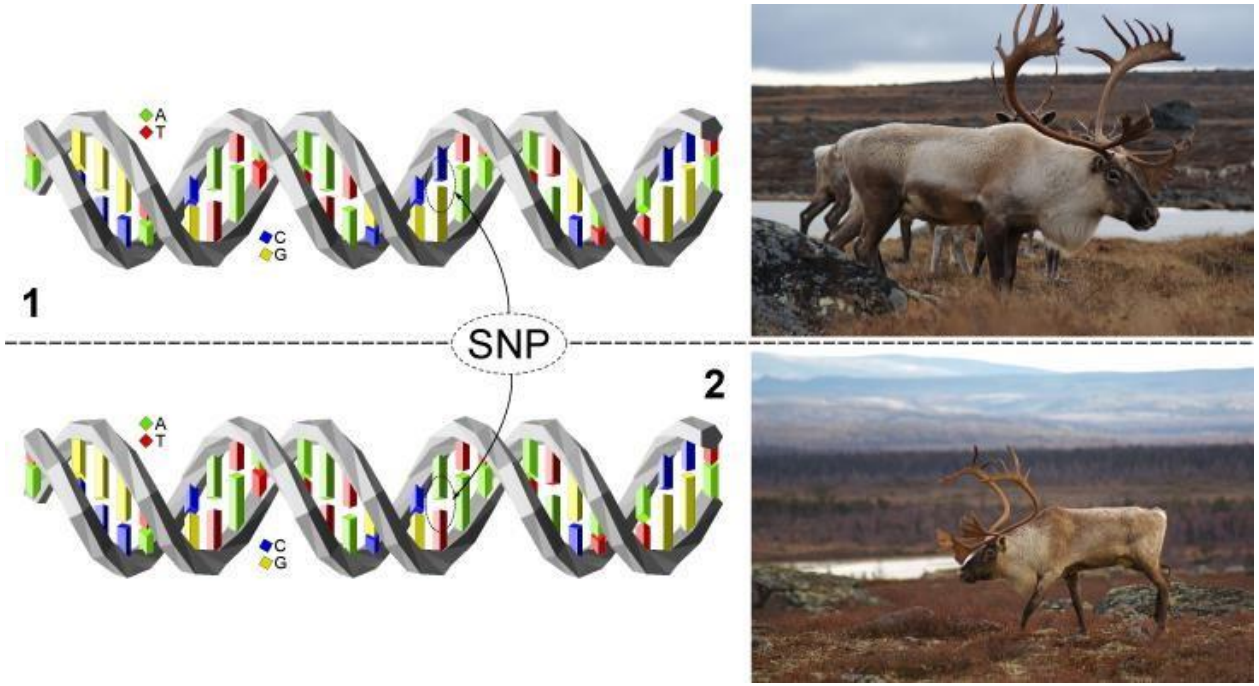
- Пан-геном (pangenome, пан-геном) = core + accessory
- Core genes: есть у всех; accessory genes: у части
- Полезно для бактерий, вирусов и популяций человека
- Снижает «референсный перекос» (reference bias, смещение референса)

Открытый и закрытый пан-геном



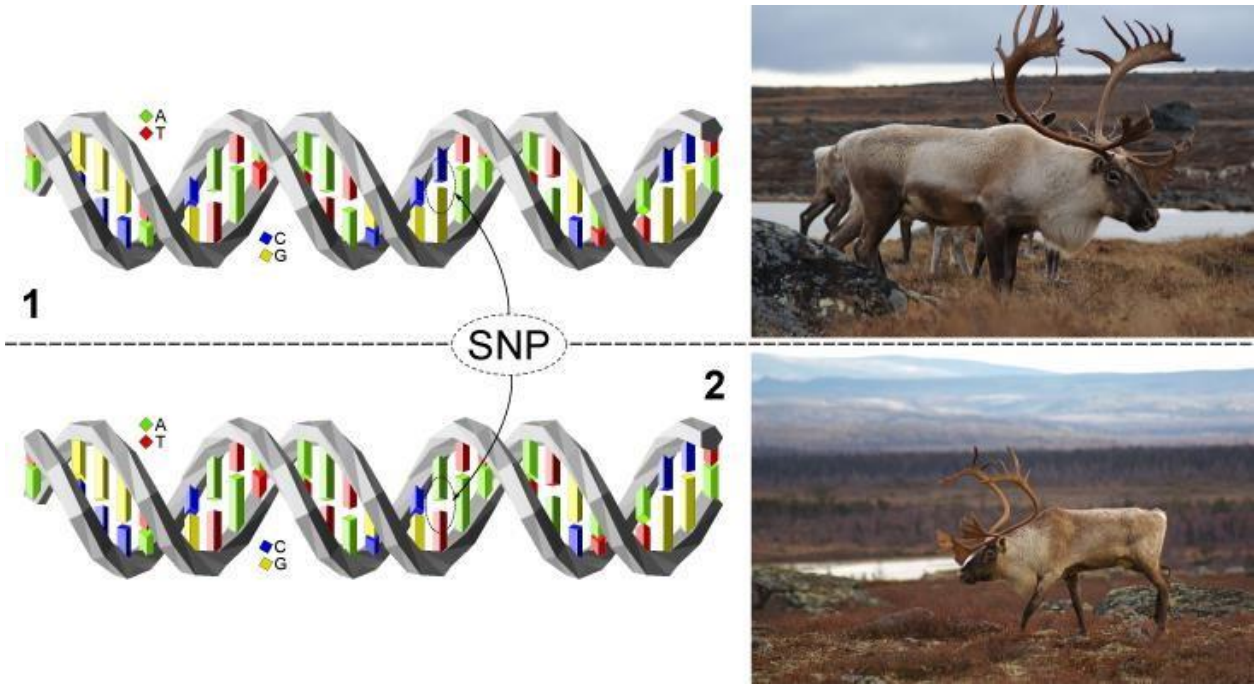
- Открытый (open) пан-геном: новые геномы добавляют новые гены
- Закрытый (closed) пан-геном: кривая выходит на плато
- Экология и обмен ДНК влияют на «открытость»
- Практика: сколько геномов нужно, чтобы «поймать» разнообразие

SNP: основной «атом» вариабельности



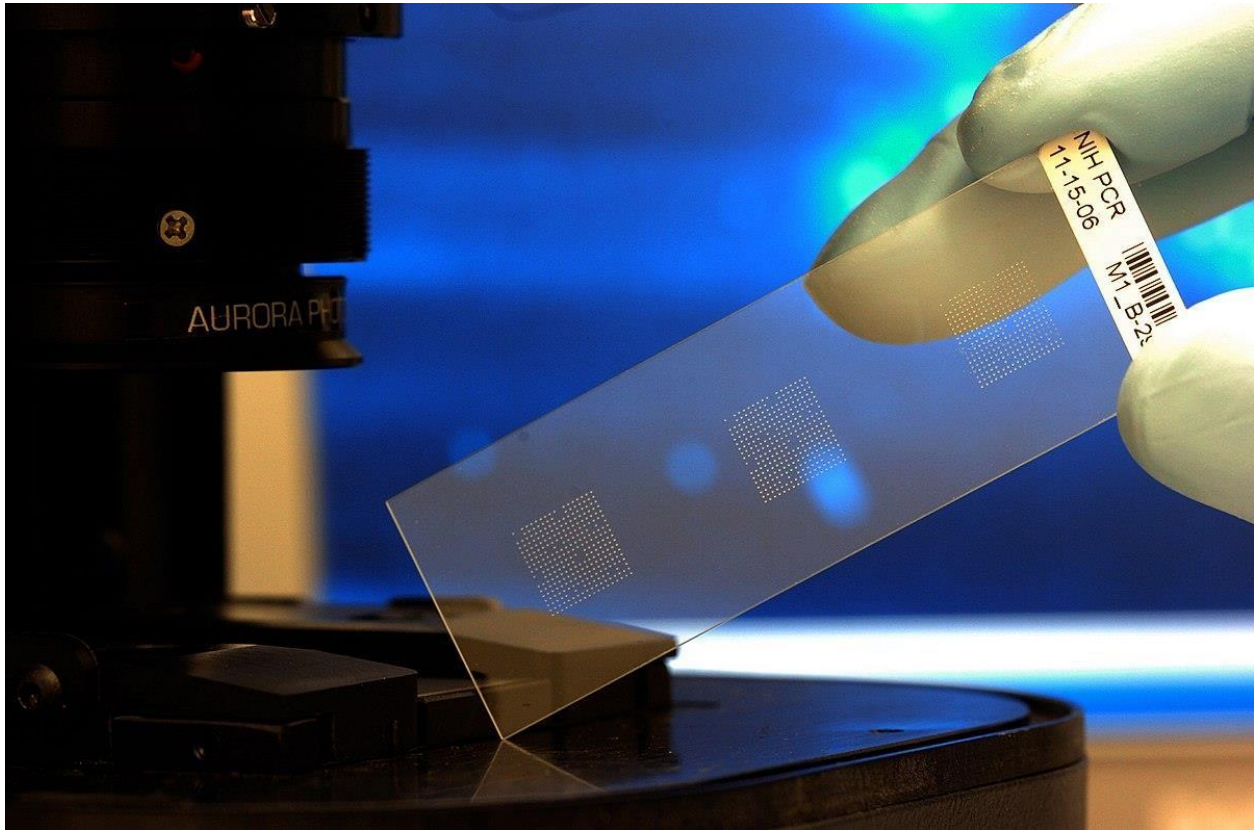
- SNP (single nucleotide polymorphism, однонуклеотидный полиморфизм)
- Может быть нейтральным или влиять на белок/регуляцию
- Используют для картирования признаков и популяционных анализов
- Важно: различать SNP (частый) и SNV (variant, вариант) в целом

Эволюция наследственной патологии: почему «вредные» варианты остаются



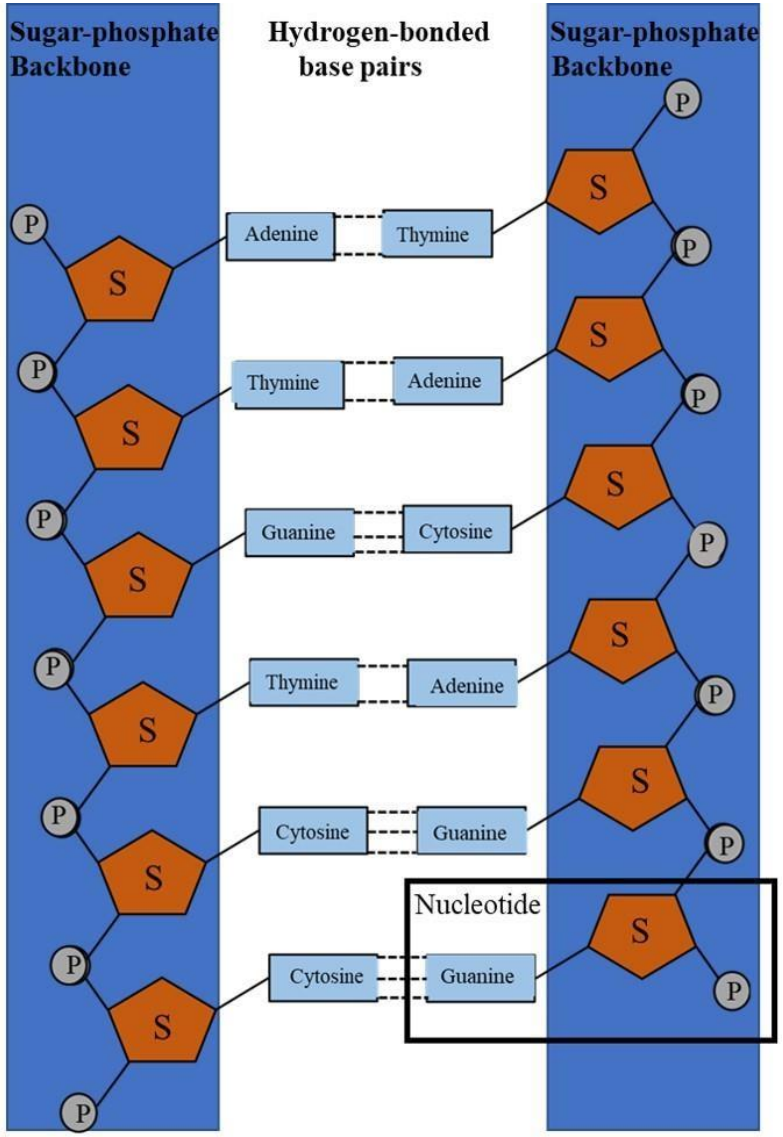
- Баланс: мутации появляются постоянно, отбор удаляет не все
- Дрейф: случайность особенно сильна в малых популяциях
- Гетерозиготное преимущество (heterozygote advantage, преимущество гетерозигот)
- Среда меняется: «нейтральное» сегодня может стать вредным завтра

SNP в молекулярной диагностике человека



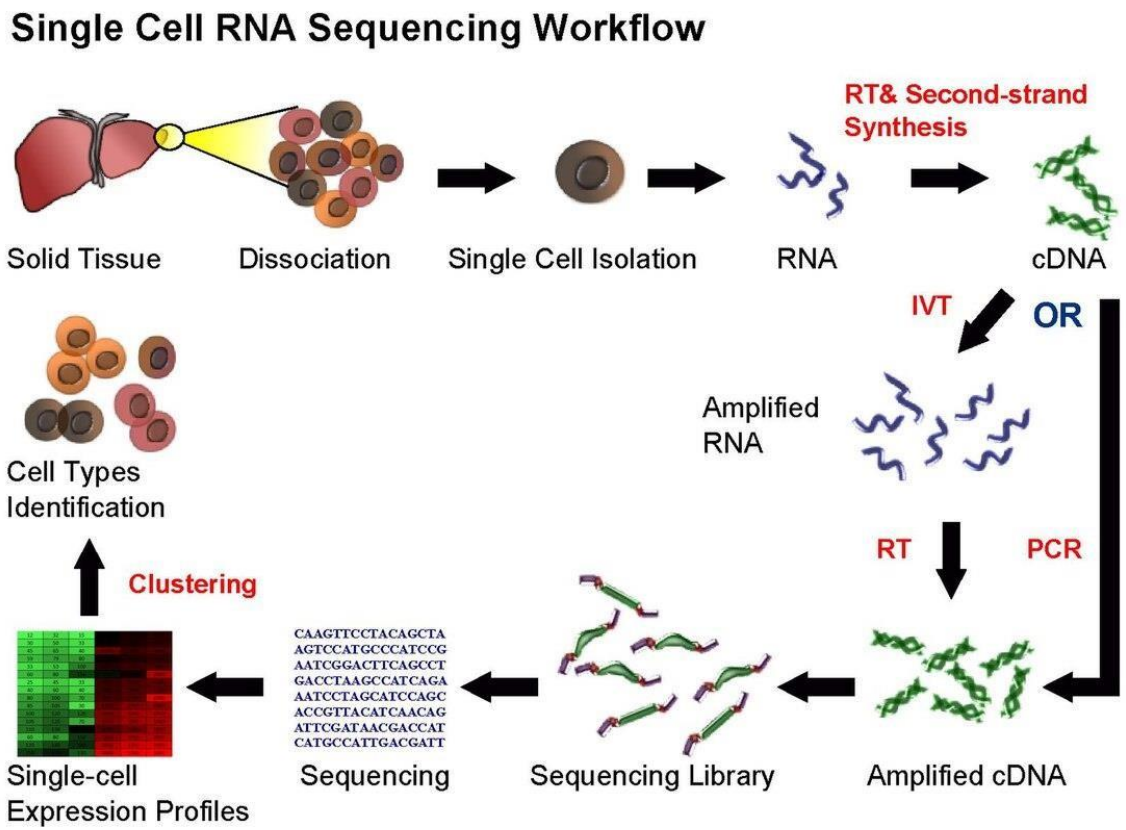
- Маркерные панели SNP для наследственных болезней и рисков
- Фармакогенетика: подбор препарата/дозы по генотипу
- Контроль качества: популяционные частоты, ложноположительные
- Этика: информированное согласие и интерпретация результатов

Спейсеры рРНК и маркерные гены



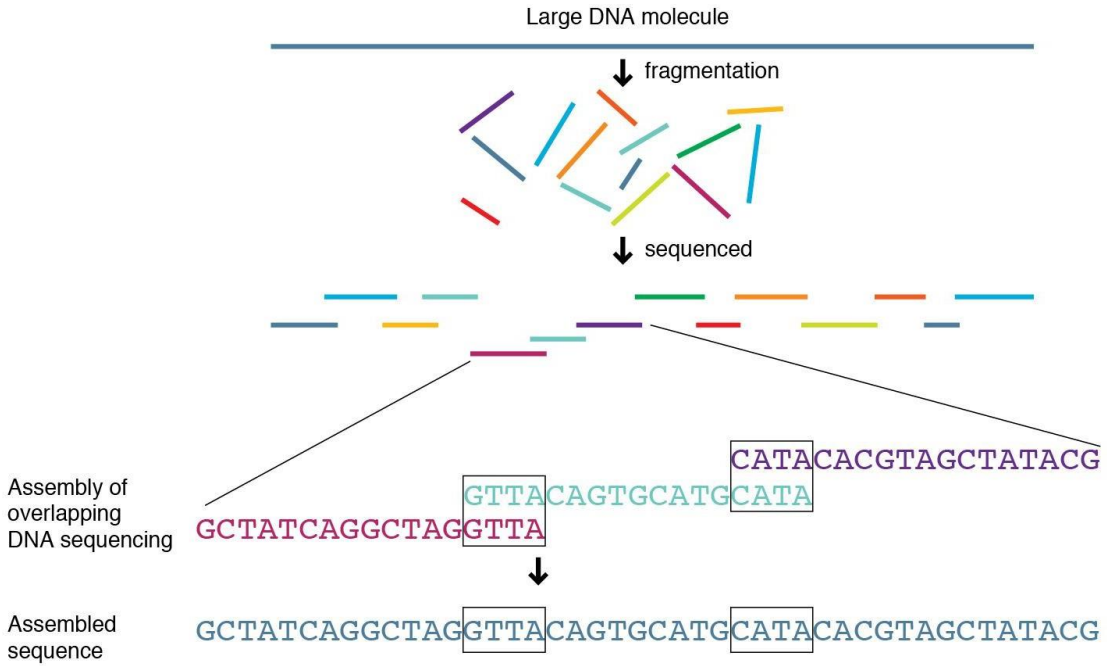
- rRNA (ribosomal RNA, рибосомная РНК) — консервативные гены
- ITS (internal transcribed spacer, внутренний транскрибируемый спейсер)
- 16S rRNA (16S ribosomal RNA, 16S рибосомная РНК) — маркер бактерий
- Выбор маркера зависит от задачи и таксона

16S рРНК секвенирование: быстрый портрет микробиома



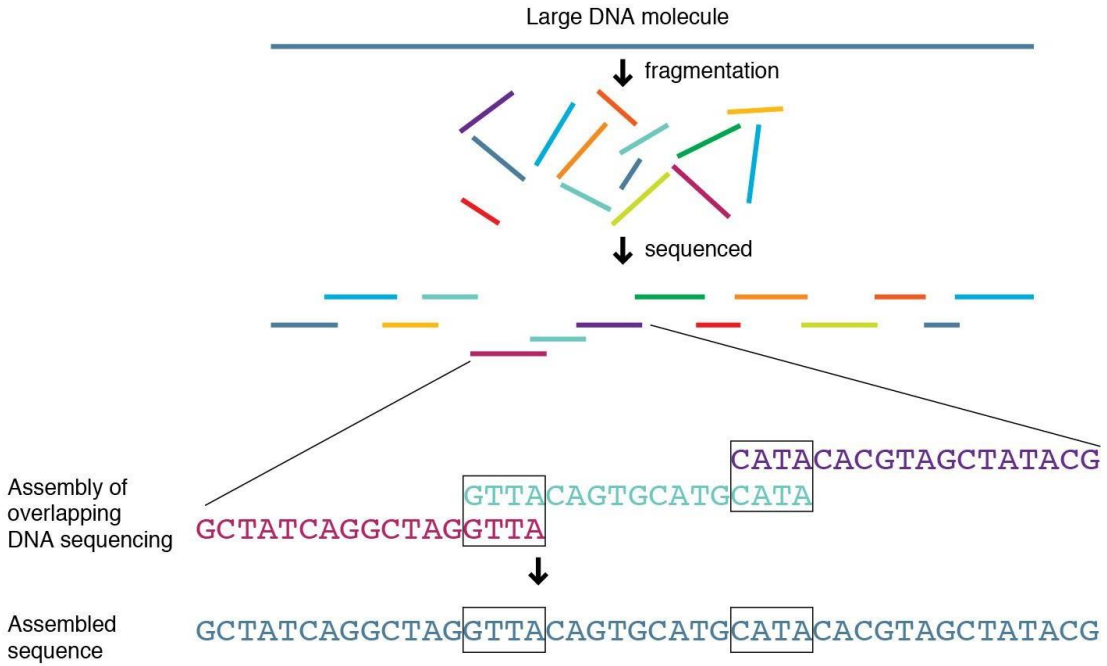
- 16S rRNA (16S ribosomal RNA, 16S рибосомная РНК) — «штрих-код» бактерий
- Ампликоны (amplicons, ампликоны) → таксономия и относительные доли
- Ограничения: разрешение до вида не всегда возможно
- Сильные смещения: ПЦР (PCR, polymerase chain reaction; полимеразная цепная реакция)

Метагеномика: «геномика окружающей среды»



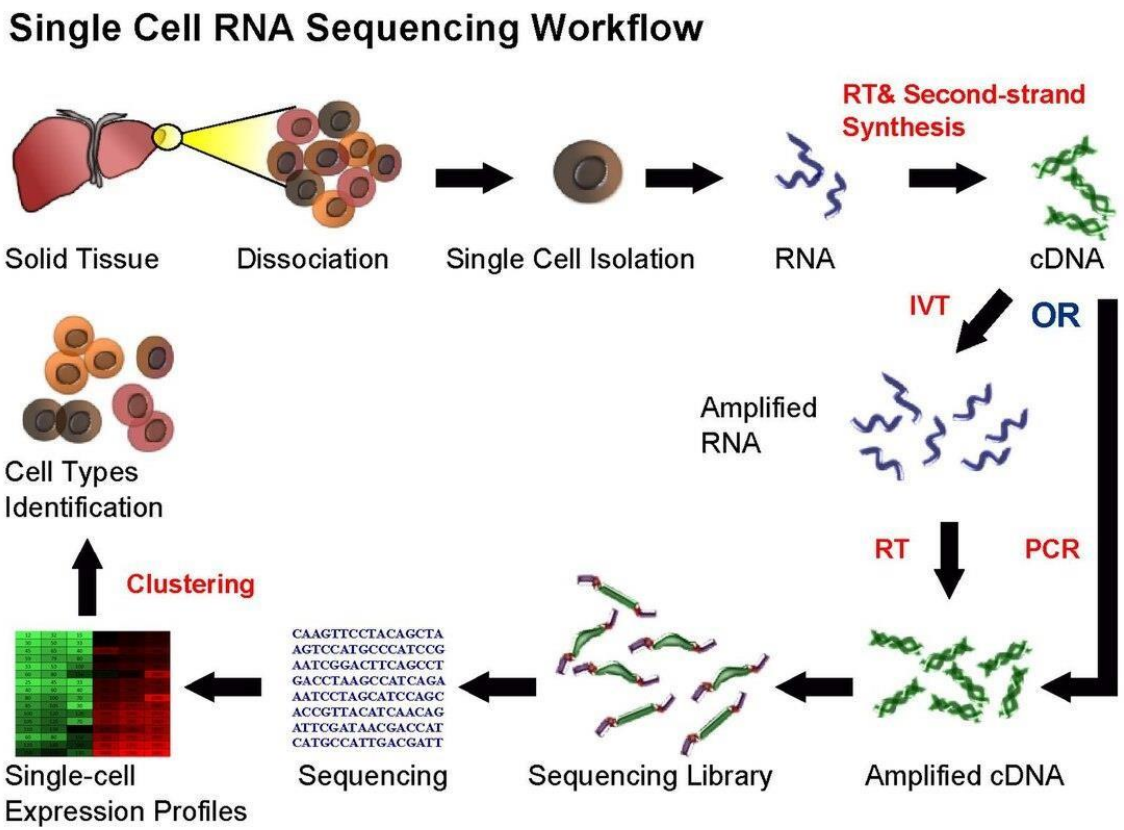
- Метагеном (metagenome, метагеном): ДНК сообщества
- Метагеномика (metagenomics, метагеномика): анализ этой смеси
- Два пути: ампликоны маркеров и тотальное секвенирование
- Цели: «кто там?» и «что они умеют делать?»

Shotgun-метагеномика: сборка и функциональный слой



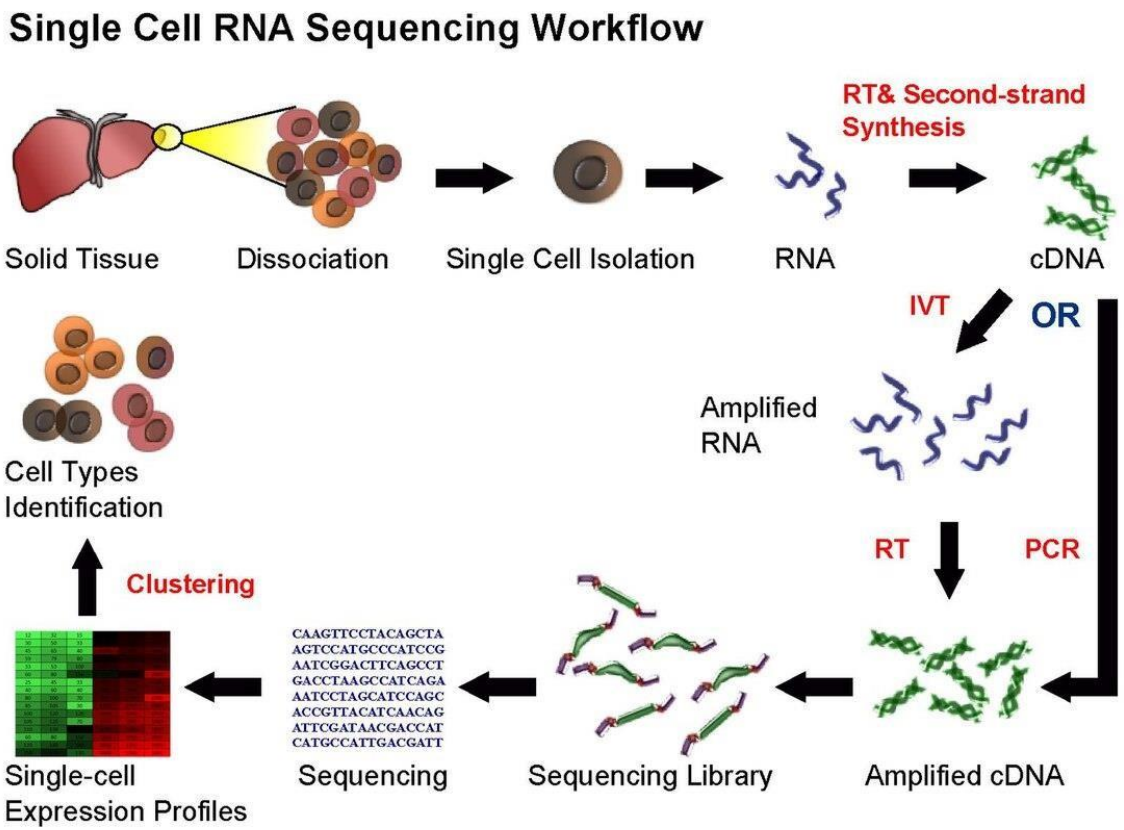
- Shotgun (shotgun, «дробовик»): секвенируем всю ДНК из образца
- Сборка (assembly, сборка) → контиги → биннинг → MAG (metagenome-assembled genome, геном из метагенома)
- Функции: ферменты, пути, резистомы, факторы вирулентности
- Проблемы: химерные сборки, контаминации, неполные геномы

Функциональная интерпретация: где чаще всего ошибаются



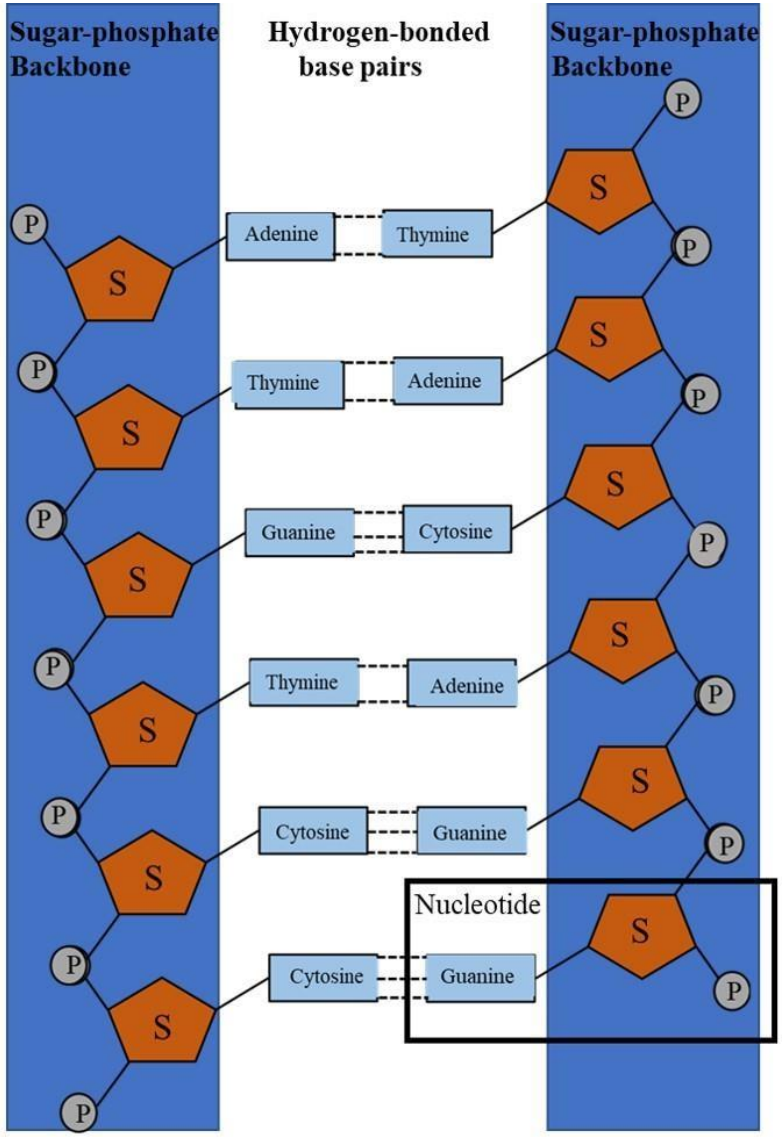
- «Есть ген» ≠ «есть функция» (экспрессия и контекст важны)
- Аналогии vs гомология: перенос функции требует осторожности
- Нормализация: глубина секвенирования и состав сообщества
- Лучше говорить языком вероятностей и проверяемых гипотез

NGS: от образца к данным (общая логика)



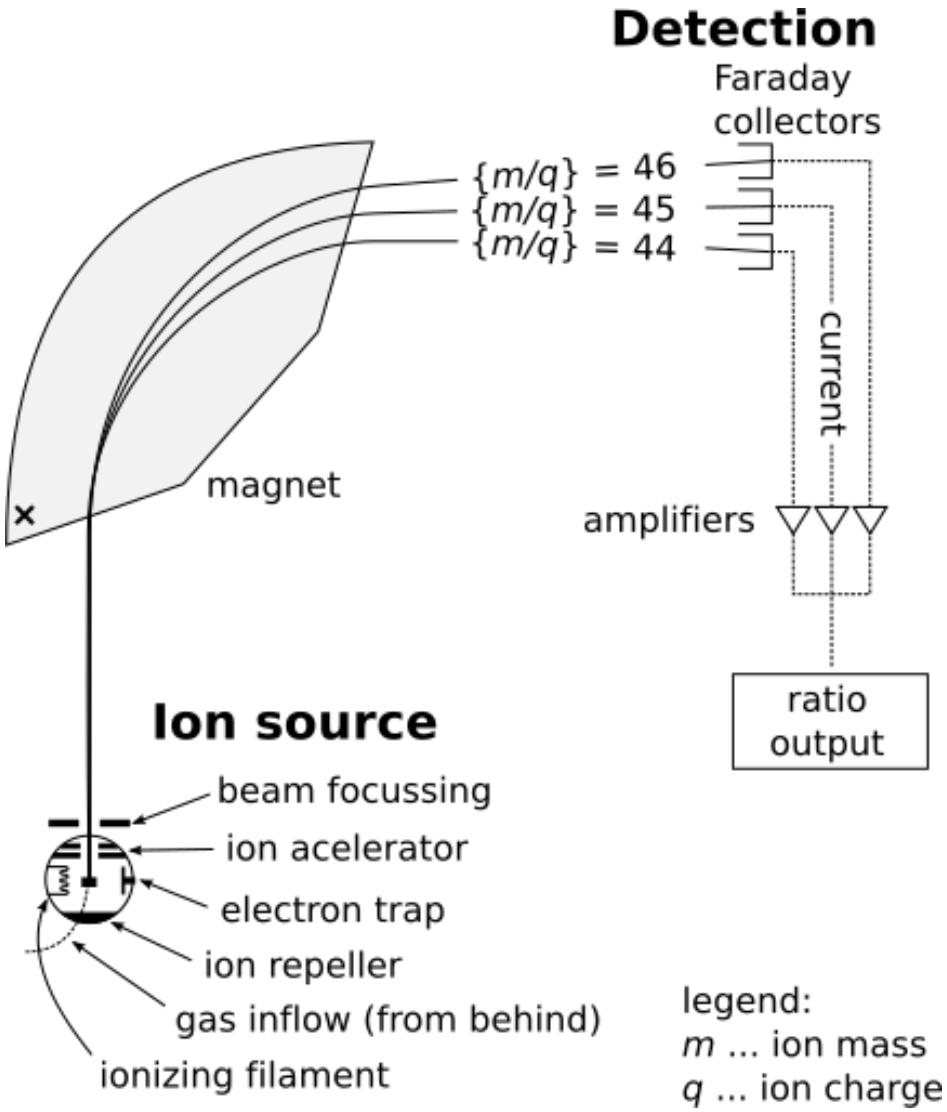
- NGS (next-generation sequencing, секвенирование нового поколения)
- Подготовка библиотеки → чтения (reads, риды) → анализ
- Контроль качества: адаптеры, ошибки, контаминации
- Дальше: выравнивание, сборка, аннотация и статистика

От генома к взаимодействиям: почему этого недостаточно «по последовательности»



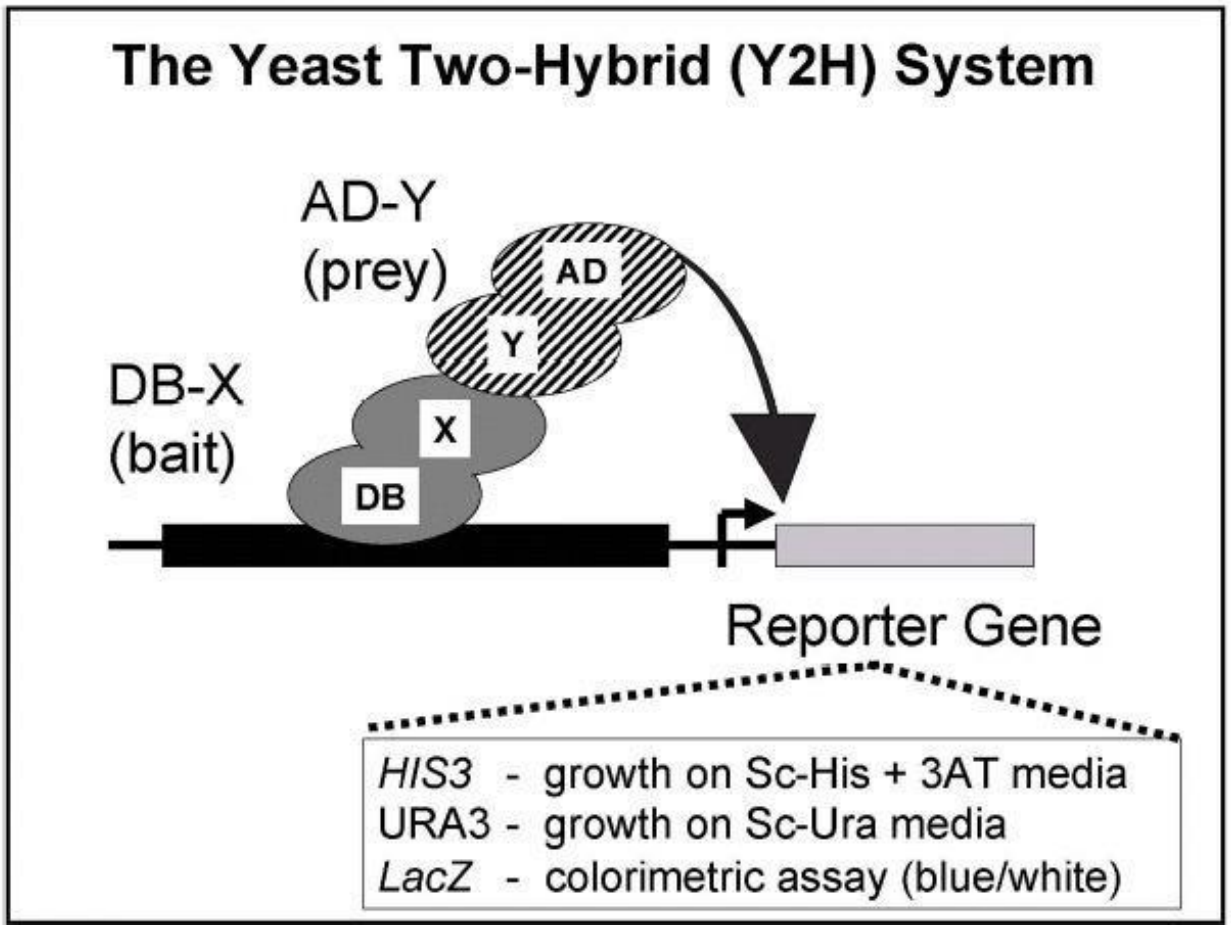
- Функция — это не только «какие гены есть», но и как они работают
- Регуляция: белки связываются с ДНК и РНК
- Комплексы: белки взаимодействуют друг с другом (PPI)
- Сетевой взгляд помогает объяснять фенотип и патологии

Белок–белковые взаимодействия (PPI): как их ищут



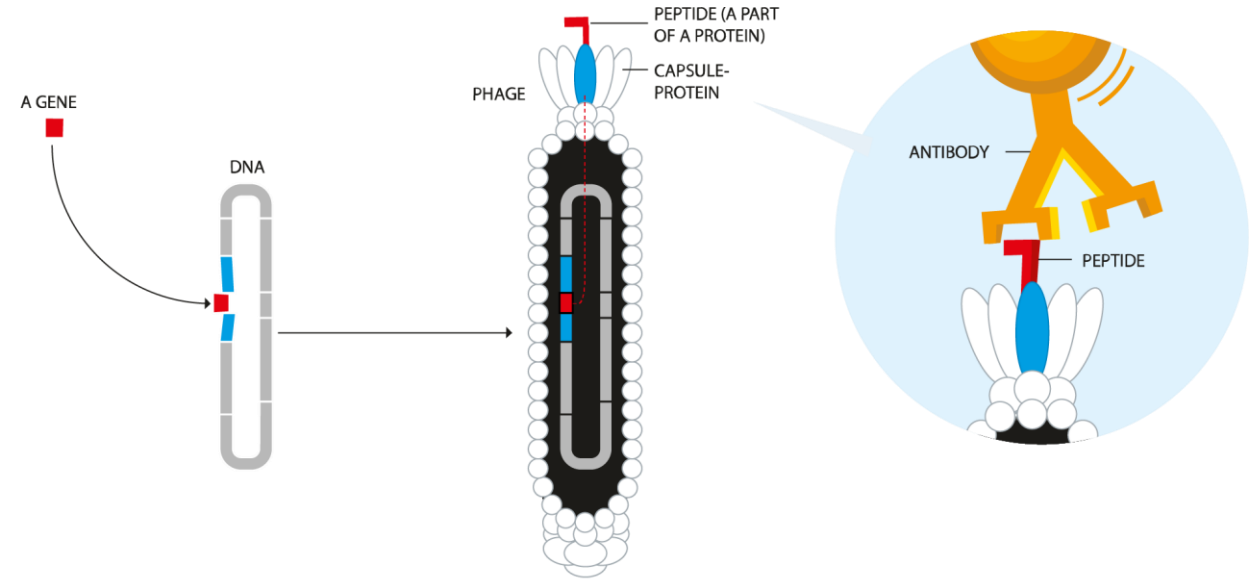
- PPI (protein–protein interactions, белок–белковые взаимодействия)
- Скрининг: Y2H, фаговый дисплей, белковые чипы
- Комплексы: co-IP (co-immunoprecipitation, ко-иммунопреципитация)
- Подтверждение: AP-MS (affinity purification–mass spectrometry, аффинная очистка–масс-спектрометрия)

Дрожжевая двугибридная система: Y2H



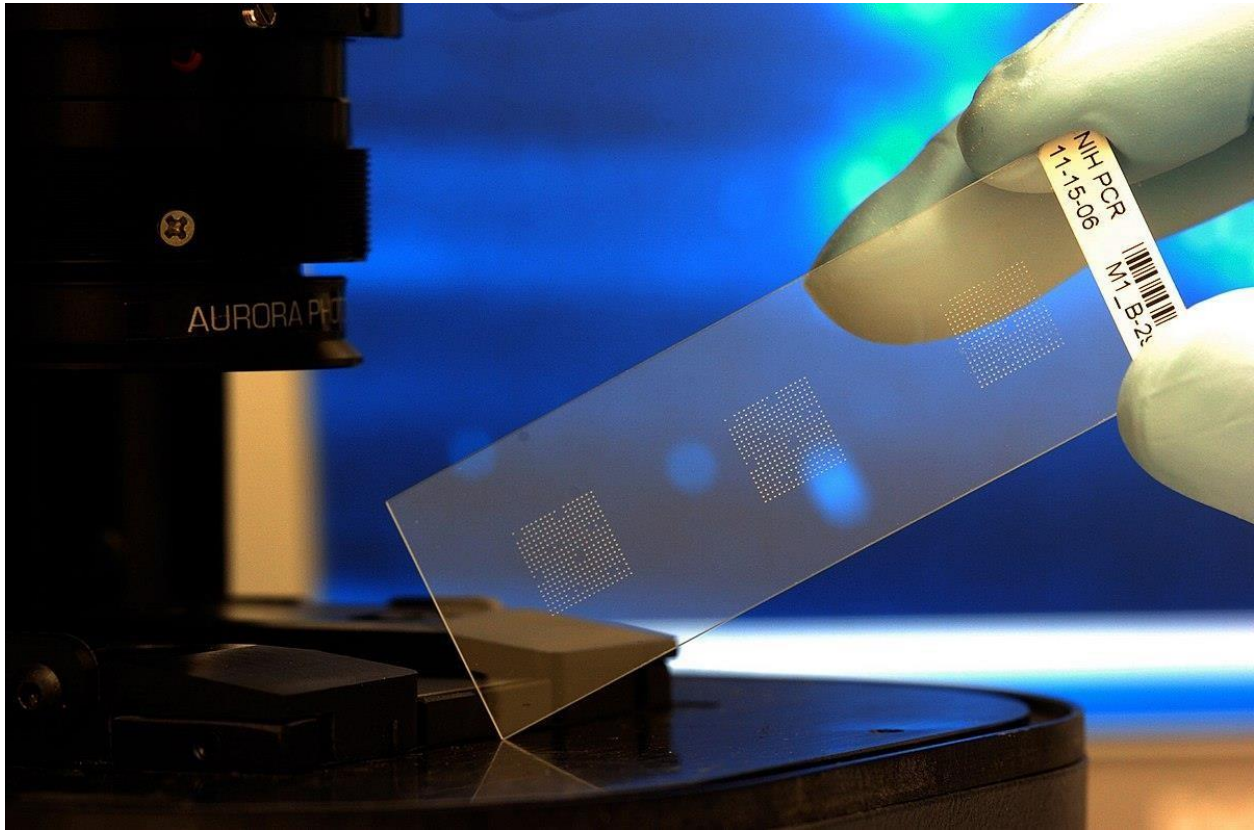
- Y2H (yeast two-hybrid, дрожжевая двугибридная система)
- «Наживка» + «добыча» → восстановление фактора транскрипции
- Сигнал: репортёрный ген (рост/окраска)
- Плюсы/минусы: дёшево и массово, но много ложных срабатываний

Фаговый дисплей: отбор лигандов и пептидов



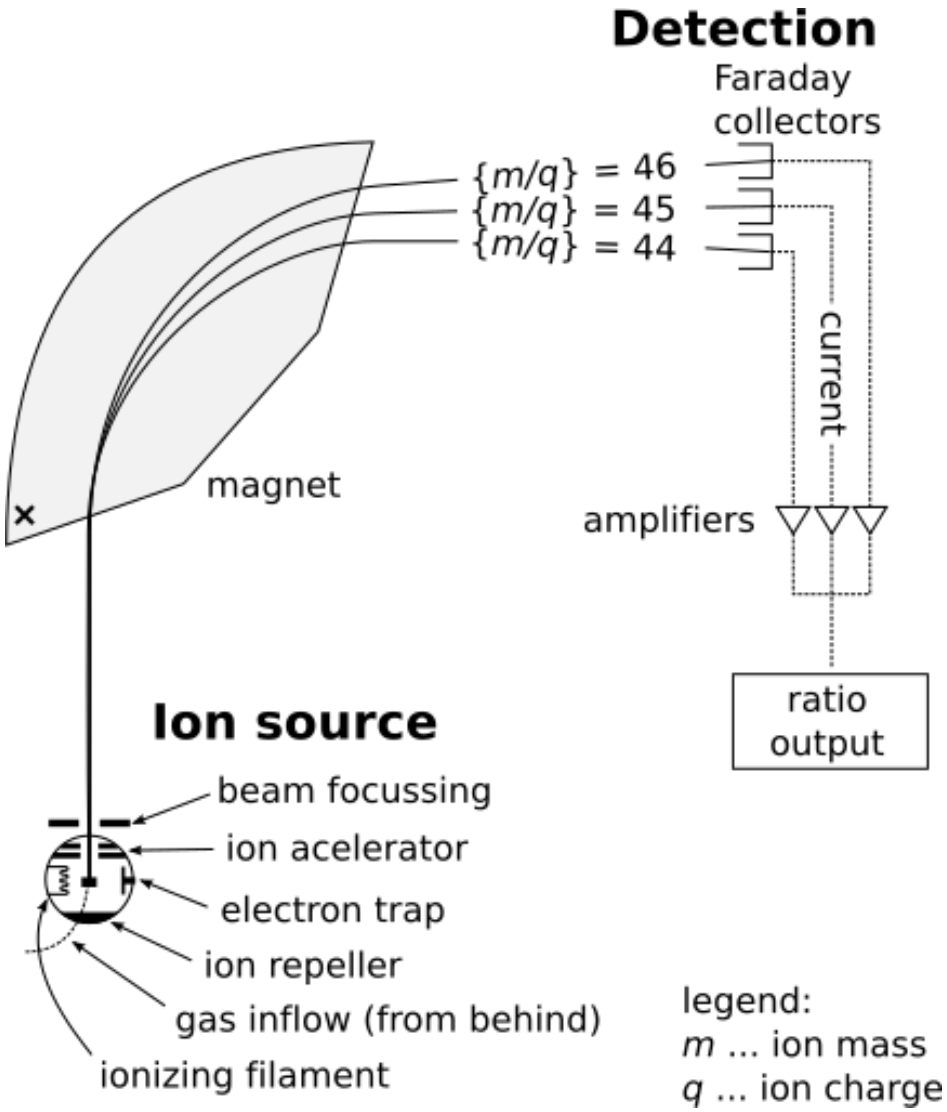
- Фаговый дисплей (phage display, фаговый дисплей)
- Пептид/белок экспонируется на поверхности фага
- Селекция на мишени → обогащение связавшихся вариантов
- Применения: эпитопы, антитела, мотивы связывания

Белковые чипы и аффинные методы



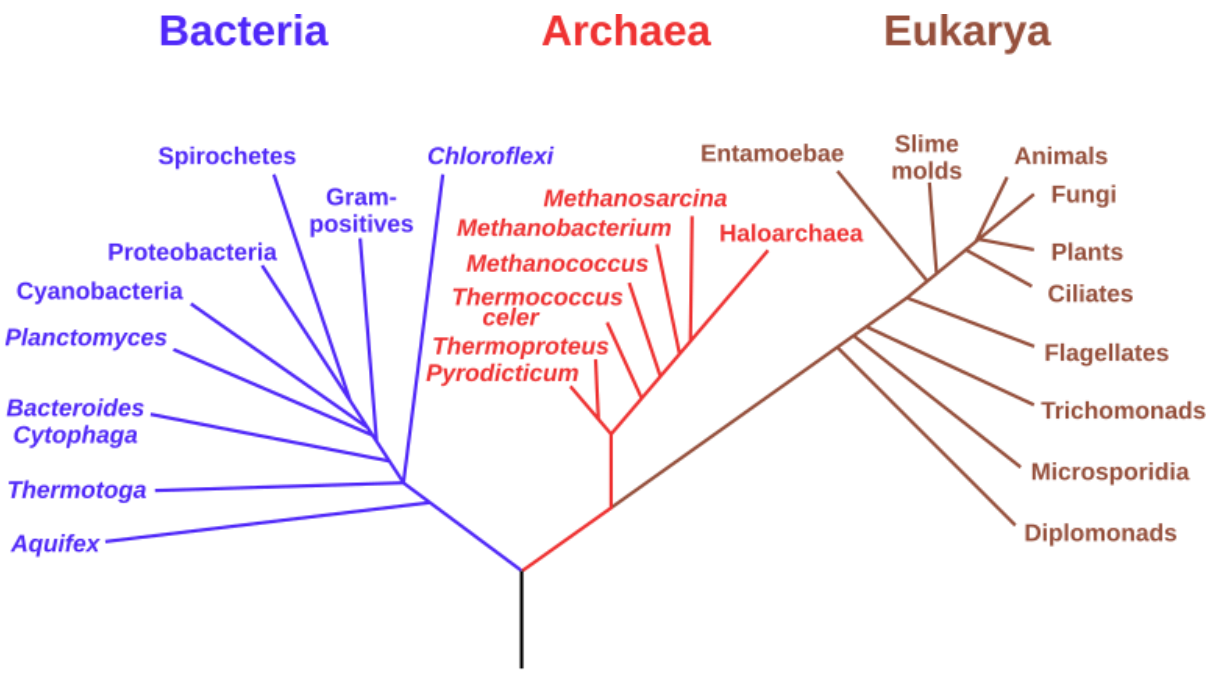
- Белковый чип (protein microarray, белковая микроматрица)
- Аффинные методы: pull-down (аффинная вытяжка), co-IP
- Вариант: AP-MS для состава комплекса
- Ключ: строгие контроли и повторяемость сигналов

Предсказание PTM и PPI: что реально можно ожидать



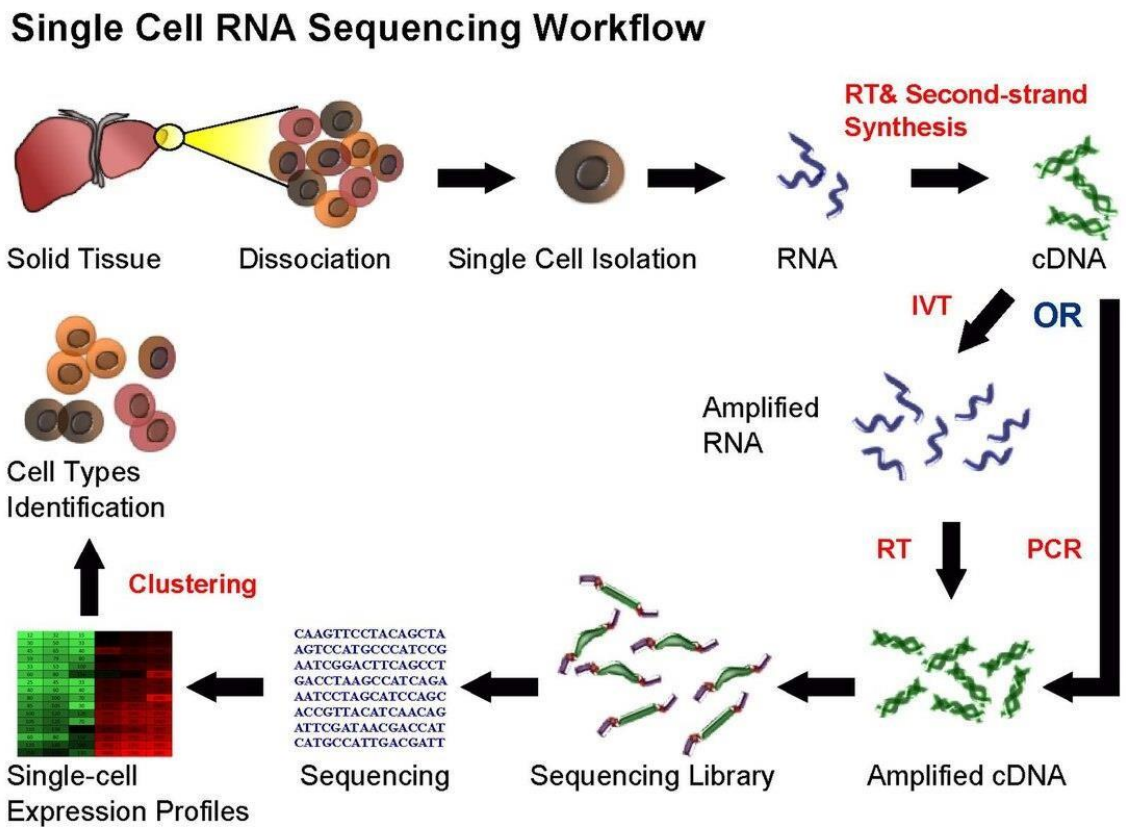
- PTM (post-translational modification, посттрансляционная модификация)
- Предсказания: мотивы, домены, коэволюция, структуры
- Важное правило: модель даёт кандидатов, не «истину»
- Лучшее сочетание: in silico → целевой эксперимент → валидация

Базы данных взаимодействий и мотивов (куда смотреть)



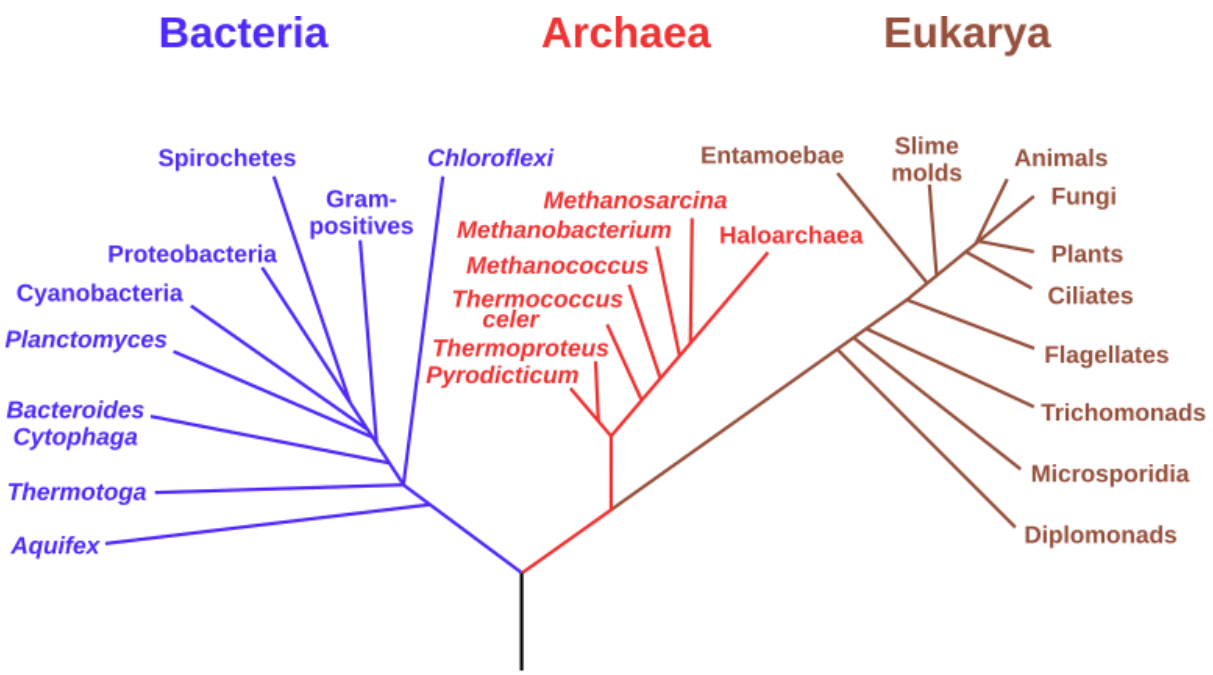
- PPI: STRING, BioGRID, IntAct — сети и подтверждения
- PTM: PhosphoSitePlus — модификации и сайты
- Мотивы ДНК: JASPAR — матрицы связывания факторов
- Важно читать «evidence»: эксперимент vs предсказание

Белок–ДНК взаимодействия и ChIP-подходы



- ChIP (chromatin immunoprecipitation, иммунопреципитация хроматина)
- ChIP-chip (ChIP + microarray, ChIP + микрочип)
- ChIP-seq (ChIP + sequencing, ChIP + секвенирование)
- Цель: найти сайты связывания факторов транскрипции и модификации хроматина

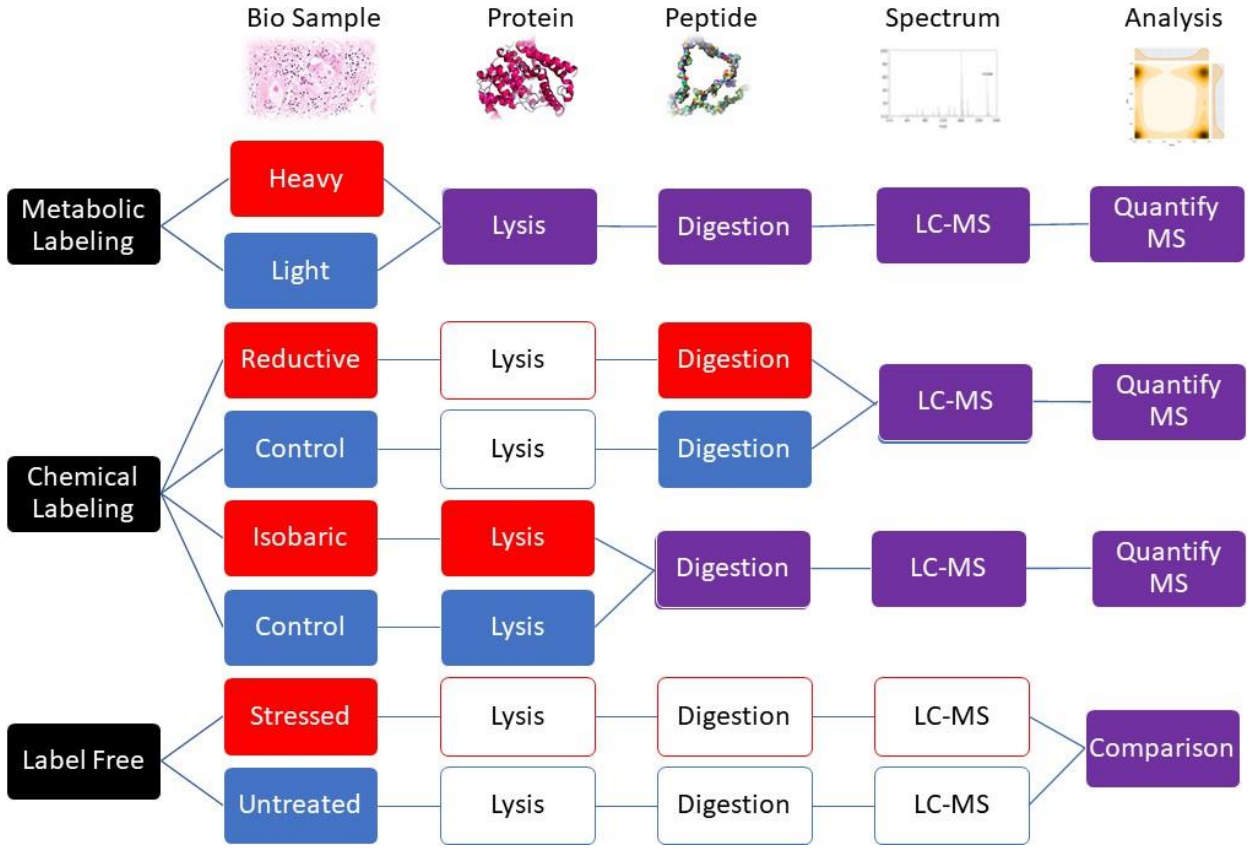
Зачем ChIP: регуляторные сети и биомедицина



- Карта сайтов связывания → модель регуляторной сети
- Сравнение условий: клеточный тип, стресс, болезнь
- Интерпретация: мотивы, промоторы, энхансеры, эпигенетика
- Связь с патологиями: варианты в регуляторных областях

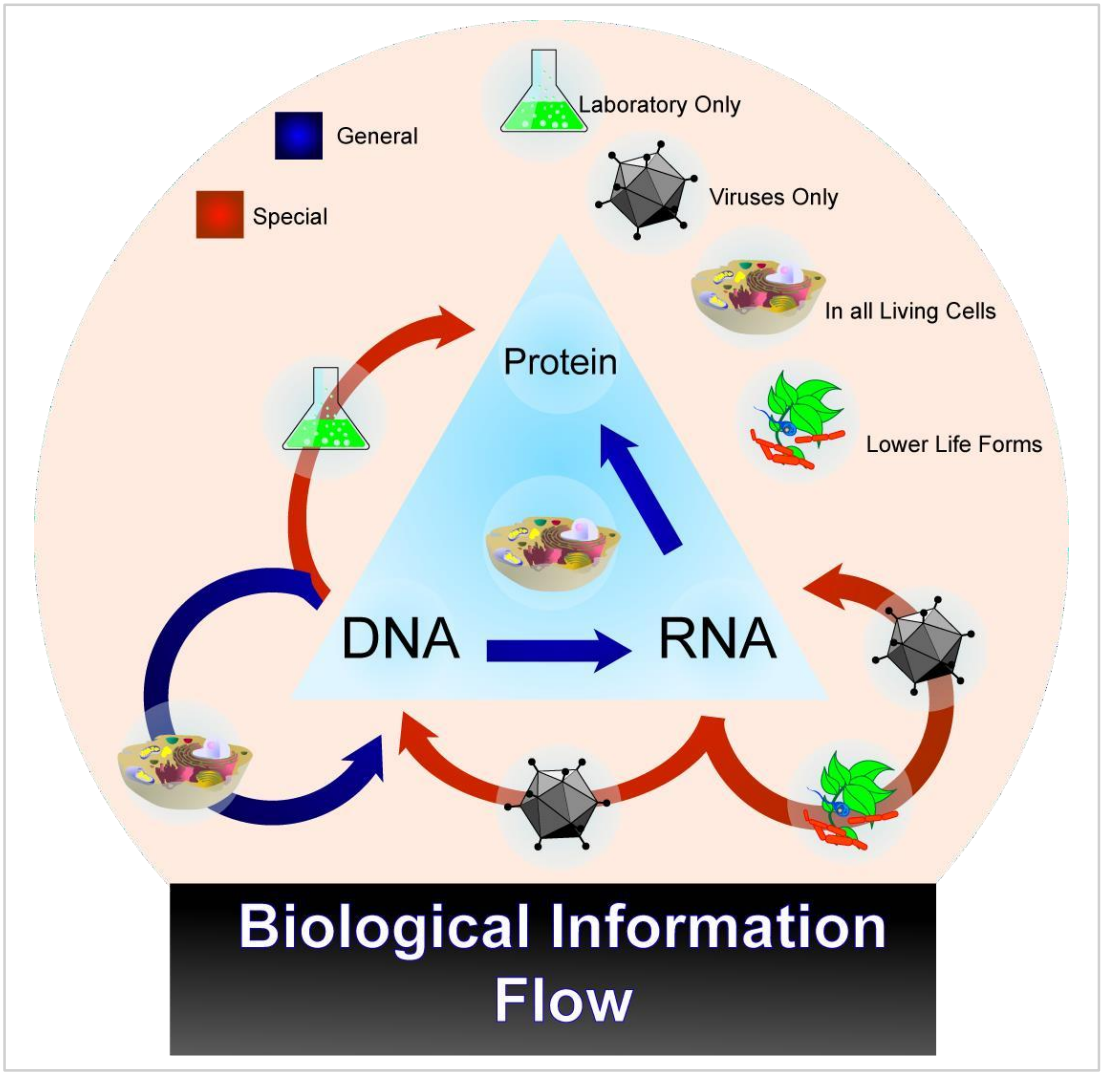
Протеомика: что мы измеряем

- Протеомика изучает состав, количество, формы, модификации и комплексы белков.
- Белки — это рабочий уровень клетки, а не просто “перевод” РНК.
- Главный инструмент — масс-спектрометрия
- Чаще всего измеряют пептиды и по ним восстанавливают белки.
- Поэтому протеомика ближе к функции и фенотипу, чем один только геном.



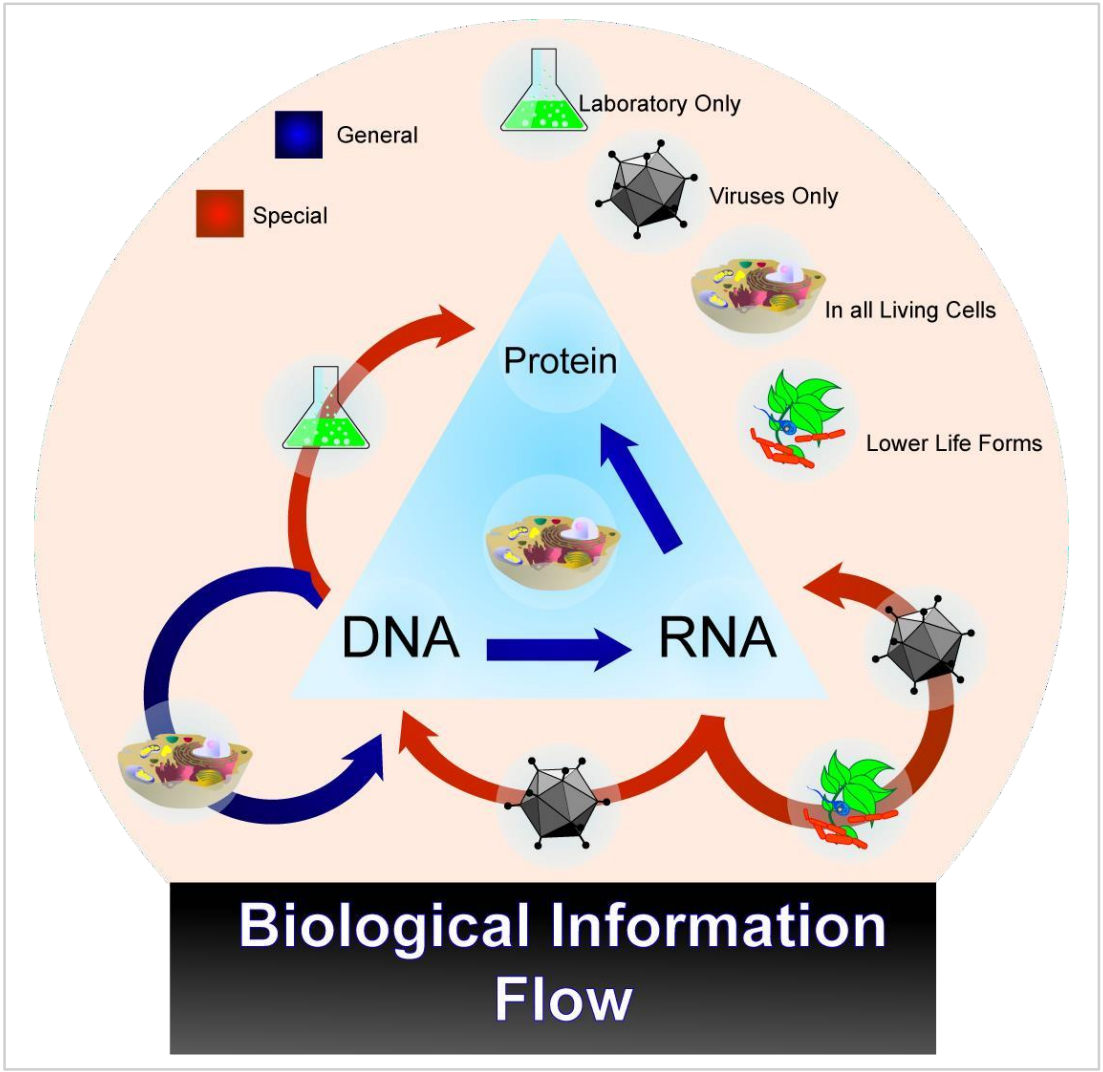
Постгеномные (негеномные) данные: общая картина

• К постгеномным данным относят транскриптом, протеом, метаболом и эпигеном.
• Геном задаёт возможности системы, а постгеномика показывает исполнение.
• Между слоями есть развилки: не вся РНК даёт белок, не всякий белок активен.
• Поэтому один уровень данных редко объясняет фенотип полностью.
• Нужны интегративные подходы и сопоставление разных “омик”.



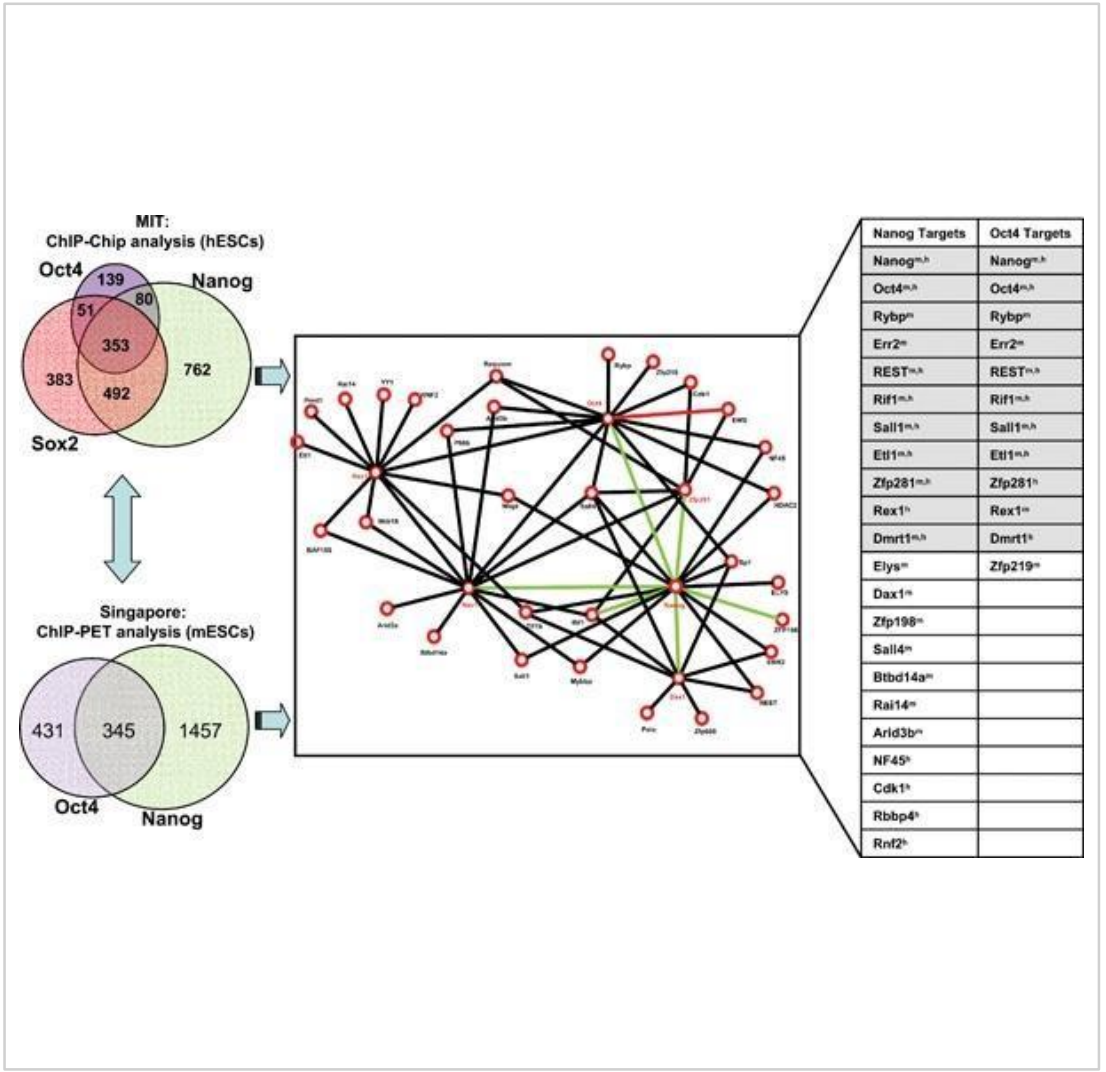
Веб-ориентированный автоматизированный мета-анализ

- Мета-анализ объединяет результаты разных исследований и наборов данных.
- Веб-платформы делают анализ воспроизводимым и удобным для повторения.
- Типовые шаги: поиск, фильтрация, унификация, статистика, интерпретация.
- Критичны метаданные, батч-эффекты и качество исходных экспериментов.
- Итог — устойчивые сигналы, а не выводы по одному датасету.



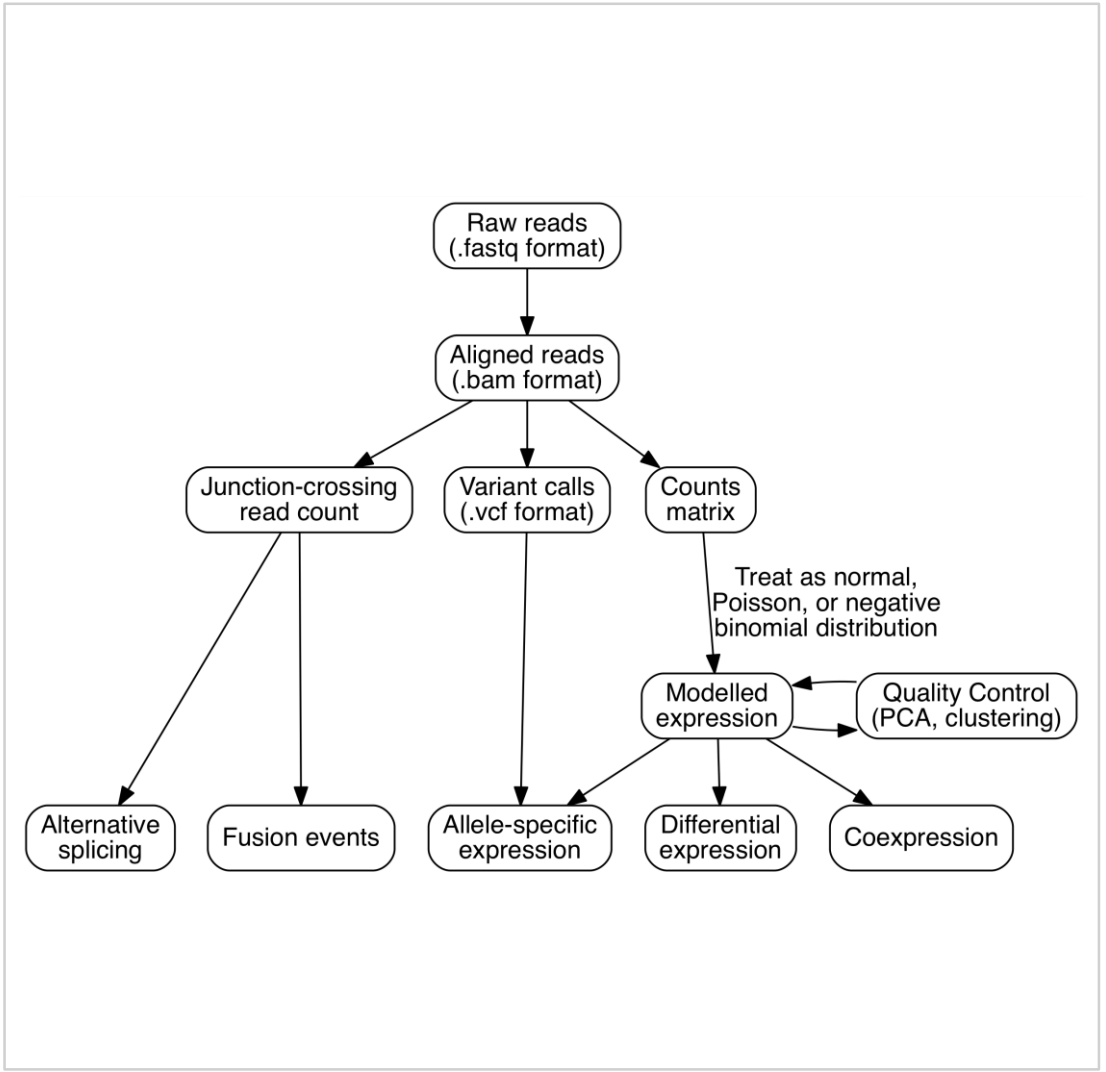
Прогнозирование взаимодействующих белков

• Цель — предсказать PPI (Protein–Protein Interaction, белок-белковое взаимодействие).
• Подсказки: домены, мотивы, ко-экспрессия, локализация, коэволюция.
• Сеть взаимодействий помогает видеть путь или комплекс, а не один белок.
• Предсказания нужно сверять с базами и независимыми данными.
• Результат — список кандидатов для последующей проверки.



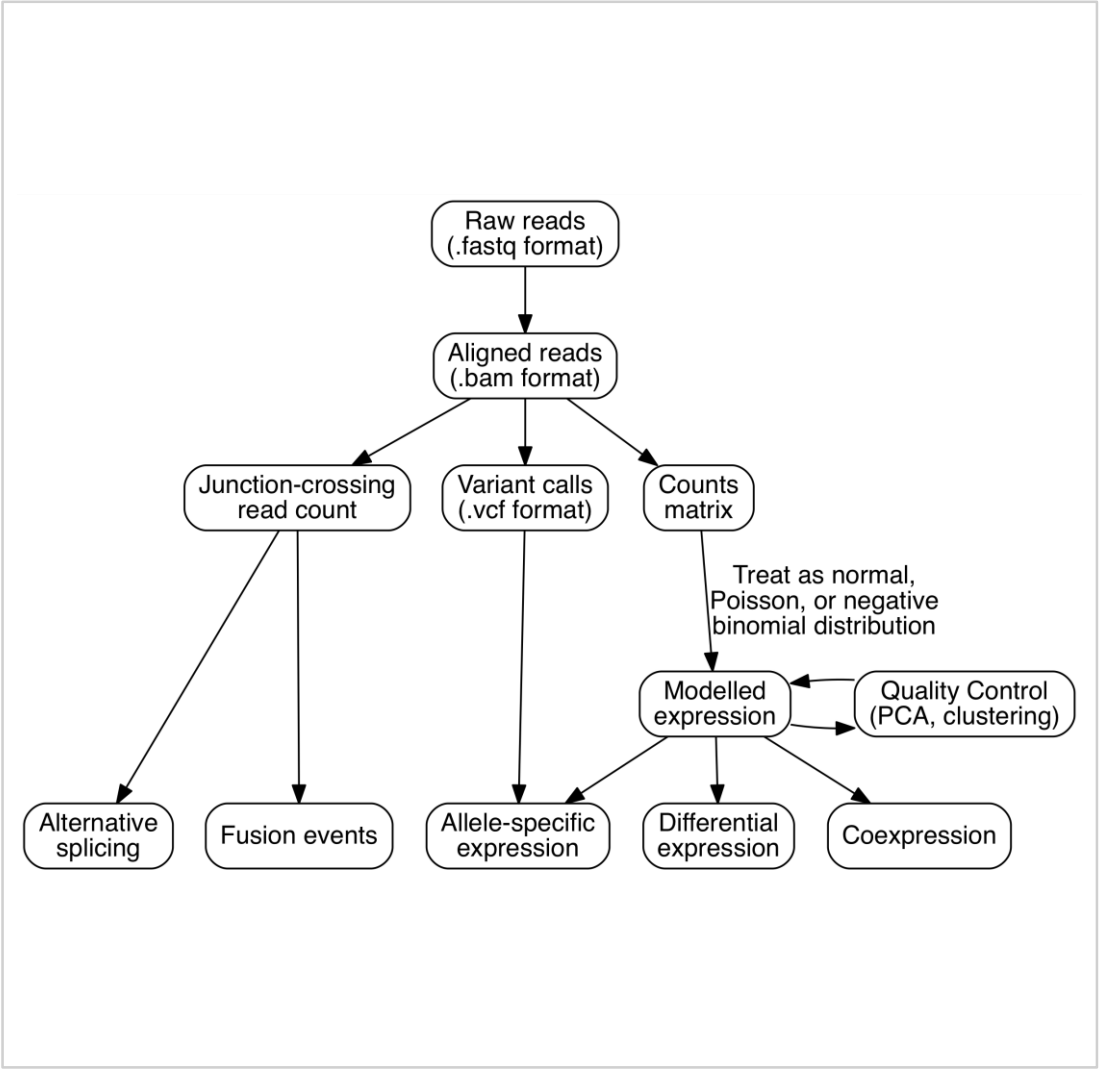
Транскриптомика: что измеряем и зачем

- Транскриптомика измеряет РНК и показывает, какие гены сейчас активны.
- Главный метод — RNA-seq (RNA sequencing, секвенирование РНК).
- Выход: уровни экспрессии, изоформы, сплайсинг и отдельные варианты.
- Метод удобен для сравнения условий, тканей и стадий развития.
- Но связь между РНК и белком не всегда линейна.



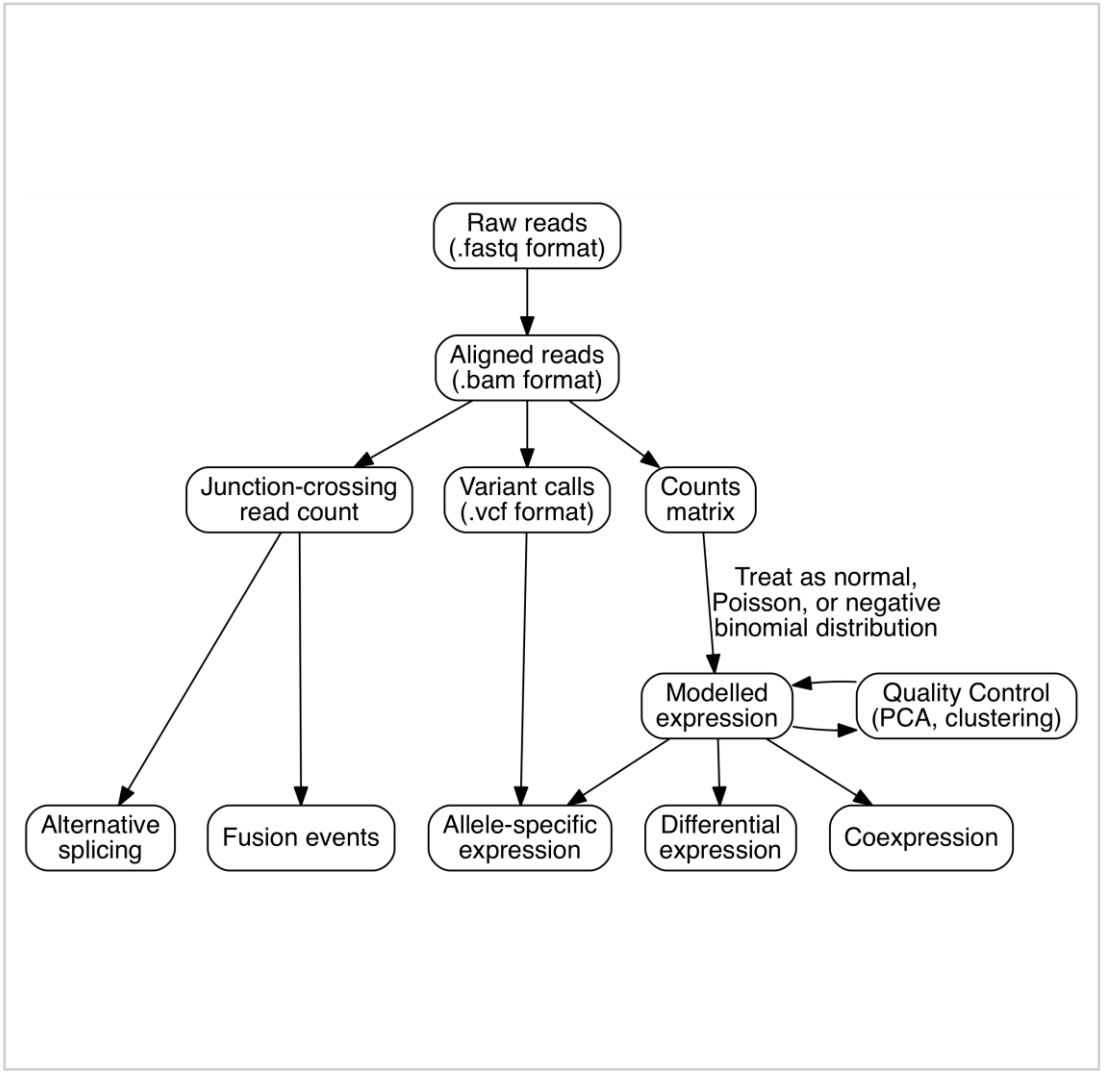
Картирование прочтений на геном

• Входные чтения записывают, например, в FASTQ (FASTQ format, формат FASTQ).
• Далее чтения выравнивают на геном или транскриптом.
• Результат обычно хранят в BAM (Binary Alignment/Map, бинарный формат выравниваний).
• После выравнивания строят матрицу счётов по генам и экзонам.
• От корректного картирования зависит весь дальнейший анализ.



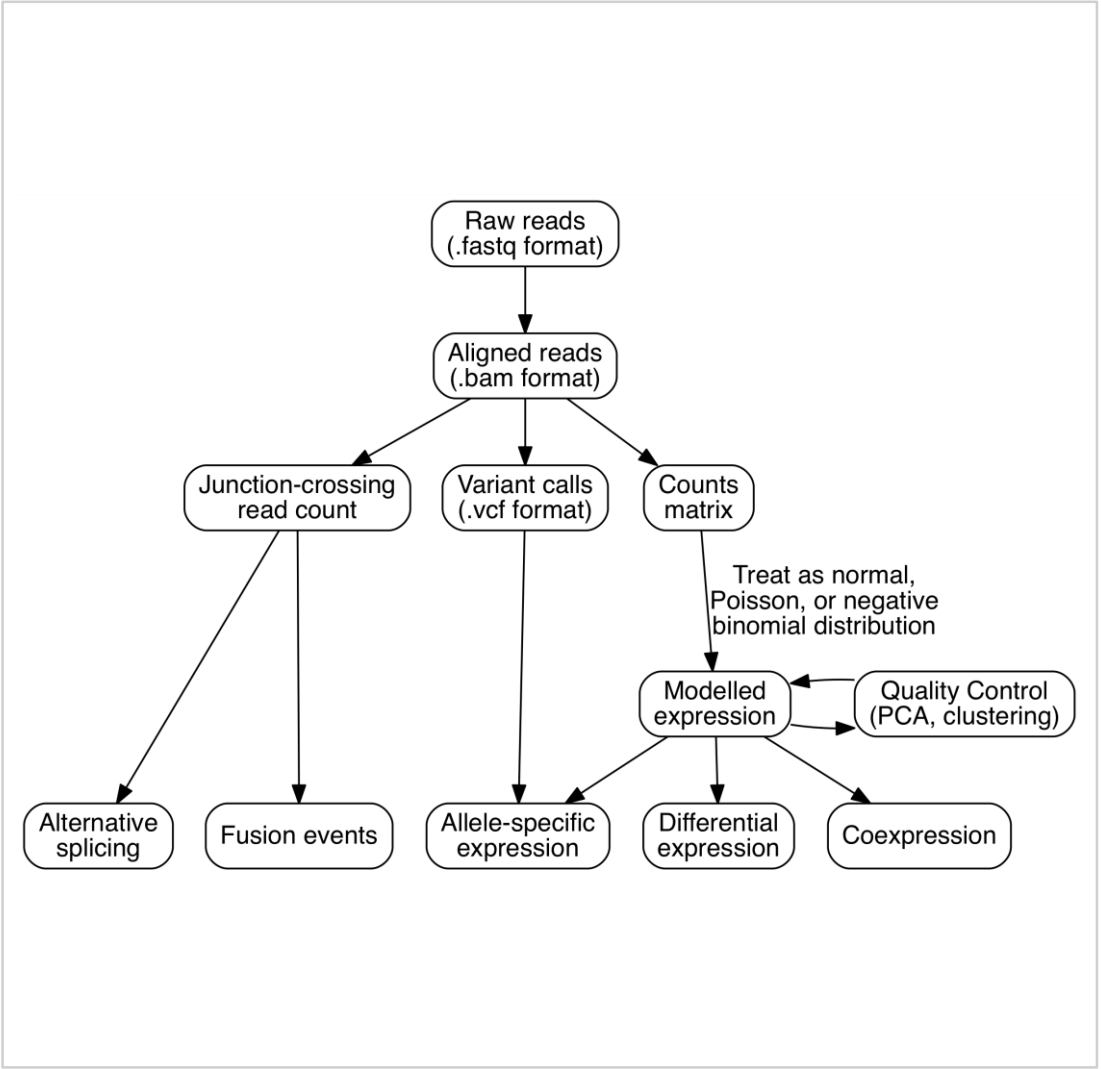
Контроль качества и фильтрация

• QC (Quality Control, контроль качества) проверяет, можно ли доверять данным.
• Смотрят качество чтений, адаптеры, длины, GC-состав и загрязнения.
• Фильтрация убирает технический шум и явный мусор.
• После очистки полезно снова смотреть метрики качества.
• Хороший QC уменьшает риск ложных биологических выводов.



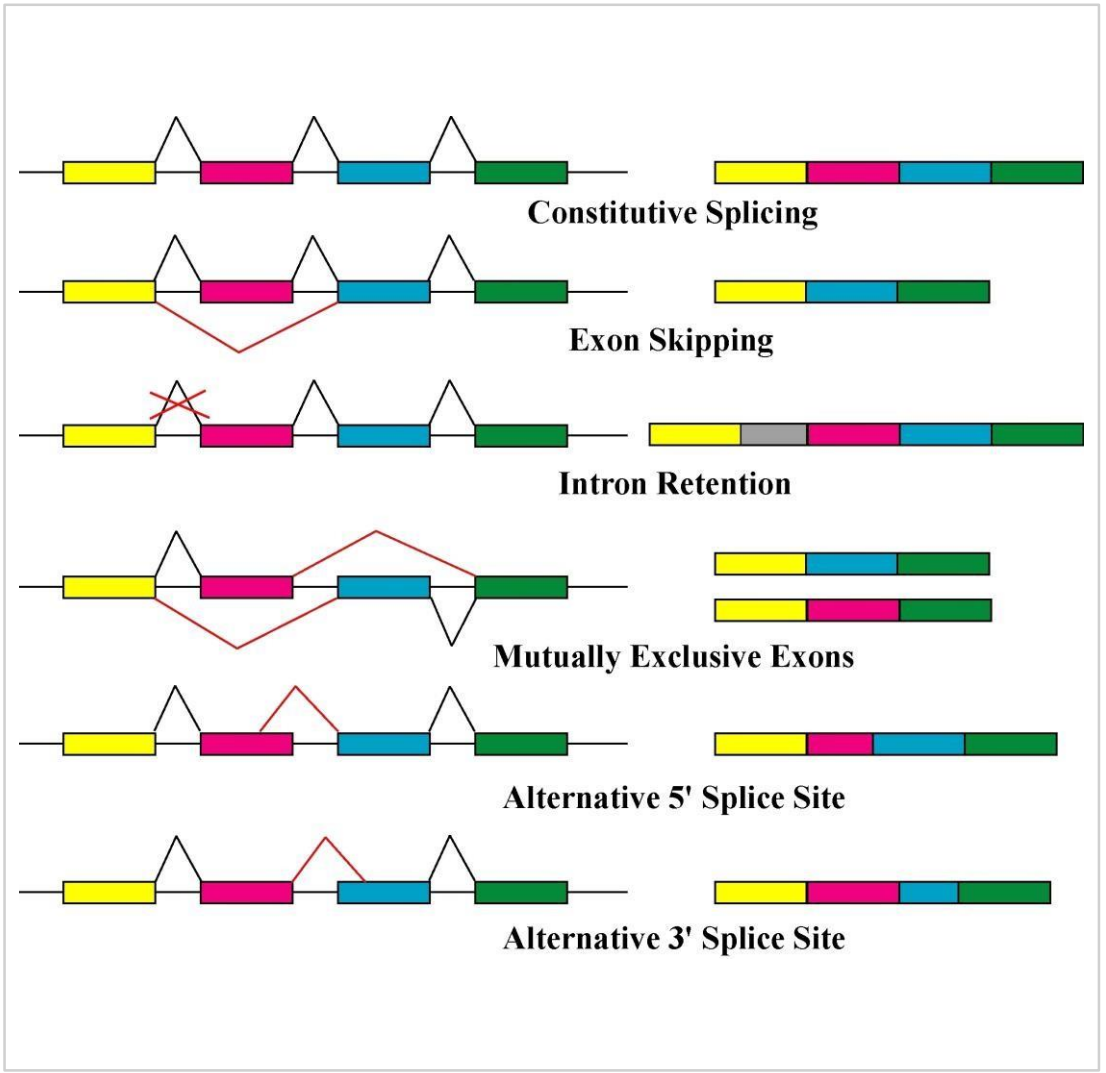
Оценка уровней экспрессии

• Сырые счёты нельзя напрямую сравнивать между образцами.
• Нужна нормализация по глубине секвенирования и длине транскриптов.
• Часто используют TPM (Transcripts Per Million, транскрипты на миллион).
• Далее сравнивают группы и оценивают статистическую значимость.
• Важно различать большой эффект и просто “значимый” эффект.



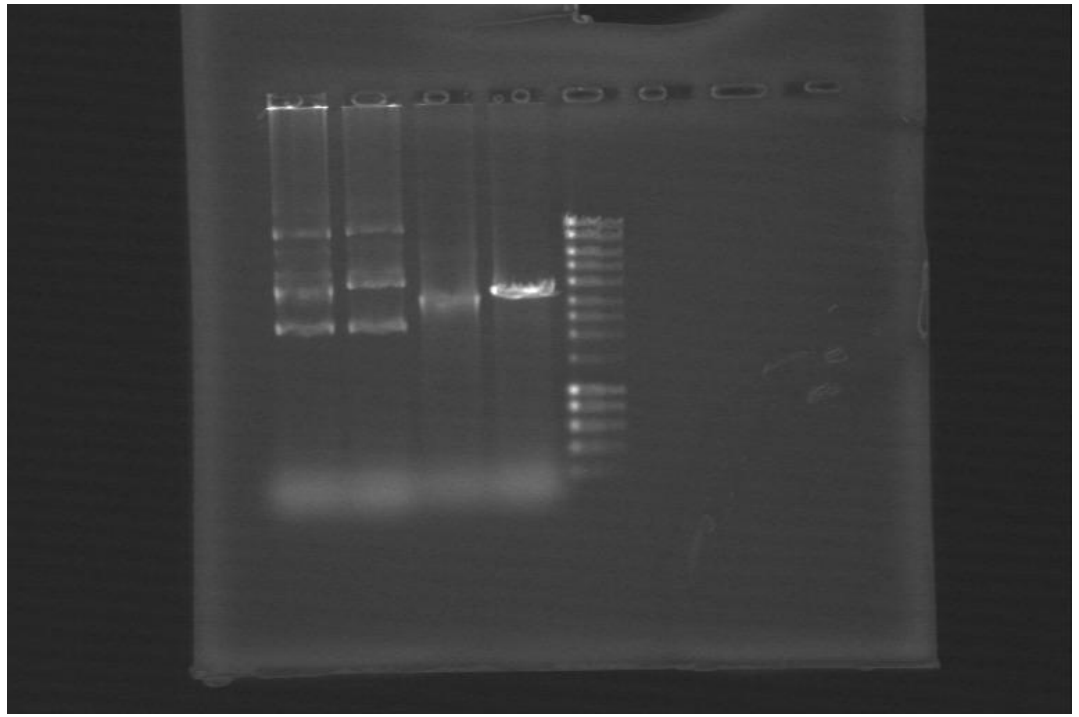
Включение экзонов и альтернативный сплайсинг

• Один ген может давать несколько вариантов зрелой РНК.
• Альтернативный сплайсинг меняет состав доменов будущего белка.
• Оценивают включение экзонов и типы сплайс-событий.
• Для такого анализа важны глубина и длина чтений.
• Это прямой мостик от транскриптомики к разнообразию белковых форм.



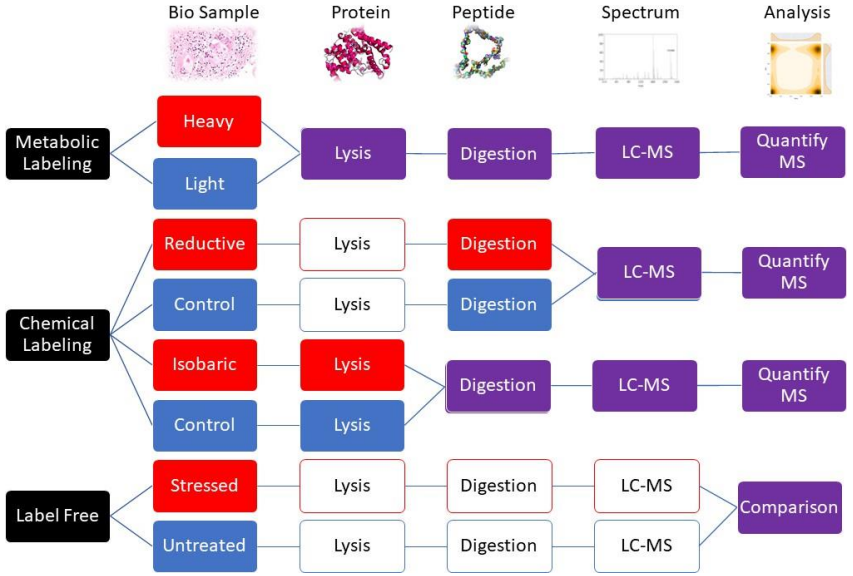
Протеомика до масс-спектрометрии: разделение белков

• Сложные смеси белков часто предварительно разделяют.
• SDS-PAGE (Sodium Dodecyl Sulfate Poly Acrylamide Gel Electrophoresis, электрофорез в полиакриламидном геле с додецилсульфатом натрия) разделяет белки по размеру.
• Гели полезны для контроля образца и выделения полос.
• Сейчас часто используют жидкостную хроматографию перед MS.
• Качество подготовки образца сильно влияет на итог.



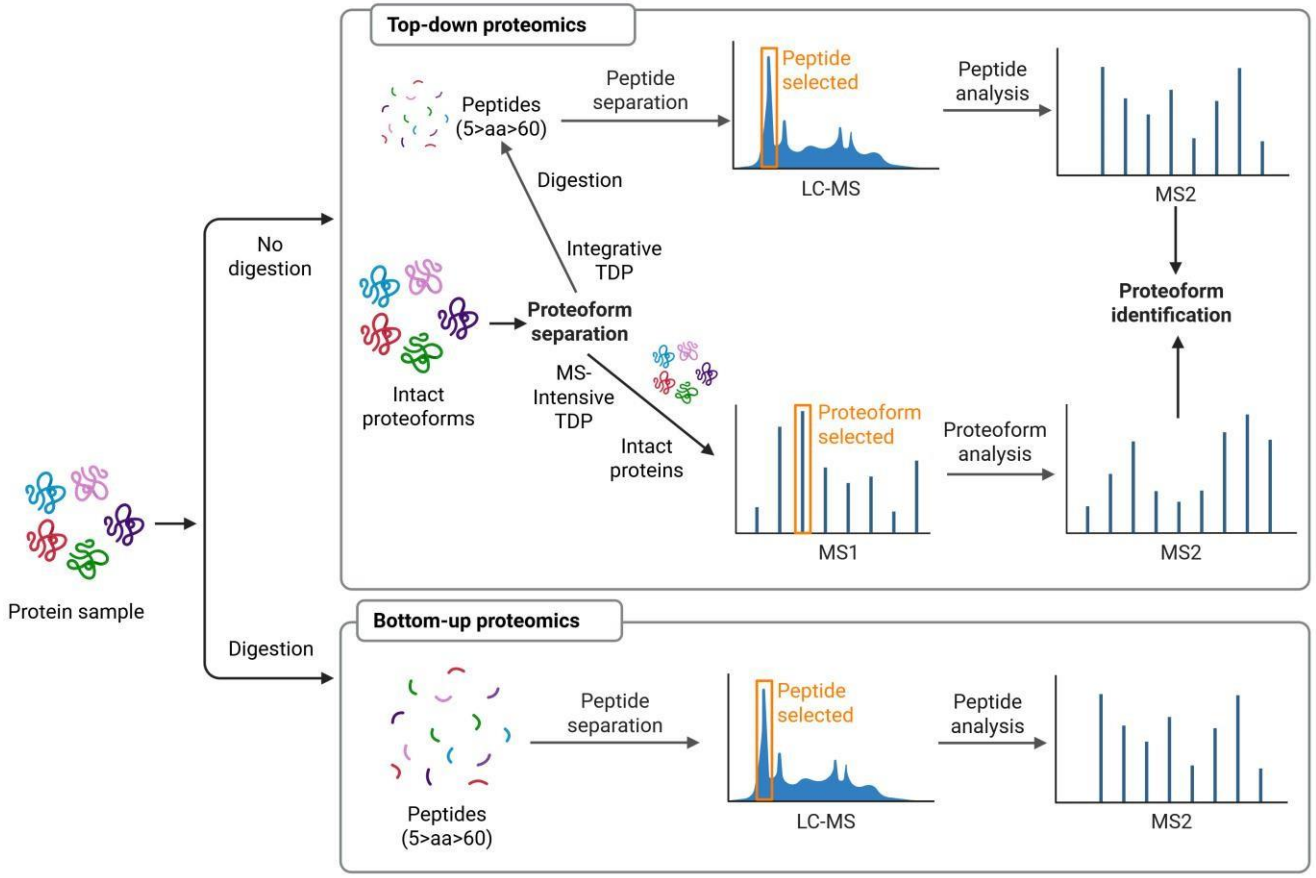
ВЭЖХ МС/МС: как читают белки через пептиды

- LC (Liquid Chromatography, жидкостная хроматография) разделяет смесь пептидов.
- MS/MS (Tandem Mass Spectrometry, тандемная масс-спектрометрия) даёт спектры фрагментов.
- По спектрам восстанавливают пептиды, а по ним — белки.
- Идентификация зависит от качества данных и базы поиска.
- Это центральная схема современной протеомики.



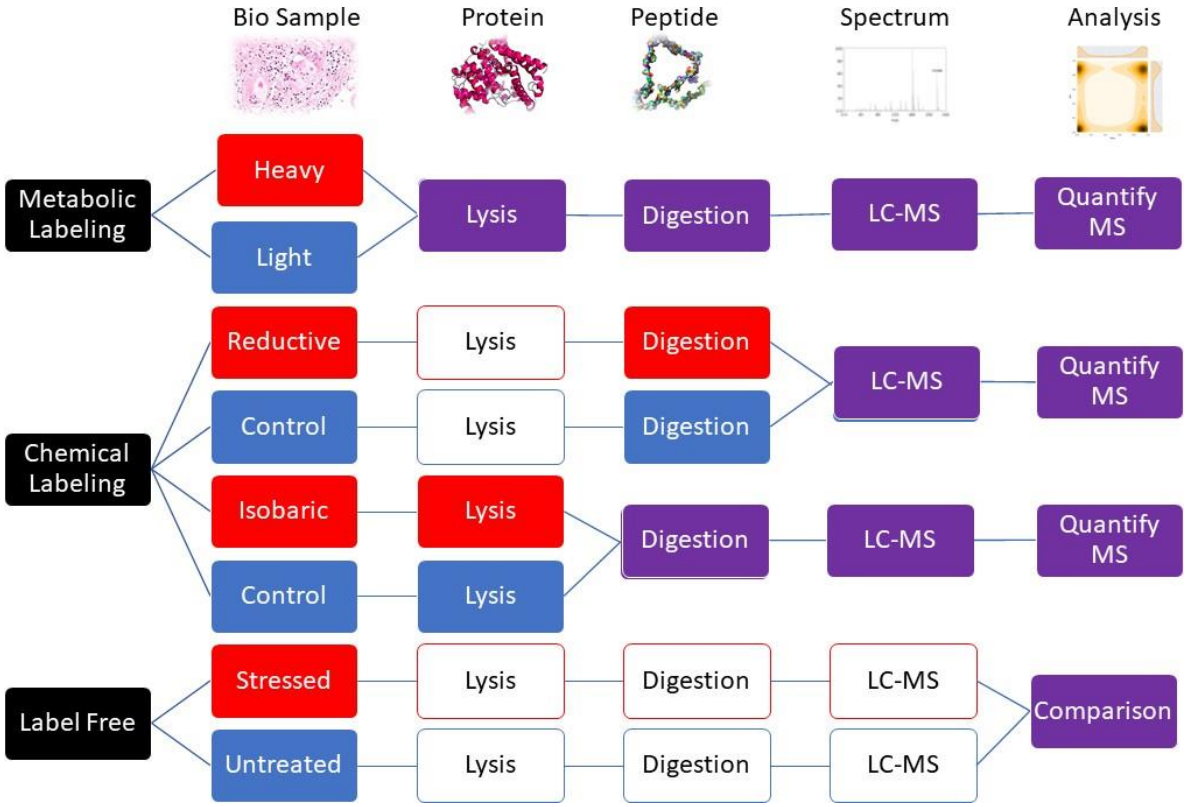
Bottom-up и Top-down протеомика

- Bottom-up: белок режут на пептиды и анализируют их массово.
- Top-down: измеряют целые протеоформы без полного расщепления.
- Bottom-up лучше подходит для сложных смесей и больших выборок.
- Top-down лучше сохраняет информацию о целой форме белка.
- Метод выбирают под задачу, а не “вообще”.



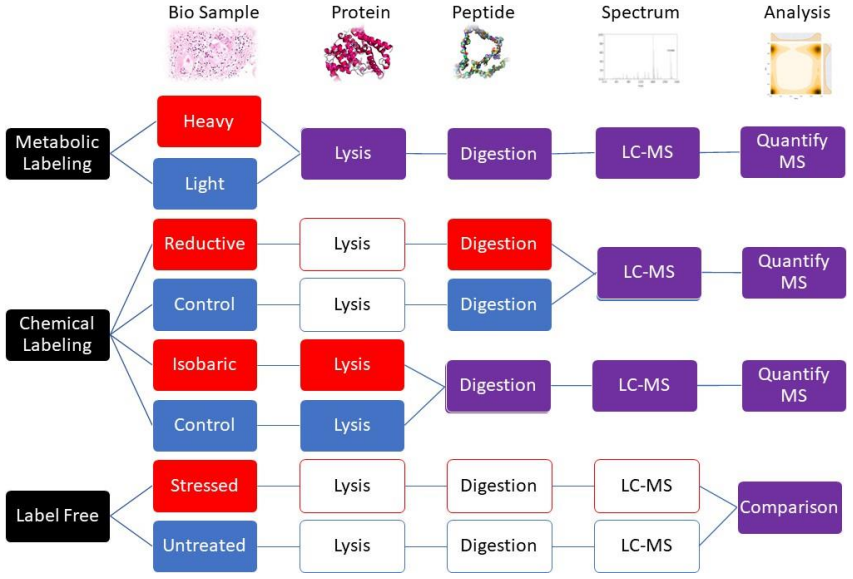
Аннотация по MS-данным: PSM и FDR

- Идентификация начинается с PSM (Peptide–Spectrum Match, соответствие “пептид–спектр”).
- Затем пептиды объединяют в белки или белковые группы.
- Нужен контроль ошибок: FDR (False Discovery Rate, доля ложных находок).
- Поэтому важны не только находки, но и уверенность в них.
- Хороший отчёт всегда показывает покрытие и статистику.



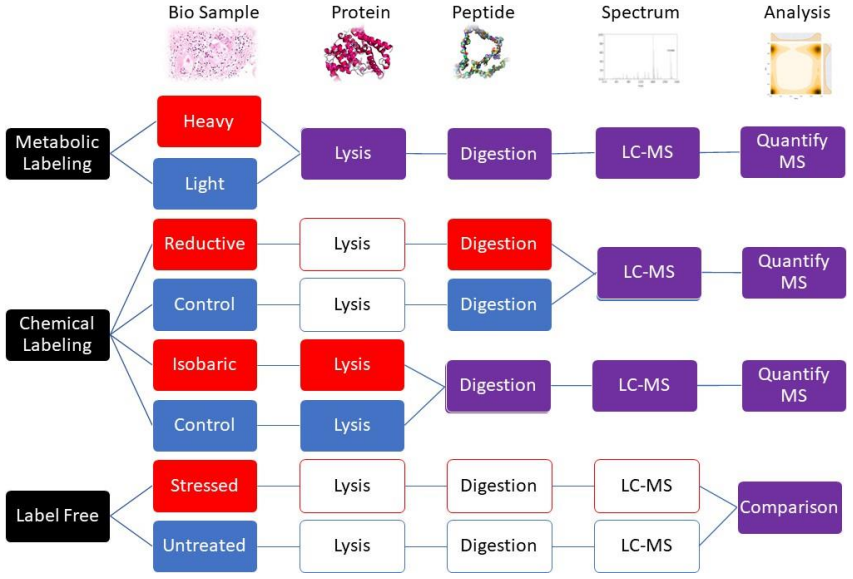
Трансляция in silico и протеолиз in silico

- In silico — это моделирование на компьютере.
- Трансляция in silico даёт теоретические белковые последовательности.
- Протеолиз in silico имитирует действие фермента, например трипсина.
- Так получают набор ожидаемых пептидов для поиска в базе.



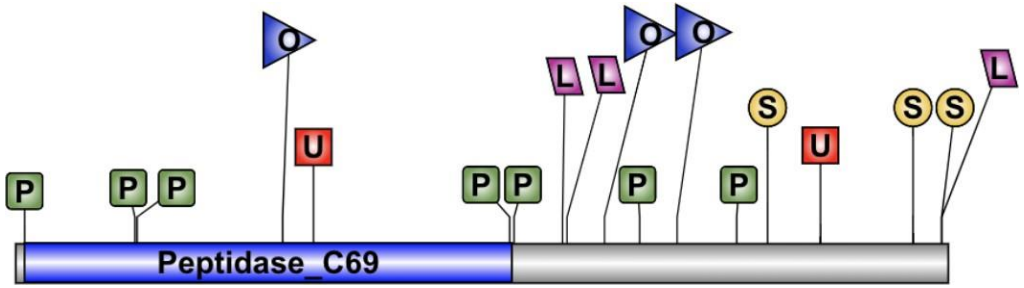
Количественная протеомика: label-free, TMT и другие подходы

• Количественная протеомика отвечает на вопрос “сколько белка стало больше или меньше”.
• Label-free сравнивает интенсивности без меток.
• TMT (Tandem Mass Tag, тандемная масс-метка) позволяет смешивать помеченные образцы.
• Выбор подхода зависит от дизайна, точности и бюджета.
• В любом случае важны реплики и аккуратная статистика.



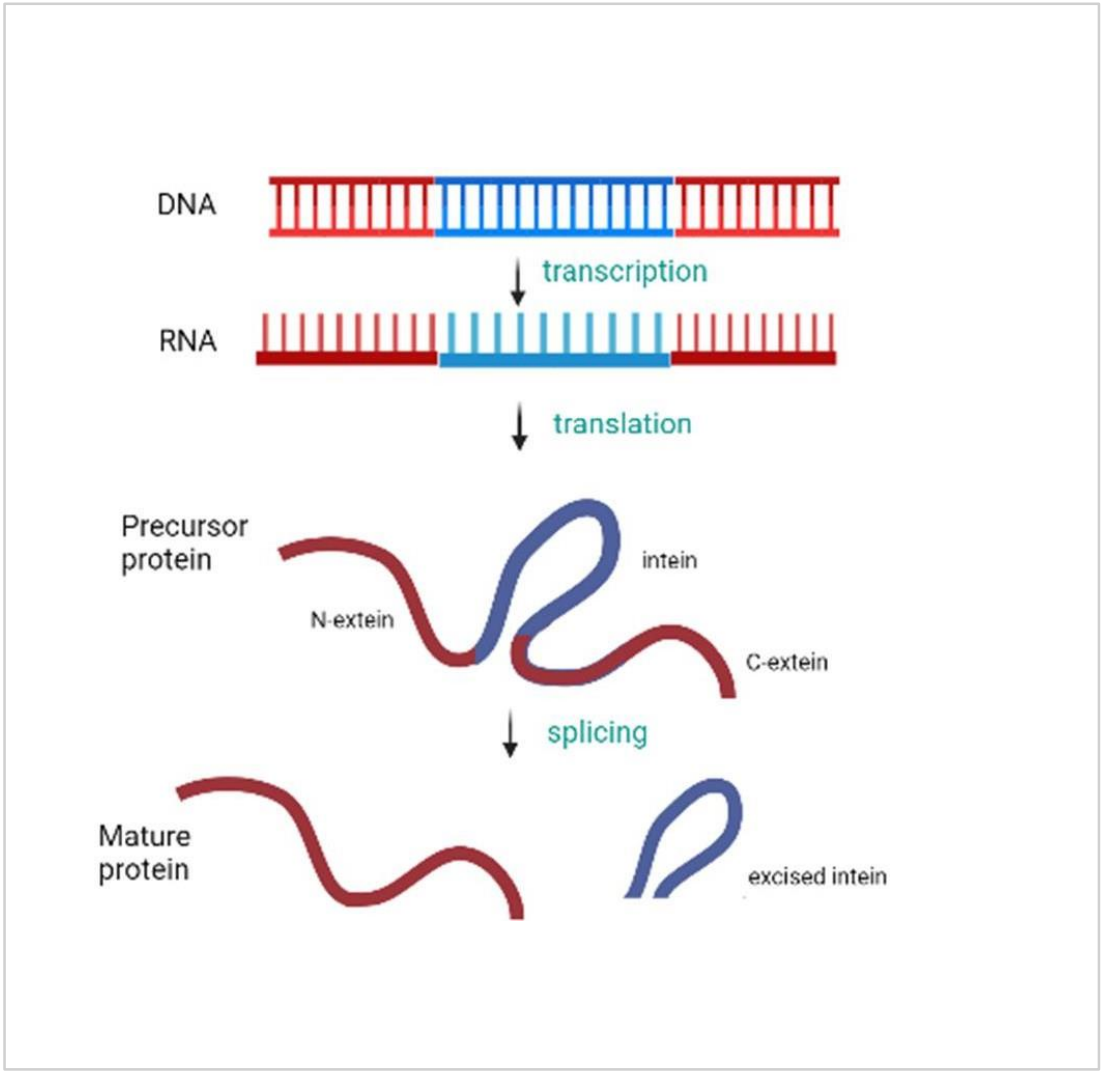
Посттрансляционные модификации белков

• PTM (Post-Translational Modification, посттрансляционная модификация) меняет свойства белка после синтеза.
• PTM влияет на активность, локализацию, стабильность и взаимодействия.
• Один белок может существовать в нескольких протеоформах.
• PTM — один из главных языков клеточного сигналинга.
• Поэтому один список белков ещё не описывает функцию полностью.



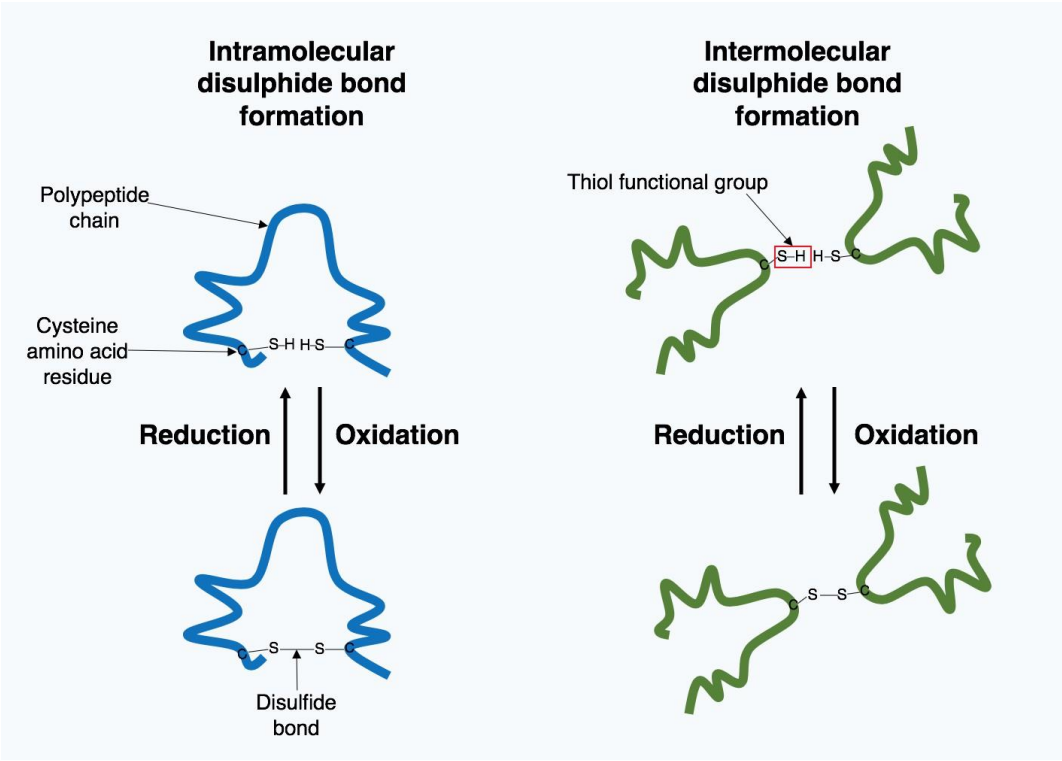
Ограниченный протеолиз и белковый сплайсинг

• Ограниченный протеолиз изменяет белок не разрушая его полностью.
• Так часто происходит созревание и активация белков-предшественников.
• Белковый сплайсинг удаляет внутренний участок и соединяет оставшиеся части.
• Такие процессы меняют структуру, активность и взаимодействия.
• Их нельзя понять только по последовательности ДНК.



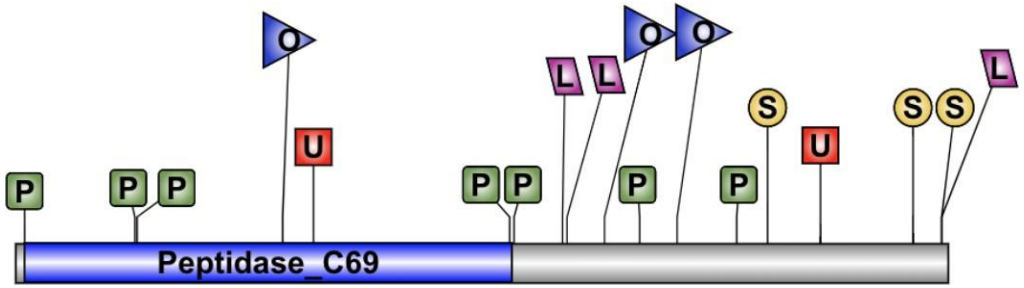
Дисульфидные связи

• Дисульфидные связи стабилизируют структуру белка.
• Они бывают внутри одной цепи и между разными цепями.
• Особенно важны для секретируемых белков и рецепторов.
• Нарушение дисульфидов может вести к неправильному сворачиванию.
• Это пример структурной посттрансляционной модификации.



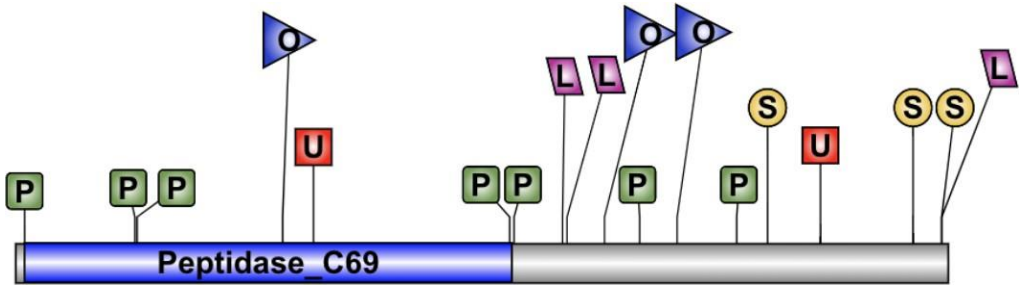
Присоединение малых химических групп

• Фосфорилирование часто служит быстрым сигнальным переключателем.
• Ацетилирование и метилирование регулируют свойства белков и хроматина.
• Гликозилирование особенно важно для мембранных и секретируемых белков.
• Карбоксилирование и другие модификации тонко настраивают функцию.
• Один сайт может участвовать в конкурирующих модификациях.



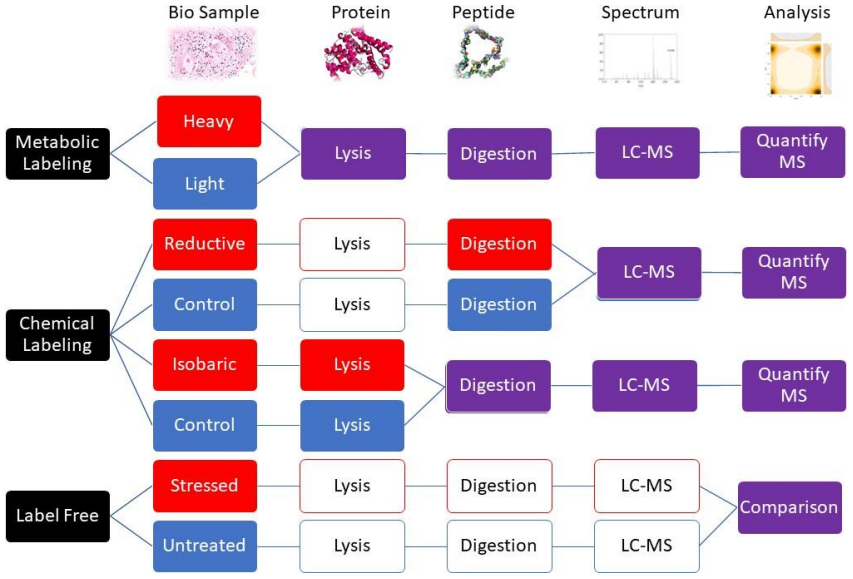
Убиквитинилирование и сумоилирование

• Убиквитинилирование часто метит белки на деградацию, но не только.
• Сумоилирование чаще связано с регуляцией взаимодействий и ядерных процессов.
• Эти метки могут образовывать цепочки с разным смыслом.
• Они регулируют стабильность, локализацию и судьбу белков.
• Нарушение этих систем важно для онкологии и нейродегенерации.



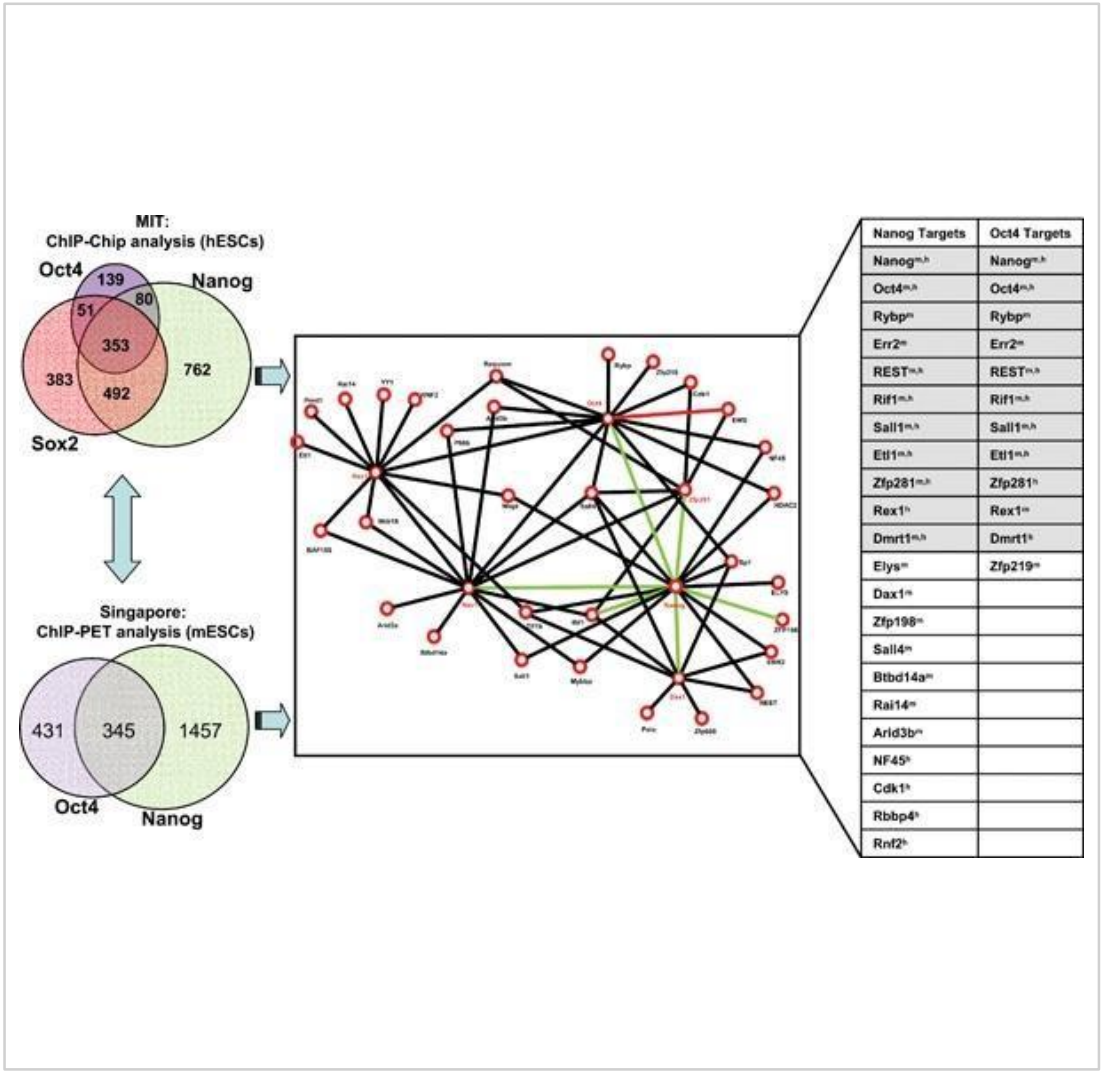
Масс-спектрометрия для анализа PTM

• Модифицированные пептиды часто редки и теряются в общей смеси.
• Поэтому для PTM часто делают обогащение целевых фракций.
• Поиск PTM требует отдельных настроек и строгого контроля ошибок.
• Итог — карта сайтов модификаций и их динамики.
• Затем PTM связывают с путями, киназами и состояниями клетки.



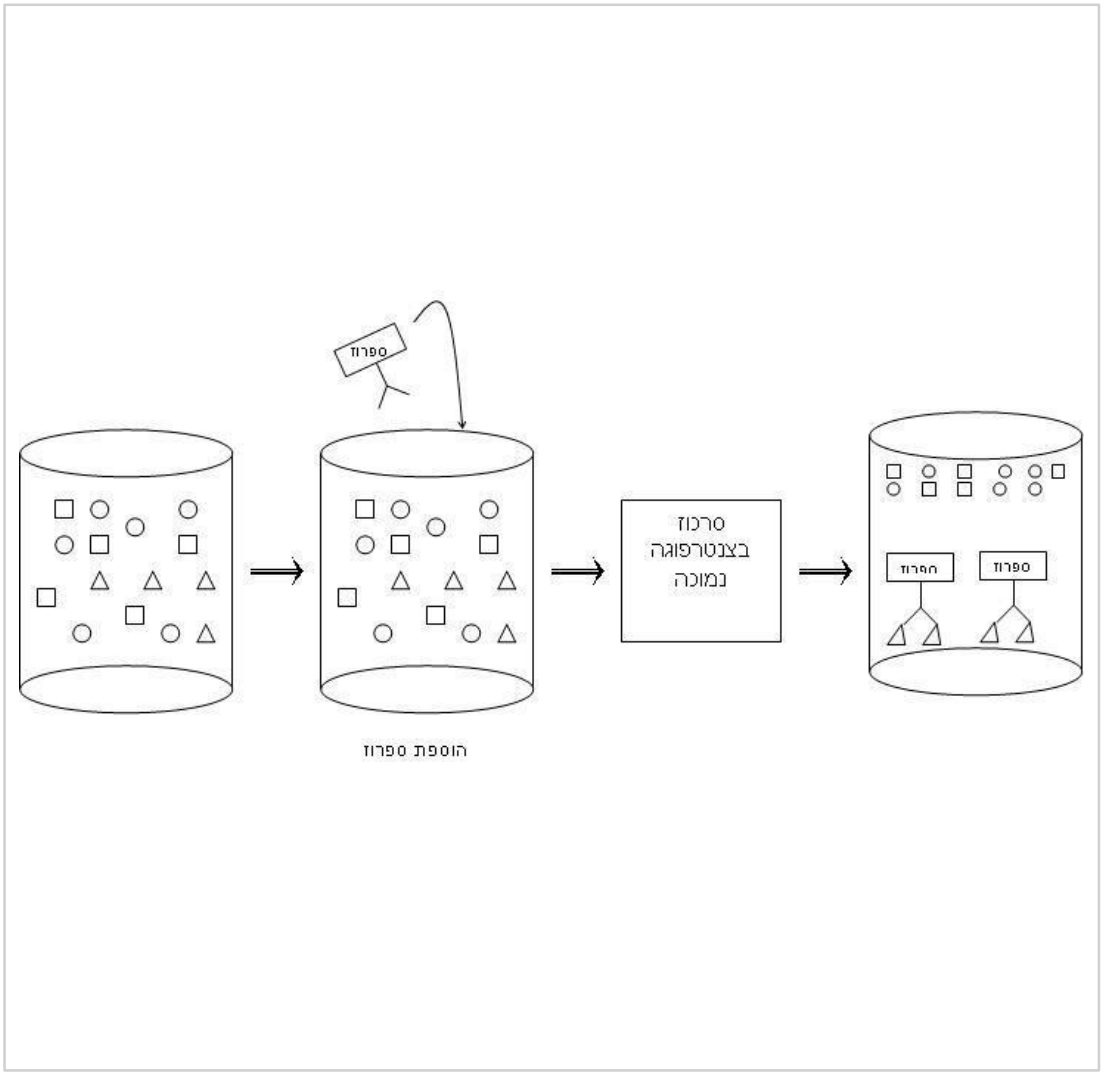
Белковые комплексы: почему они важны

• Во многих случаях функциональная единица клетки — это комплекс, а не один белок.
• Один и тот же белок может входить в разные комплексы.
• Состав комплексов меняется между условиями и стадиями клетки.
• Изучение комплексов помогает объяснять механизм фенотипа.
• Поэтому протеомика всё чаще работает на сетевом уровне.



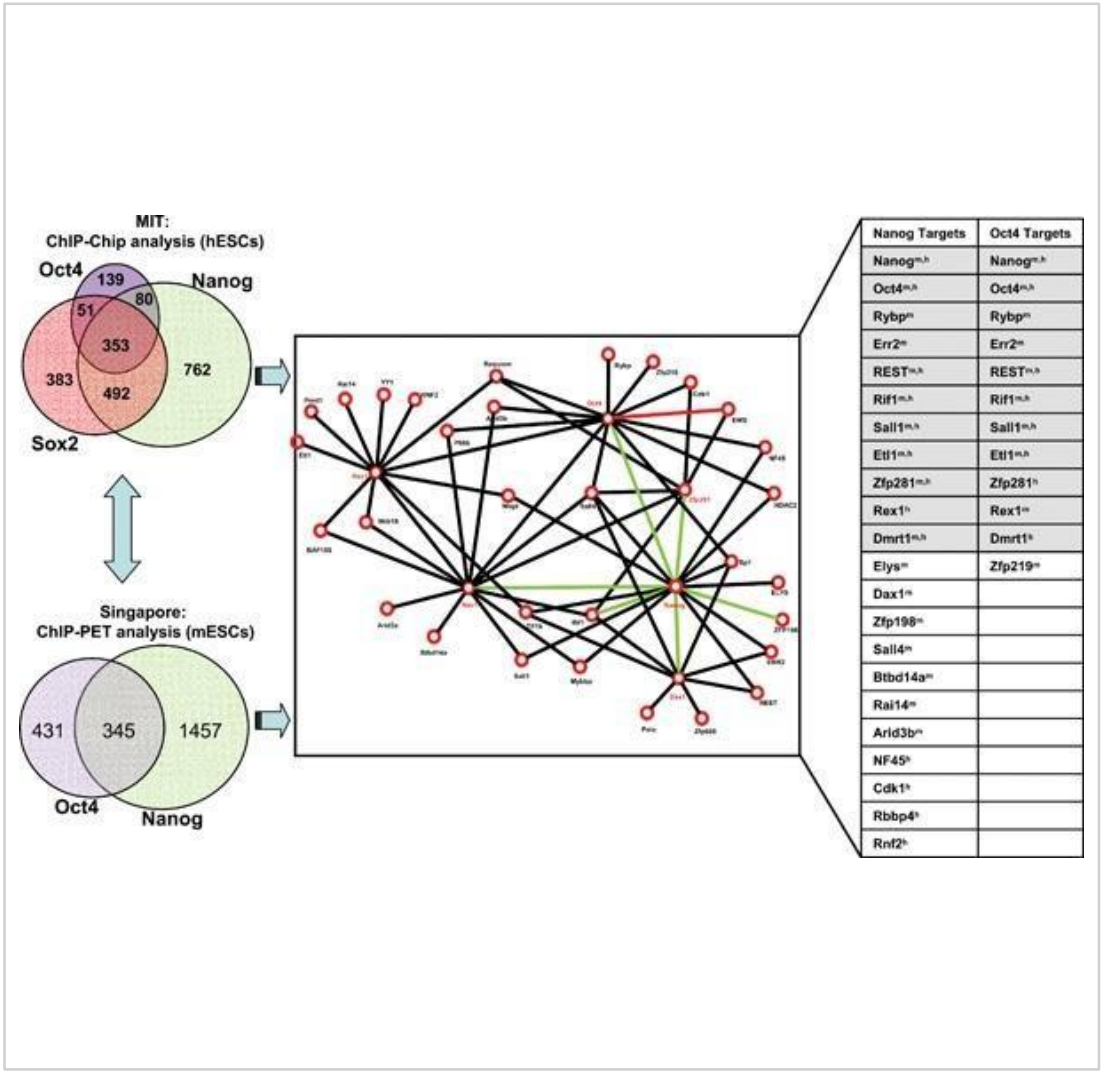
Аффинные методы: Co-IP и AP-MS

• Co-IP (Co-immunoprecipitation, ко-иммунопреципитация) вытягивает белок и его партнёров антителом.
• AP-MS (Affinity Purification–Mass Spectrometry, аффинная очистка с масс-спектрометрией) затем определяет состав комплекса.
• Нужны контроли на фон и неспецифическое связывание.
• Такие методы дают экспериментальные карты взаимодействий.
• Затем их сравнивают между условиями и клеточными состояниями.



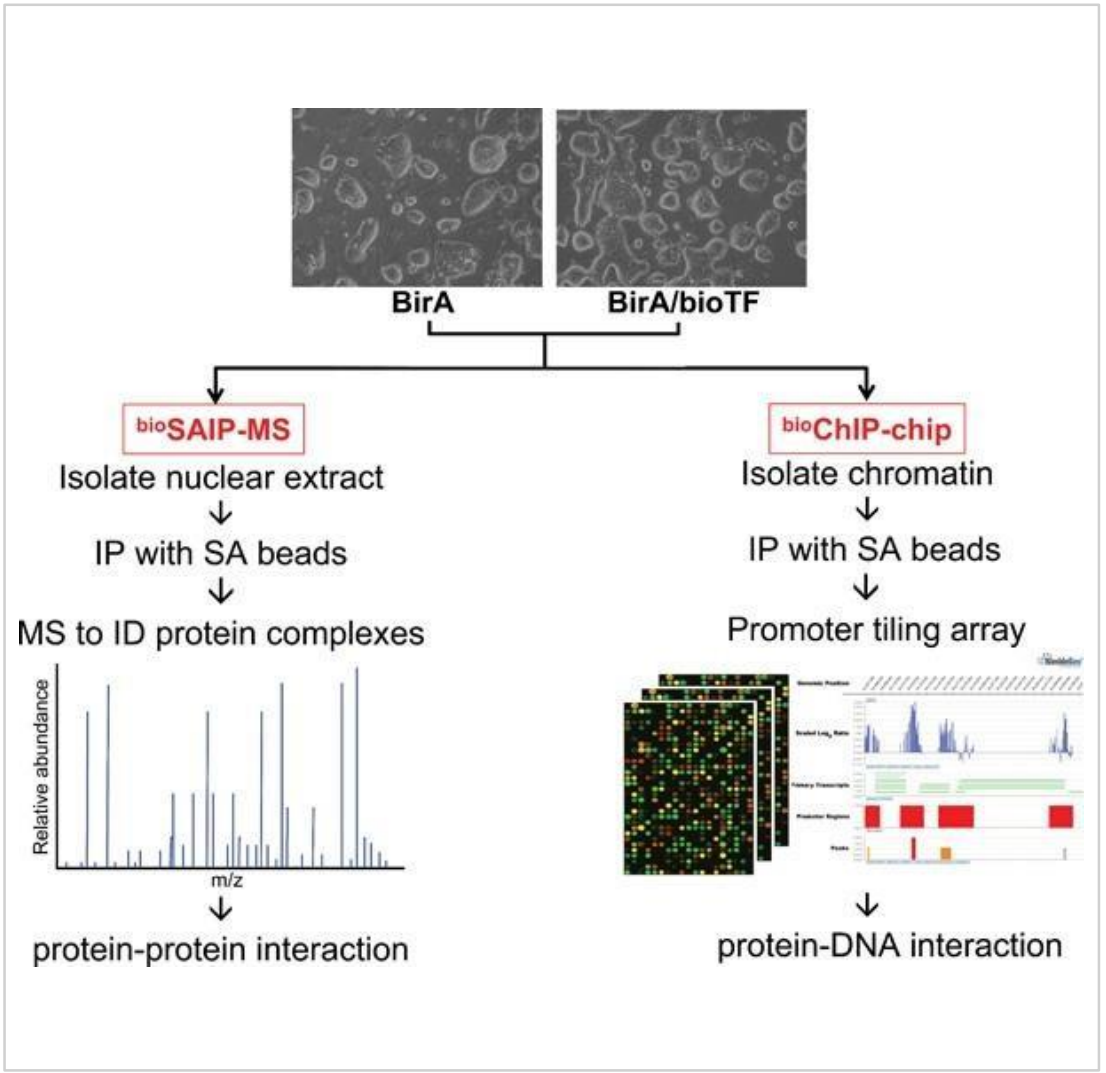
Карты взаимодействий между белками в клетке

• Сеть взаимодействий показывает модули, кластеры и “узлы” регуляции.
• Высокосвязные узлы часто оказываются функционально важными.
• Сети можно сравнивать между условиями и состояниями клетки.
• Но любая сеть зависит от метода и неполноты данных.
• Поэтому сеть — это рабочая карта, а не окончательная истина.



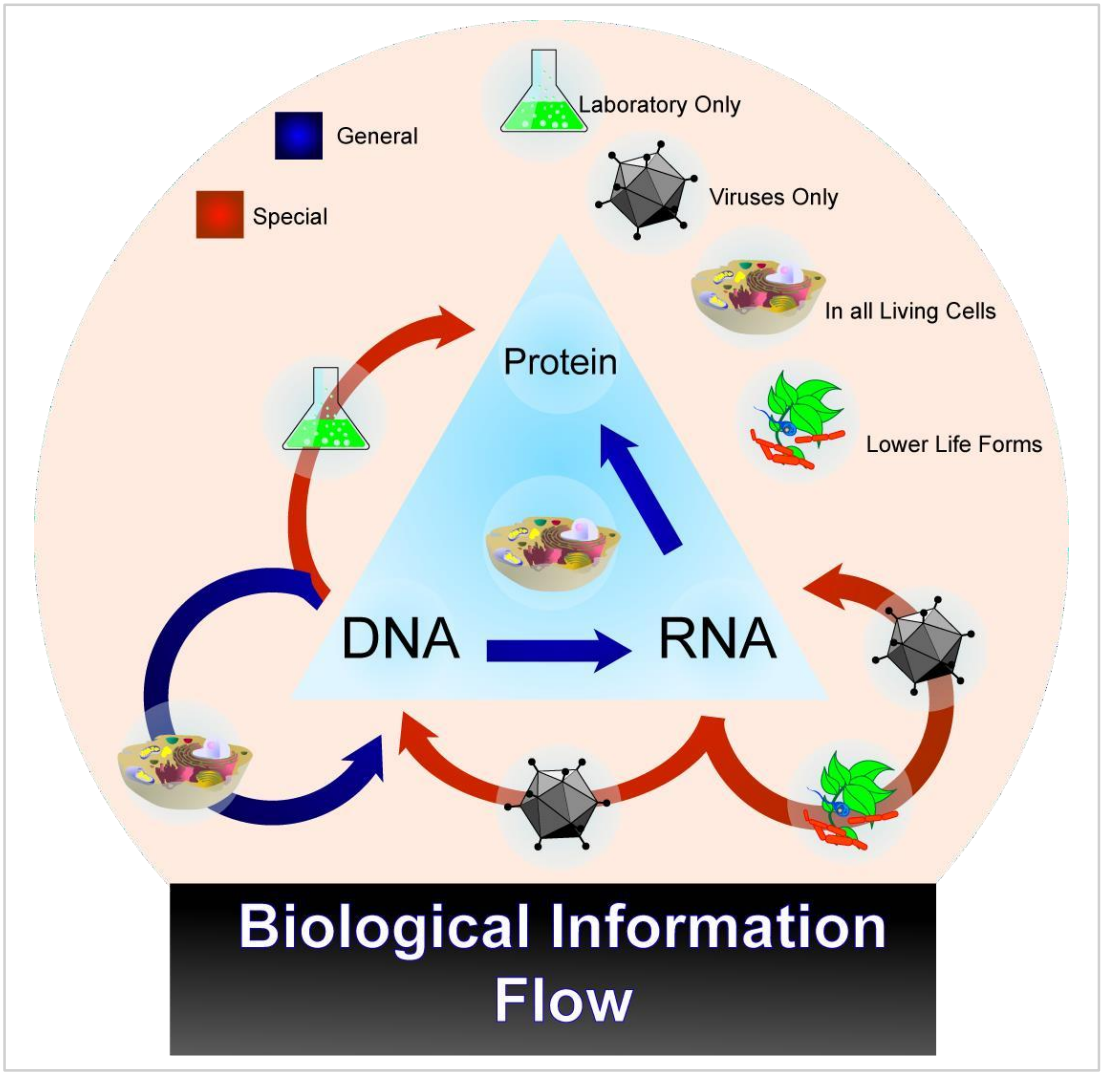
Прогнозирование и аннотирование PPI (белок-белковые взаимодействия)

• Предсказания дополняют экспериментальные сети там, где данных мало.
• Используют доменные модели, структуры, коэволюцию и ко-экспрессию.
• Аннотирование связывает взаимодействия с путями и функциями.
• Нужна проверка по базам и независимым наборам данных.
• Результат — ранжированный список наиболее правдоподобных связей.



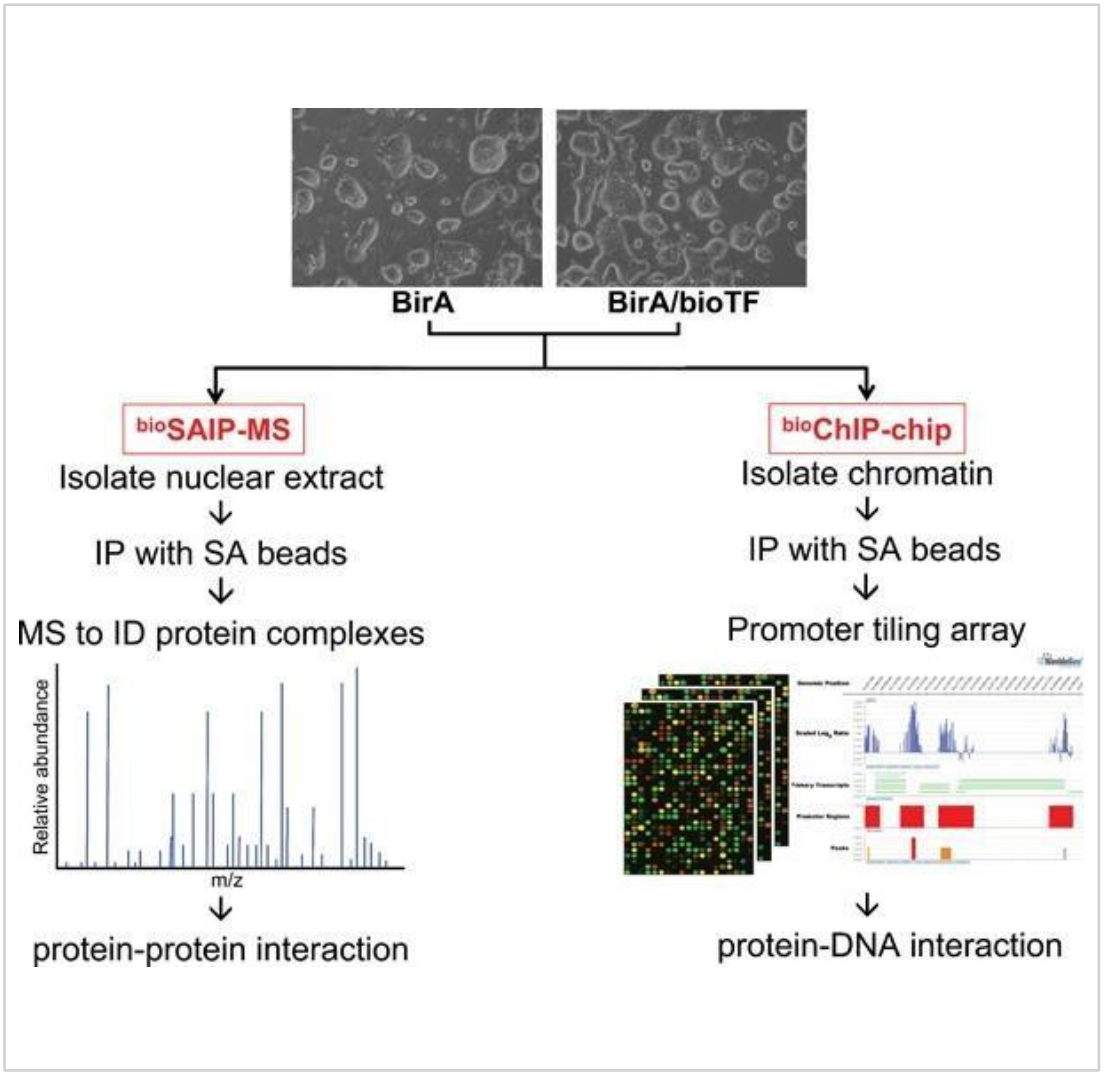
Функциональная интерпретация протеомных данных

• Списки белков превращают в функции и пути, а не оставляют “как есть”.
• Используют GO (Gene Ontology, генная онтология), KEGG и другие базы.
• Смотрят обогащение процессов, путей и модулей.
• Для сетей оценивают кластеры и их биологический смысл.
• Интерпретация зависит от качества аннотаций и выбранного фона.



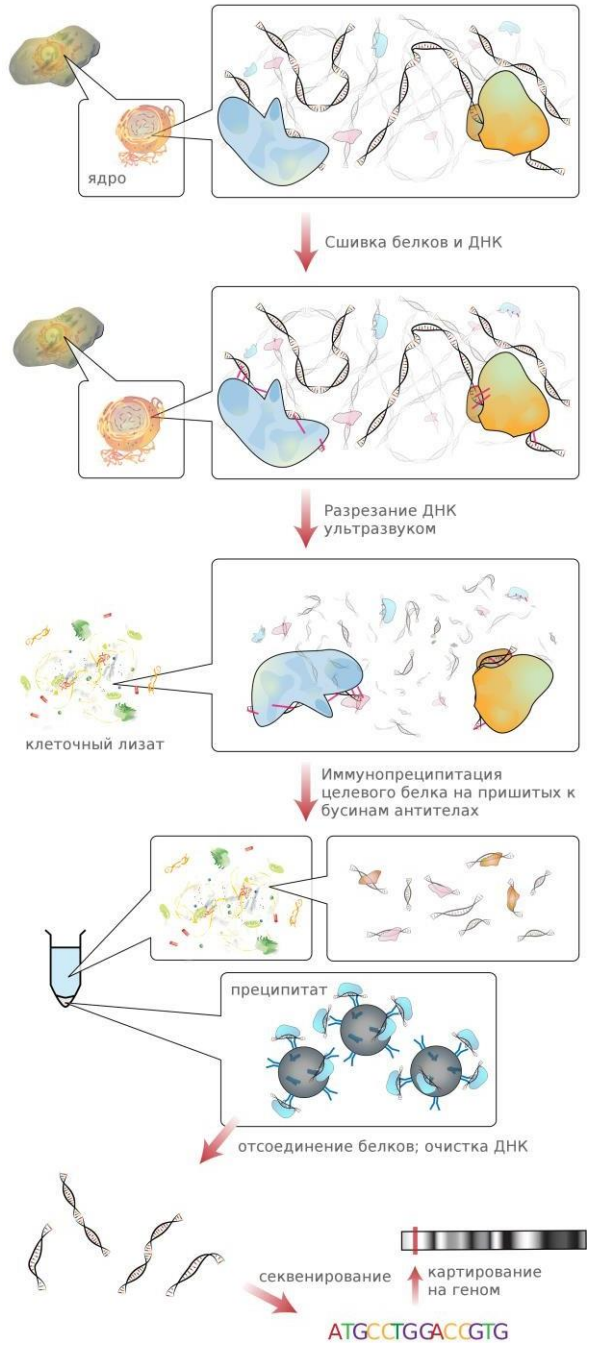
Белок-ДНК взаимодействия

• Белки регулируют транскрипцию через связывание с ДНК.
• Особенно важны TF (Transcription Factor, фактор транскрипции).
• Поиск сайтов связывания помогает понять регуляторные сети.
• Это связывает протеом регуляторов и транскриптом клеточных ответов.
• Дальше для этого используют ChIP-подходы.



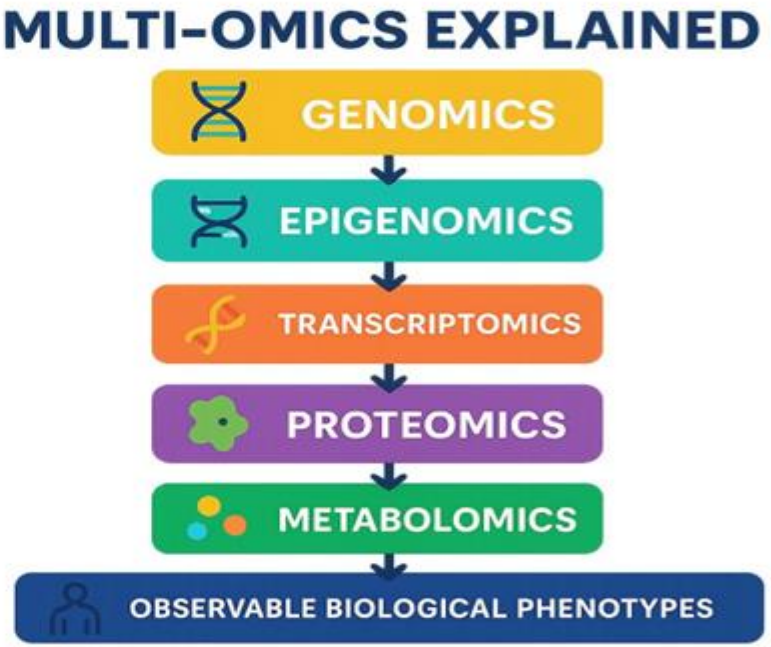
ChIP-chip и ChIP-seq

• ChIP (Chromatin Immunoprecipitation, иммунопреципитация хроматина) “ловит” участки ДНК, связанные с белком.
• ChIP-chip читает их через микрочип, ChIP-seq — через секвенирование.
• Типовые шаги: фиксация, фрагментация, иммунопреципитация, анализ ДНК.
• Выход — пики связывания, мотивы и предполагаемые мишени.
• Ключевые факторы качества: антитело, контроль и обработка данных.



Интеграция транскриптомики и протеомики

- Интеграция объединяет РНК, белки, PTM и взаимодействия в одну модель.
- Так можно строить причинные цепочки: регулятор → мишени → фенотип.
- Используют корреляции, сети, факторные модели и машинное обучение.
- Трудности: разные масштабы данных, идентификаторы и батч-эффекты.
- Но именно multi-omics даёт наиболее целостную картину системы.



Lifebit
lifebit.ai