«Анализ транскриптомных данных»

Лекция #1

NGS и RNA-Seq

Серёжа Исаев

аспирант **MedUni Vienna**

Преподаватели курса

Серёжа Исаев

аспирант **MedUni Vienna**
выпускник **ФББ МГУ**

tg: `@sergisa`

Даня Бобровский

студент **EPFL**
студент **ФББ МГУ**

tg: `@daniilbobrovskiy`

Домашние задания и зачёт

— Кто может получить зачёт в зачётку?
— Студенты четвёртого и пятого курса ФББ МГУ. Остальные могут получить сертификат о прохождении курса.

— Нужно ли ходить на курс для зачёта?
— *Нет, но зачем тогда вам этот курс?*

— Что нужно для зачёта?
— *Есть две опции:*

- *выполнение более чем 12 домашних заданий на оценку как минимум “удовлетворительно”,*
- *выполнение двух проектов по каждому из блоков (bulk RNA-Seq и scRNA-Seq) — это будет обговорено дополнительно*

Домашние задания и зачёт

— Что такое оценка “удовлетворительно”?

— Это значит, что я посчитал, что вы выполнили задание верно как минимум концептуально.

— Как мне надо будет сдавать задания?

— Вся информация будет потом, но главное то, что все домашние задания должны быть выложены в вашем личном GitHub-репозитории.

— Как получить фидбэк по домашке?

— По всей видимости, почти никак, потому что “вас много, а я один”, проверять я буду домашние задания только в конце и только у тех, кто хотя бы формально может претендовать на зачёт.

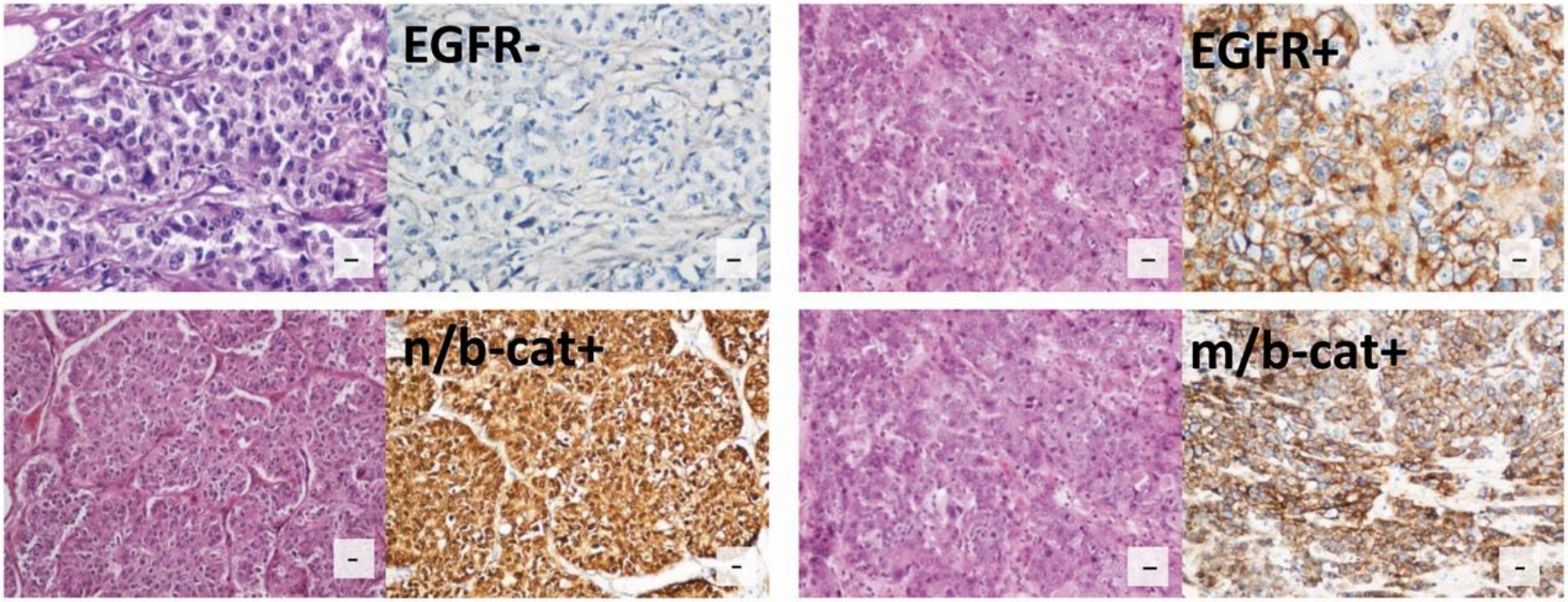
Содержание курса

1. Bulk RNA-Seq:
    a. **экспериментальные подходы,**
    b. выравнивания и псевдовыравнивания,
    c. анализ дифференциальной экспрессии,
    d. функциональный анализ;

1. Single-cell RNA-Seq:
    b. экспериментальные подходы,
    c. отличия от процессинга bulk RNA-Seq,
    d. методы снижения размерности,
    e. кластера и траектории,
    f. мультимодальные омики одиночных клеток.

Зачем мы изучаем РНК?

Иммуногистохимические окрашивания являются одной из основной оптик изучения гетерогенности тканей



Lakis et al., **Anticancer Res** , 2016

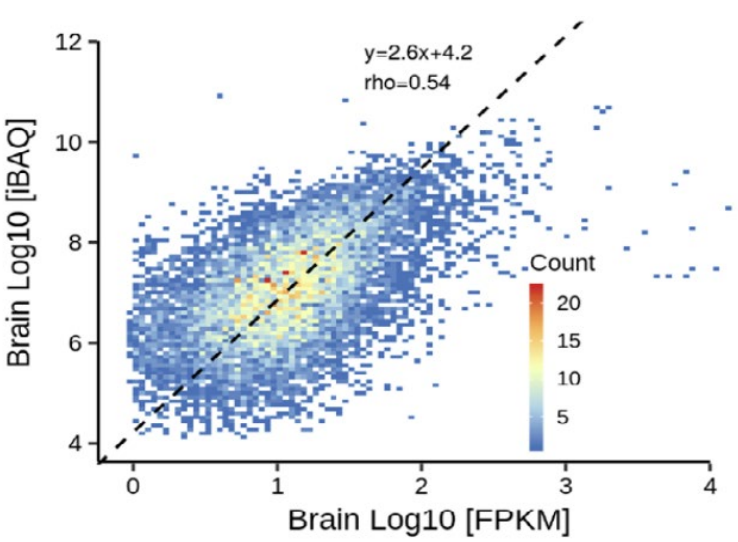
Зачем мы изучаем РНК?

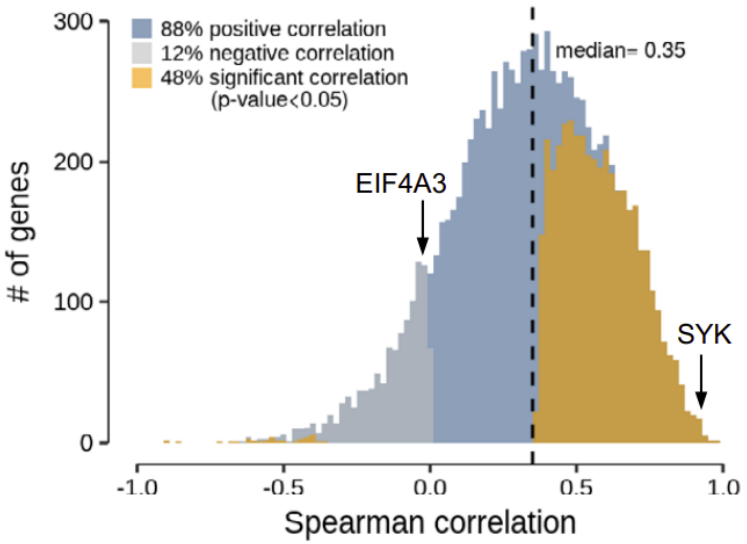
РНК ≈ белки ≈ фенотип

Для того, чтобы понять фенотип ткани, нам важно понимать, сколько и каких белков там было

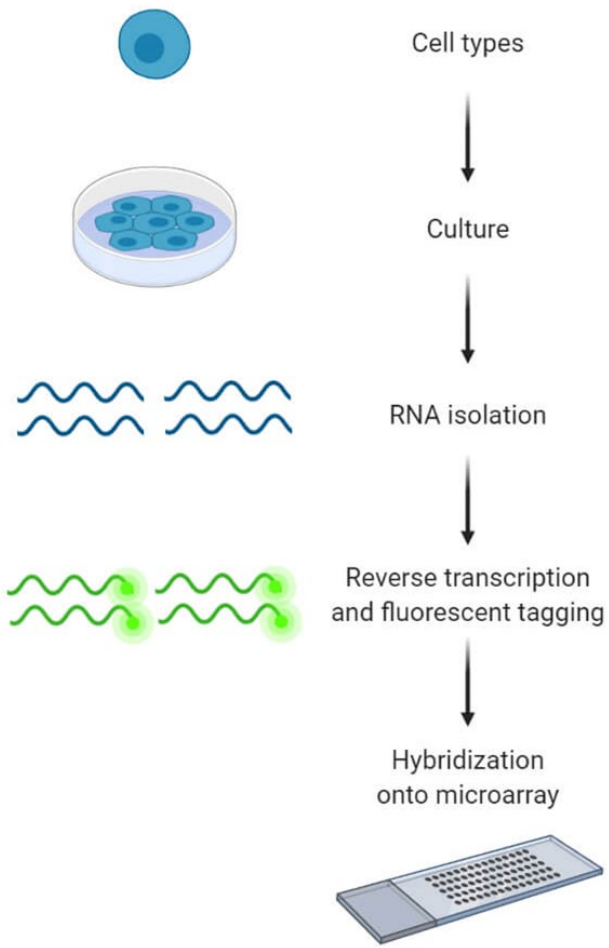
Работать с белками напрямую сложно и дорого, поэтому работают с РНК

Wang et al., **Mol Syst Biol** , 2019



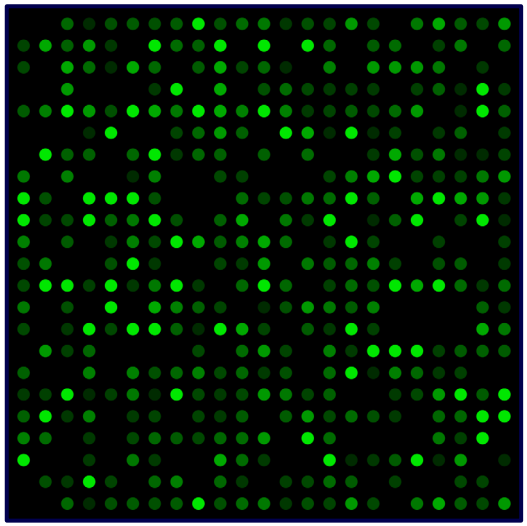


Одноканальные микрочипы

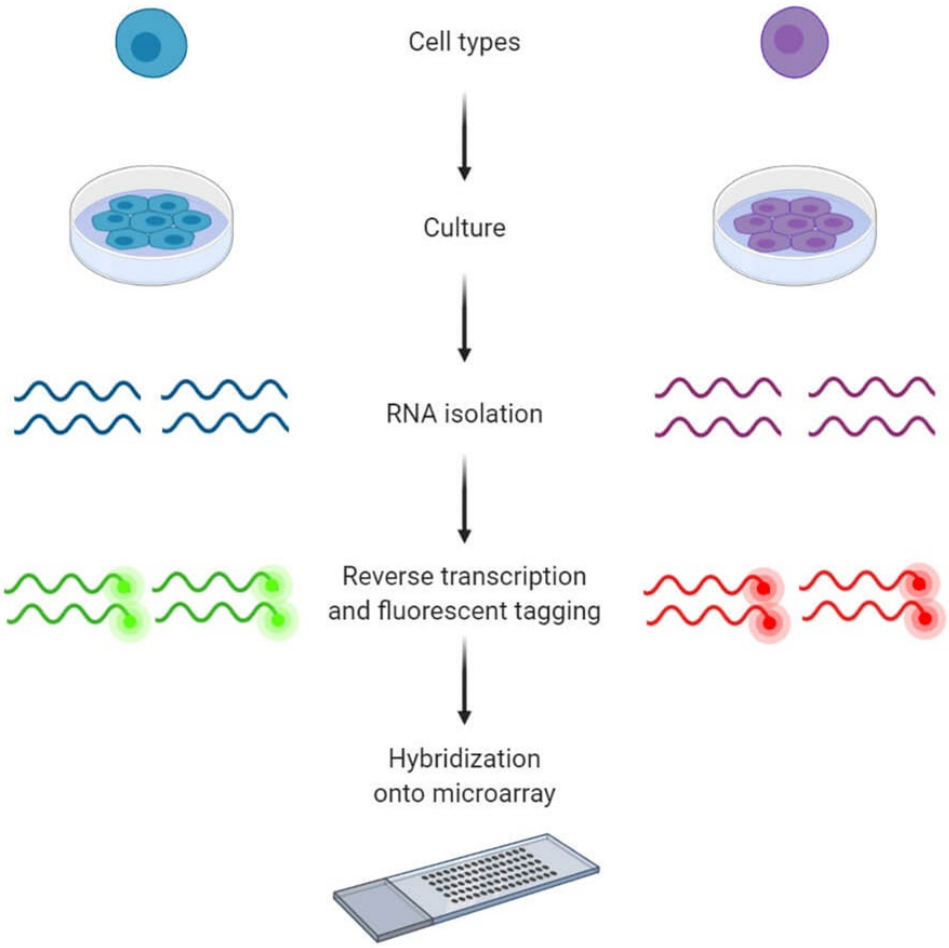


Каждое пятно (спот) содержит зонды для определённого гена

Интенсивность и цвет сигнала отражают количество транскриптов определённого гена в образце

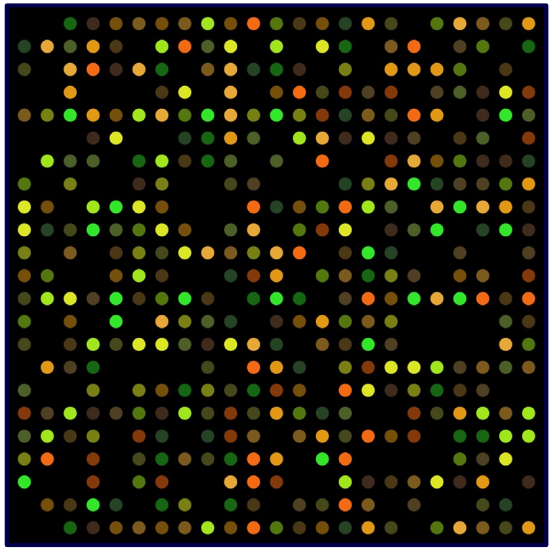


Двуканальные микрочипы



Гибридизуются два образца, каждый со своим красителем

Образцы можно сразу же сравнить



Анализ микрочиповых данных

В этом курсе мы не затрагиваем тонкости анализа микрочипов

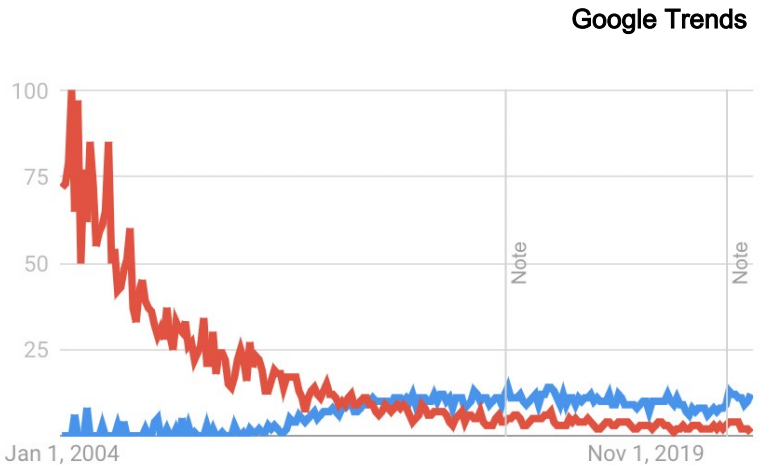
В этом анализе есть много неочевидных тонкостей, и подходы, работающие для работы с RNA-Seq, могут быть абсолютно нерелевантными для микрочипов

В зависимости от платформы чипа (Affymetrix, Illumina, …) подходы к анализу могут слегка различаться

Плюсы и минусы микрочипов

Плюсы микрочипов :

+ Относительно дешёвый метод
+ Обработка и анализ в целом проще, чем в случае с RNA-Seq



Минусы микрочипов :

— Нужны априорные знания о последовательности
— Нужны достаточно большие концентрации исходной РНК (~ 1μg)
— Нельзя исследовать новые события сплайсинга или какие-то трансформации, связанные с последовательностями

Секвенирование

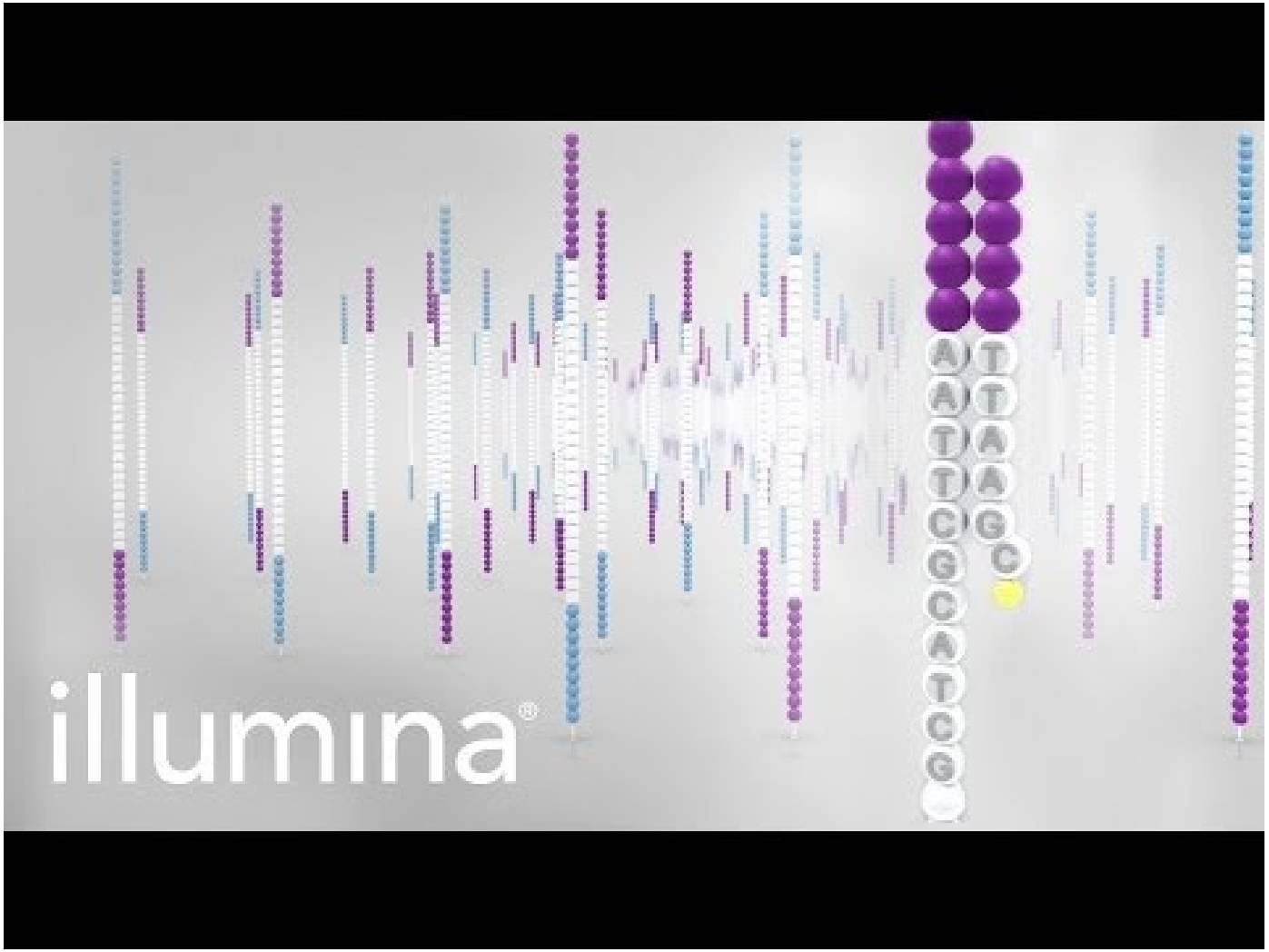
Секвенирование — это метод определения нерегулярного гетерополимера (чаще всего говорят про секвенирование нуклеиновых кислот, однако это относится также и к белкам)

Секвенирование “предыдущего” поколения : можем за раз прочитать очень мало последовательностей

Секвенирование “следующего” поколения (NGS) : можем за раз прочитать сразу очень много последовательностей

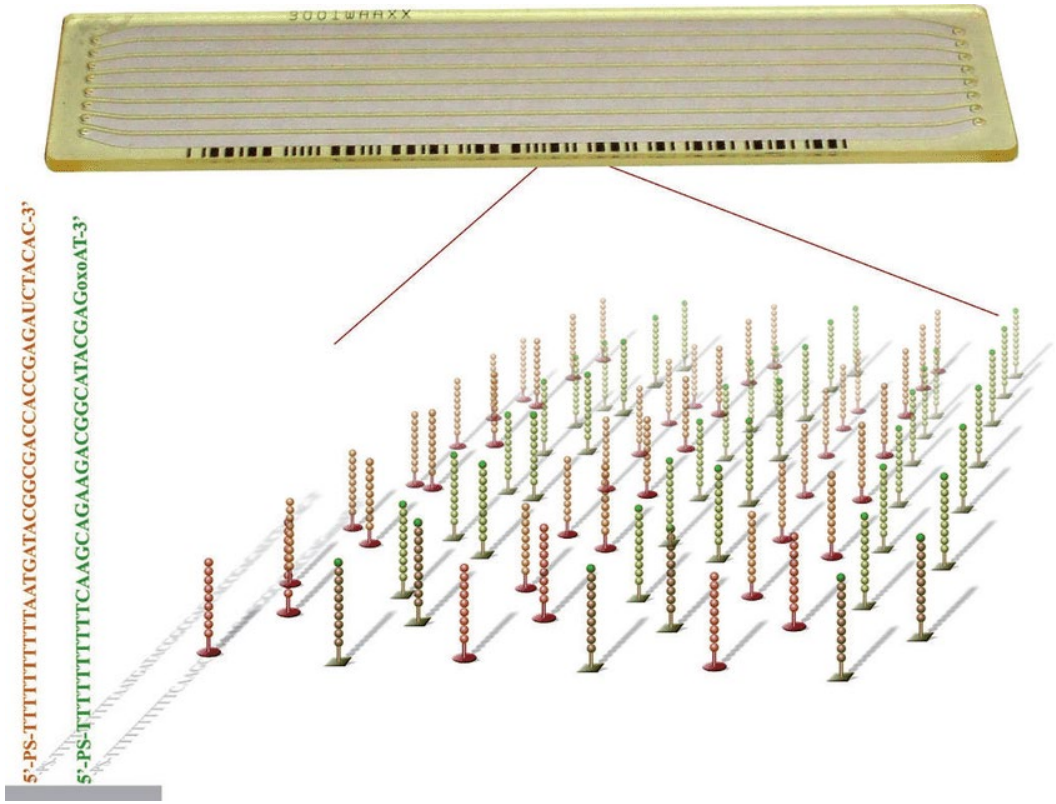
Те фрагменты ДНК, которые непосредственно загружают в секвенатор, называют библиотекой

Принцип работы Illumina



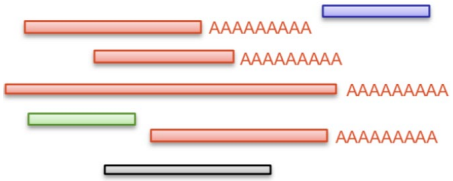
Lane на проточной ячейке

Каждая проточная ячейка состоит из восьми дорожек, с каждой дорожки будет генерироваться свой .fastq-файл



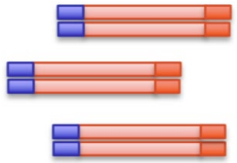
Дорожная карта подготовки библиотеки

Выделение РНК из образца

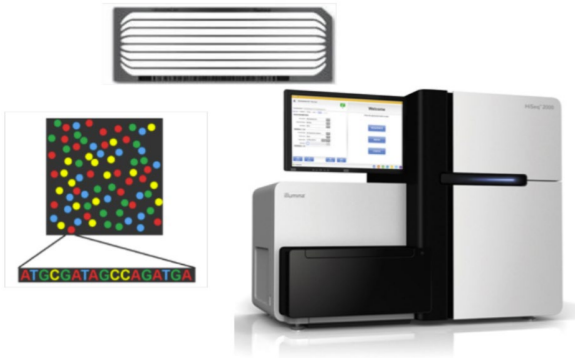


Отбор транскриптов

Синтез кДНК, создание библиотеки



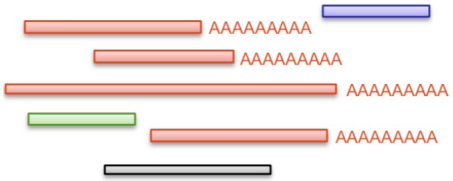
Секвенирование



Анализ RNA -Seq

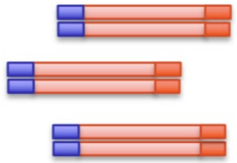
Дорожная карта подготовки библиотеки

Выделение РНК из образца

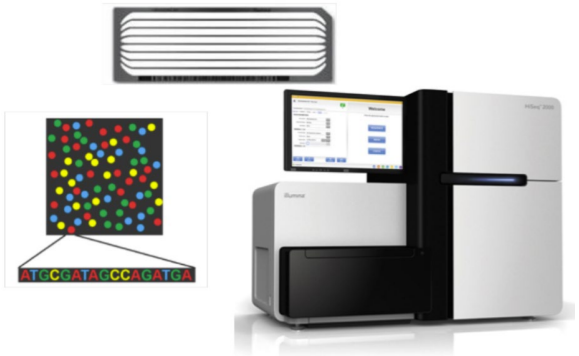


Отбор транскриптов

Синтез кДНК, создание библиотеки

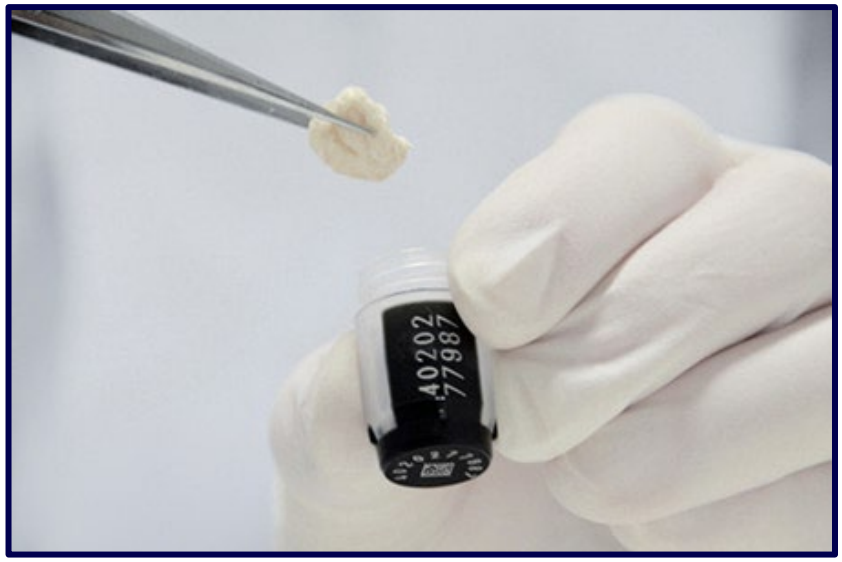


Секвенирование



Анализ RNA -Seq

Хранение образца



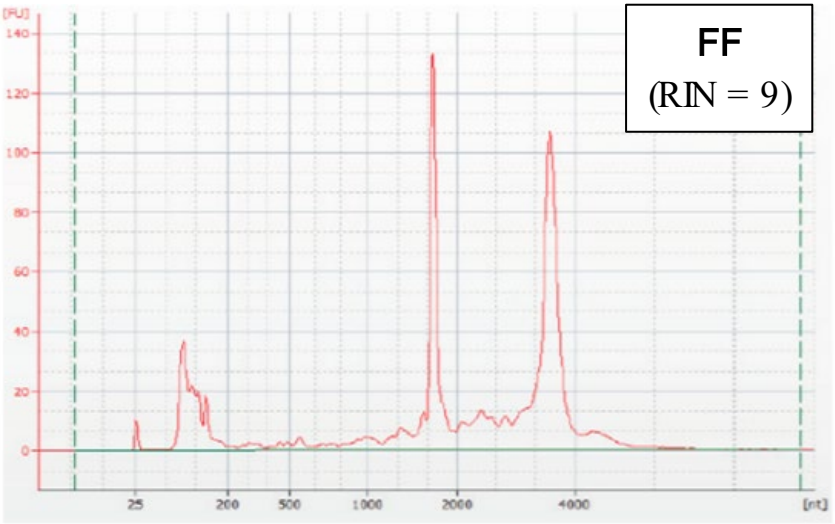
Заморозка в жидком азоте
(**FF** — fresh frozen)

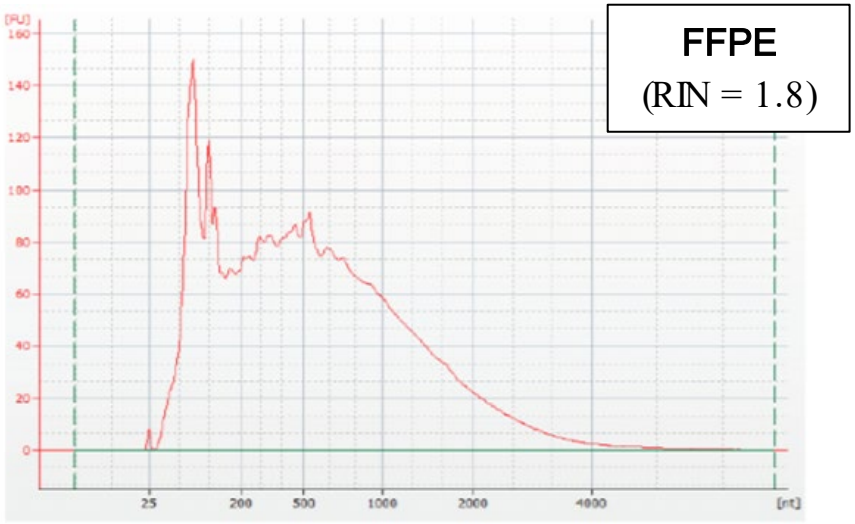
Формалин и парафин
(**FFPE** — formalin-fixed, paraffin-embedded)

RIN

RIN — это метрика, которая по различным параметрам хроматограммы определяет степень деградации РНК в образце

Обычно RIN у FFPE образцов заметно хуже, чем у FF





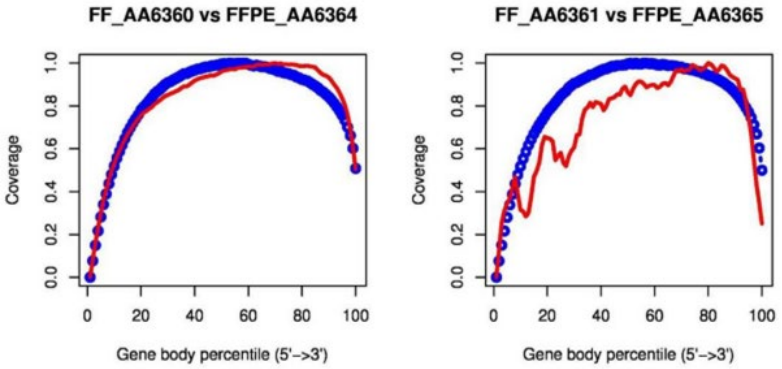
thermofisher.com

Аналоги RIN

Не всегда есть возможность посчитать RIN на стадии подготовки образцов. В этом случае можно воспользоваться косвенными метриками качества РНК в образце

Метрика **Gene Body Coverage** из пакета RSeQC отражает неравномерность покрытия генов ридами

Метрика **TIN (Transcript Integrity Number)** , тоже из RSeQC, является некоторым аналогом RIN ($1 \leq RIN \leq 10$, $1 \leq TIN \leq 100$)

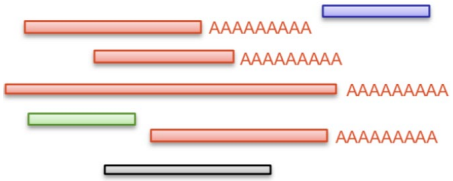


	TIN
	median
FF_AA6360	72
FFPE_AA6364	23
FF_AA6361	72
FFPE_AA6365	1

Esteve-Codina et al., **PLoS ONE**, 2017

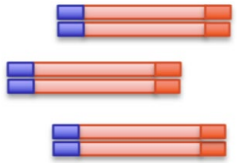
Дорожная карта подготовки библиотеки

Выделение РНК из образца

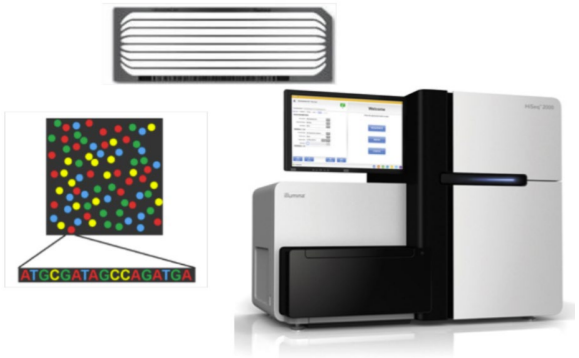


Отбор транскриптов

Синтез кДНК, создание библиотеки



Секвенирование

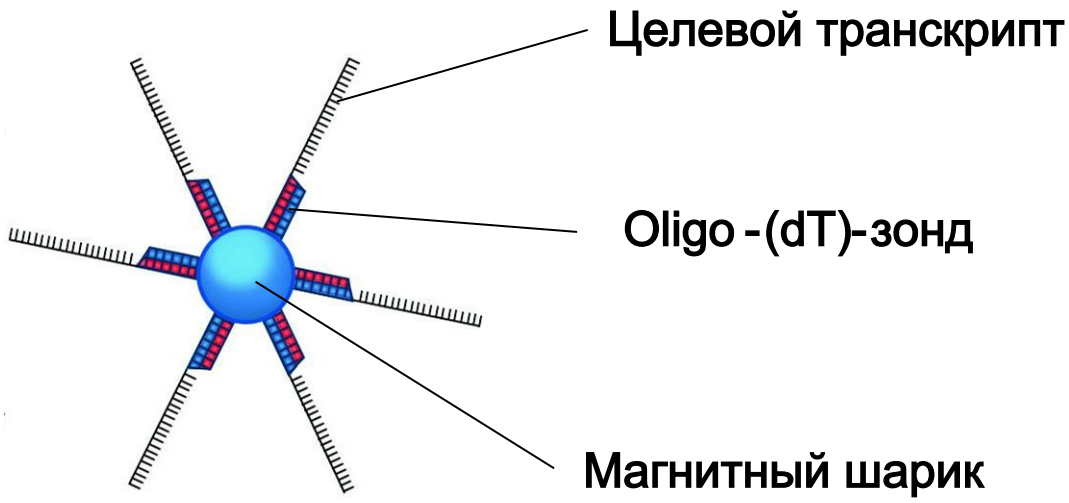


Анализ RNA -Seq

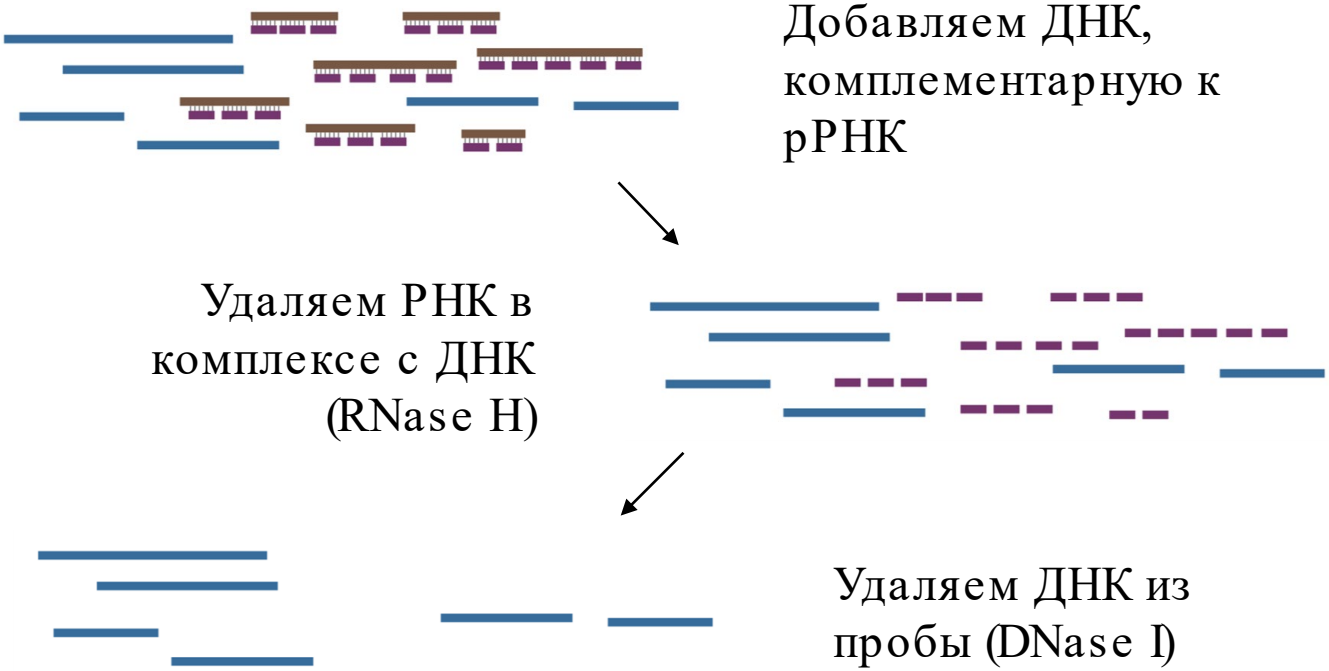
Oligo-(dT)-гибридизация (= polyA)

Отбираются только полиаденилированные транскрипты

Всё равно остаётся небольшая примесь рРНК и других нкРНК

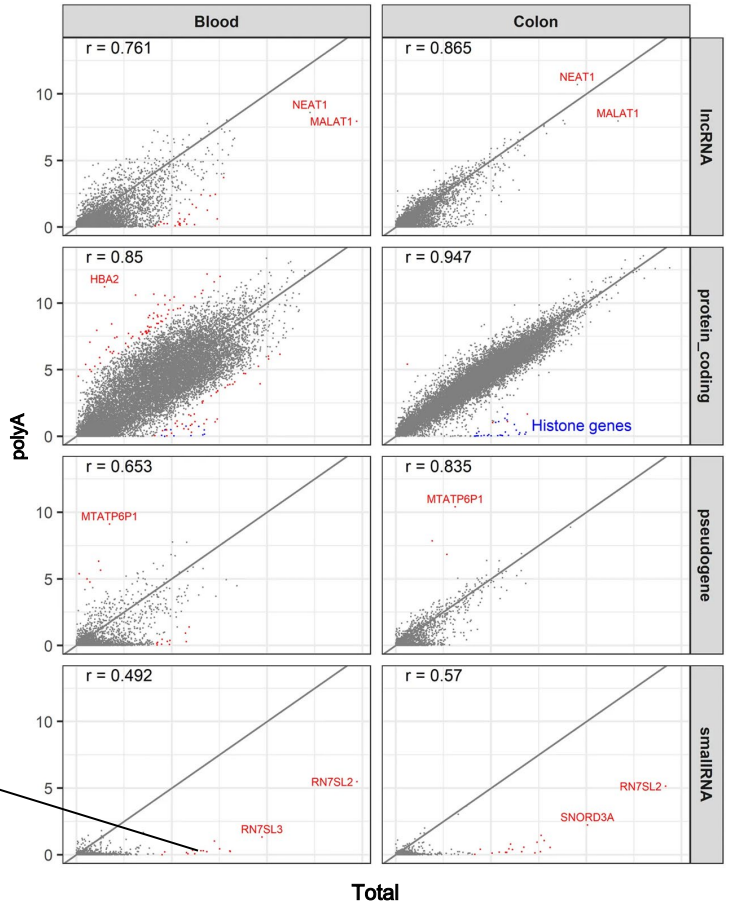


Деплеция рРНК (= total)

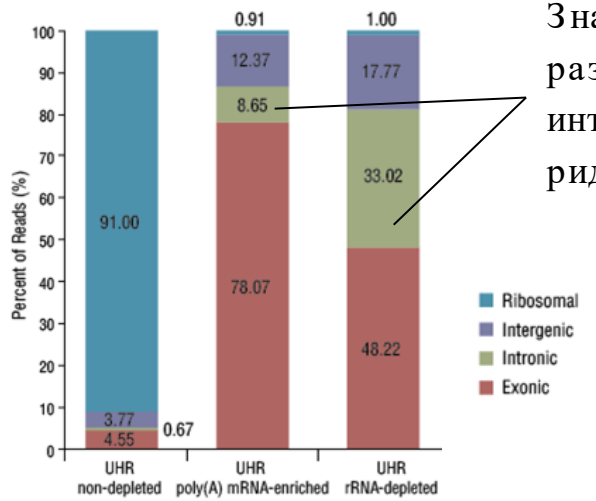


Метод позволяет исследовать экспрессию неполиаденилированных транскриптов

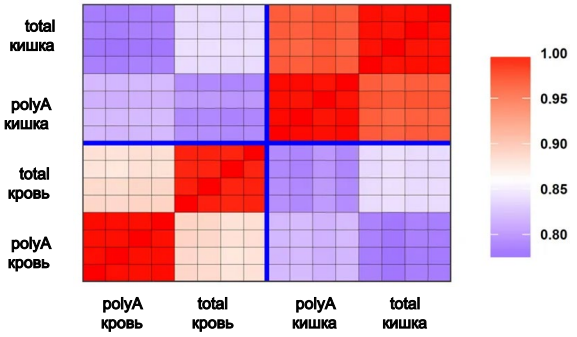
Сходимость методов



Значительная разница в оценке экспрессии малых РНК



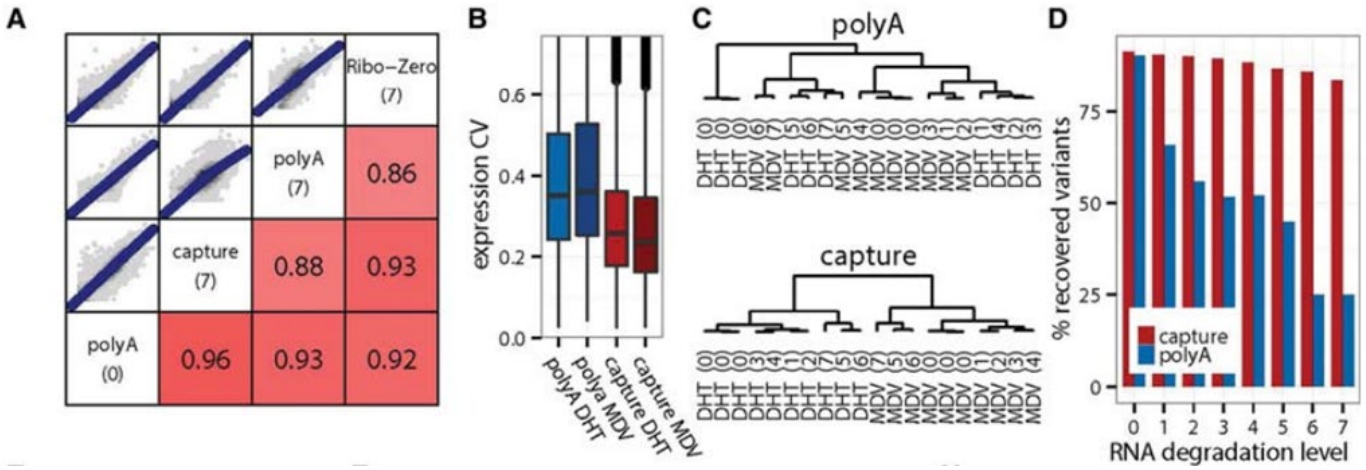
Значительная разница в интронных ридах



Zhao et al., **Sci Rep**, 2018

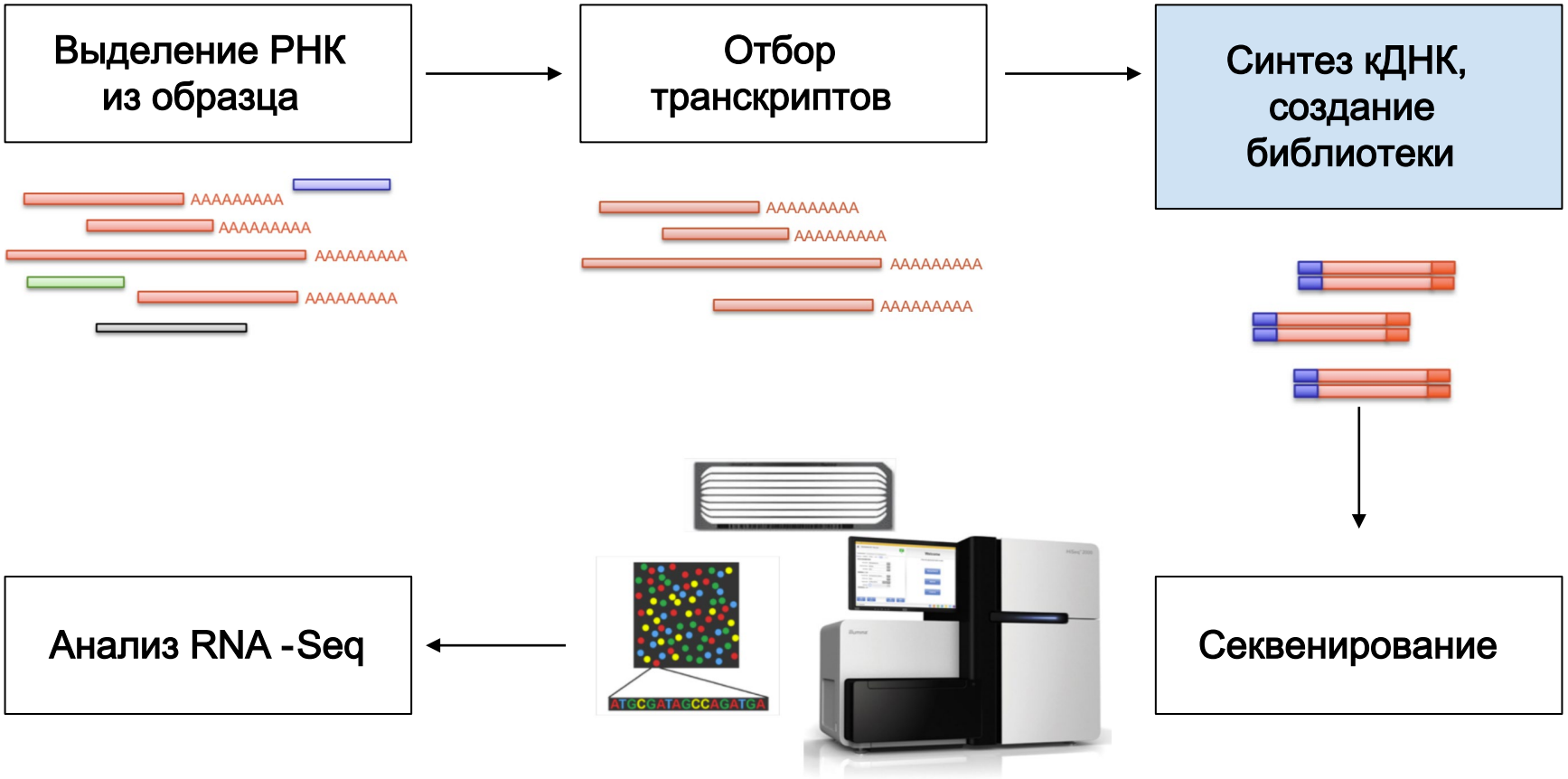
Exome capture

При работе с FFPE образцами низкого качества обычно выделяют необходимые транскрипты при помощи зондов на экзом

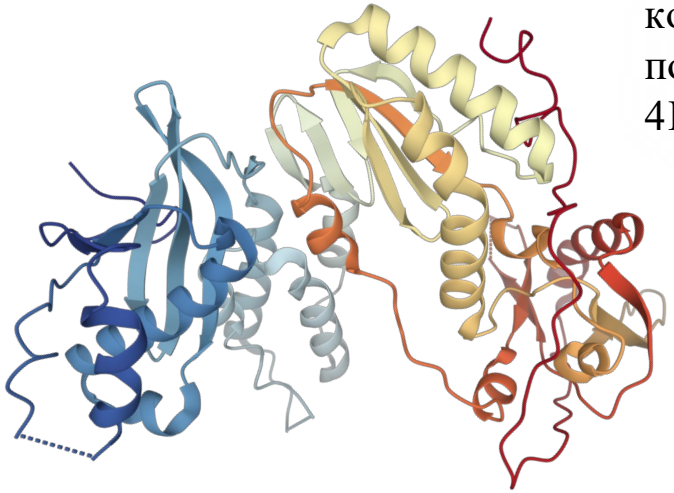


Marcin et al., **Genome Res** , 2015

Дорожная карта подготовки библиотеки



Обратная транскрипция



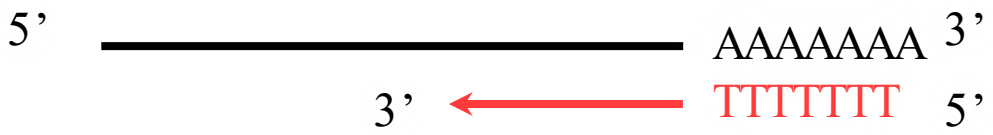
Структура RT из MMLV, которую используют при подготовке кДНК (PDB ID: 4MH8)

Перед секвенированием необходимо синтезировать кДНК на матрице ДНК при помощи **обратной транскриптазы** (RT)

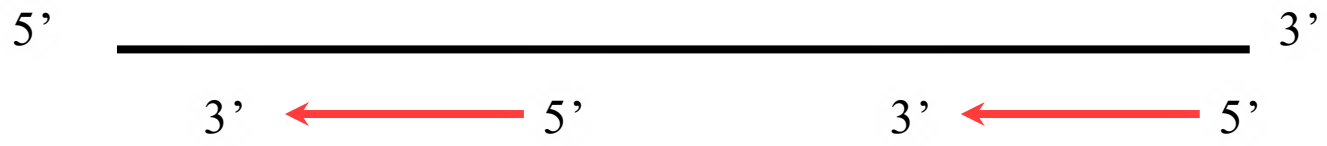
Для того, чтобы обратная транскриптаза начала синтез ДНК, ей необходим праймер

Праймеры для обратной транскрипции

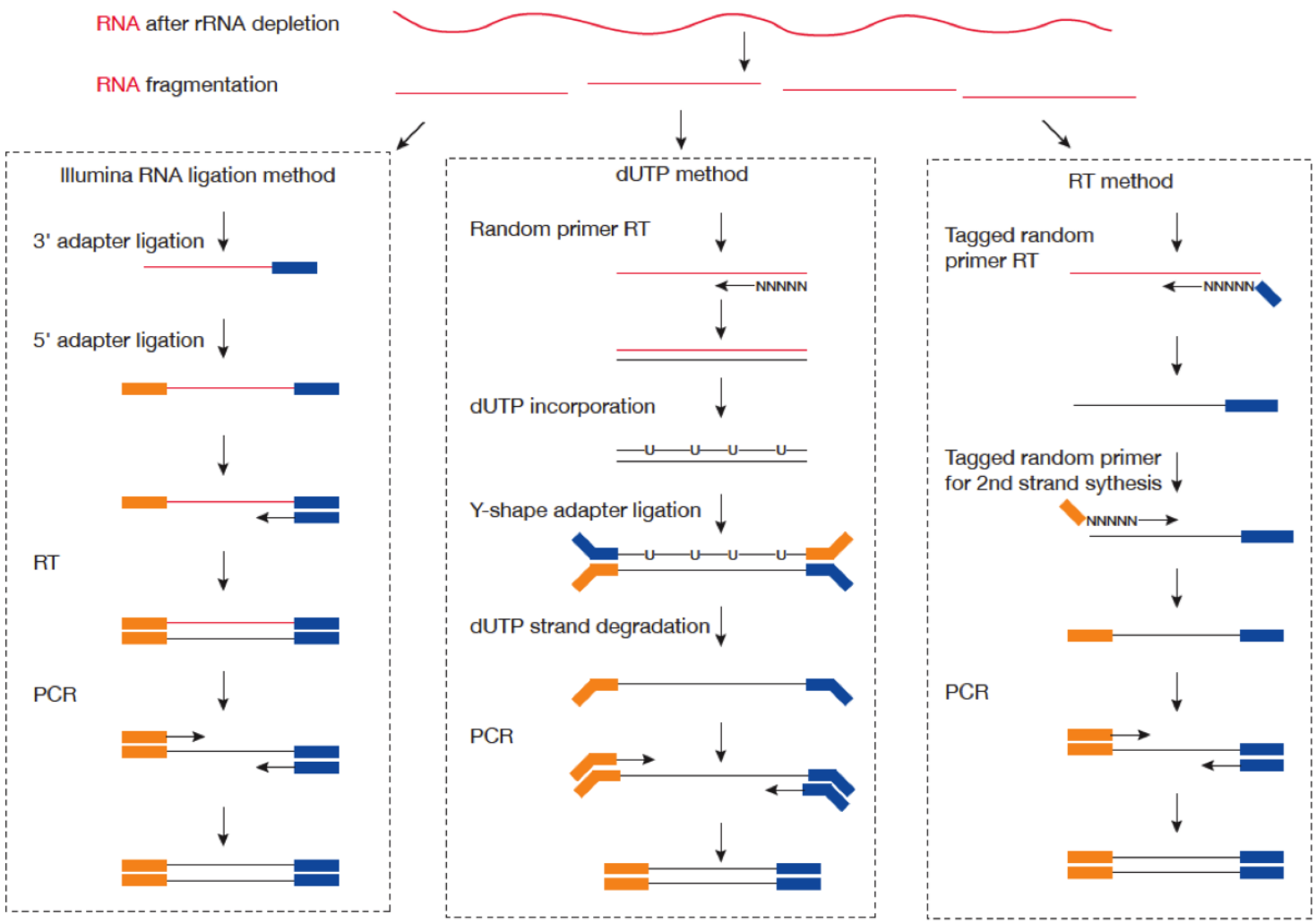
В качестве праймера можно использовать **poly-A** от мРНК



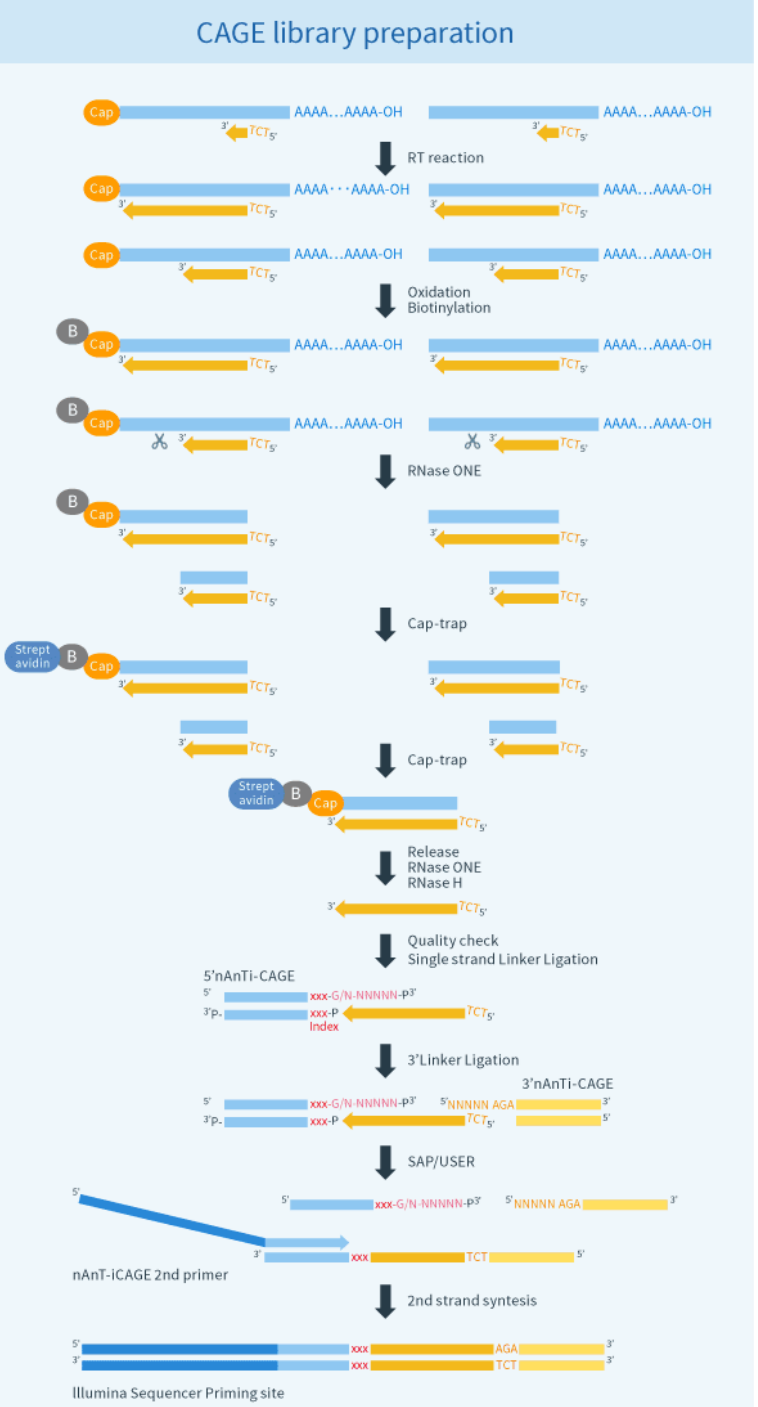
В основном (почти всегда) используют **случайные праймеры** . Если в описании метода указано, что он сделан при помощи polyA, чаще всего речь идёт о методе селекции транскриптов



Stranded RNA -Seq



biostars.org



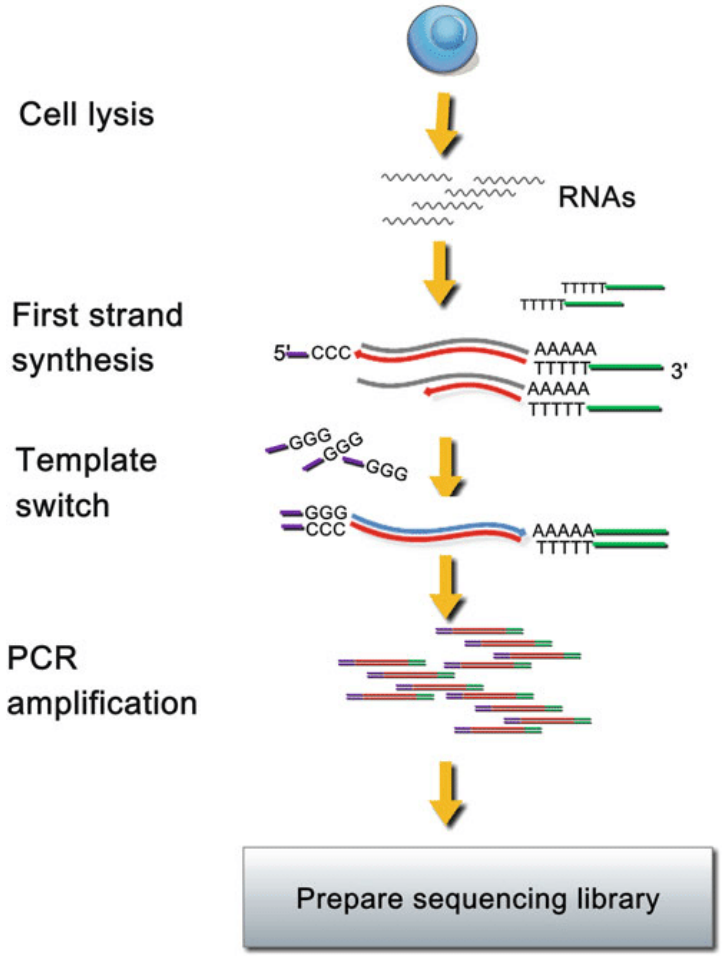
CAGE

При помощи подготовки библиотек методом CAGE приходит обогащение на те участки транскриптов, которые расположены ближе к 5’-концу

Это позволяет увеличить покрытие 5’-конца, а также задетектировать многие **малые регуляторные** РНК (в результате экспрессии энхансеров и т. п.)

cage-seq.com

SMART-Seq



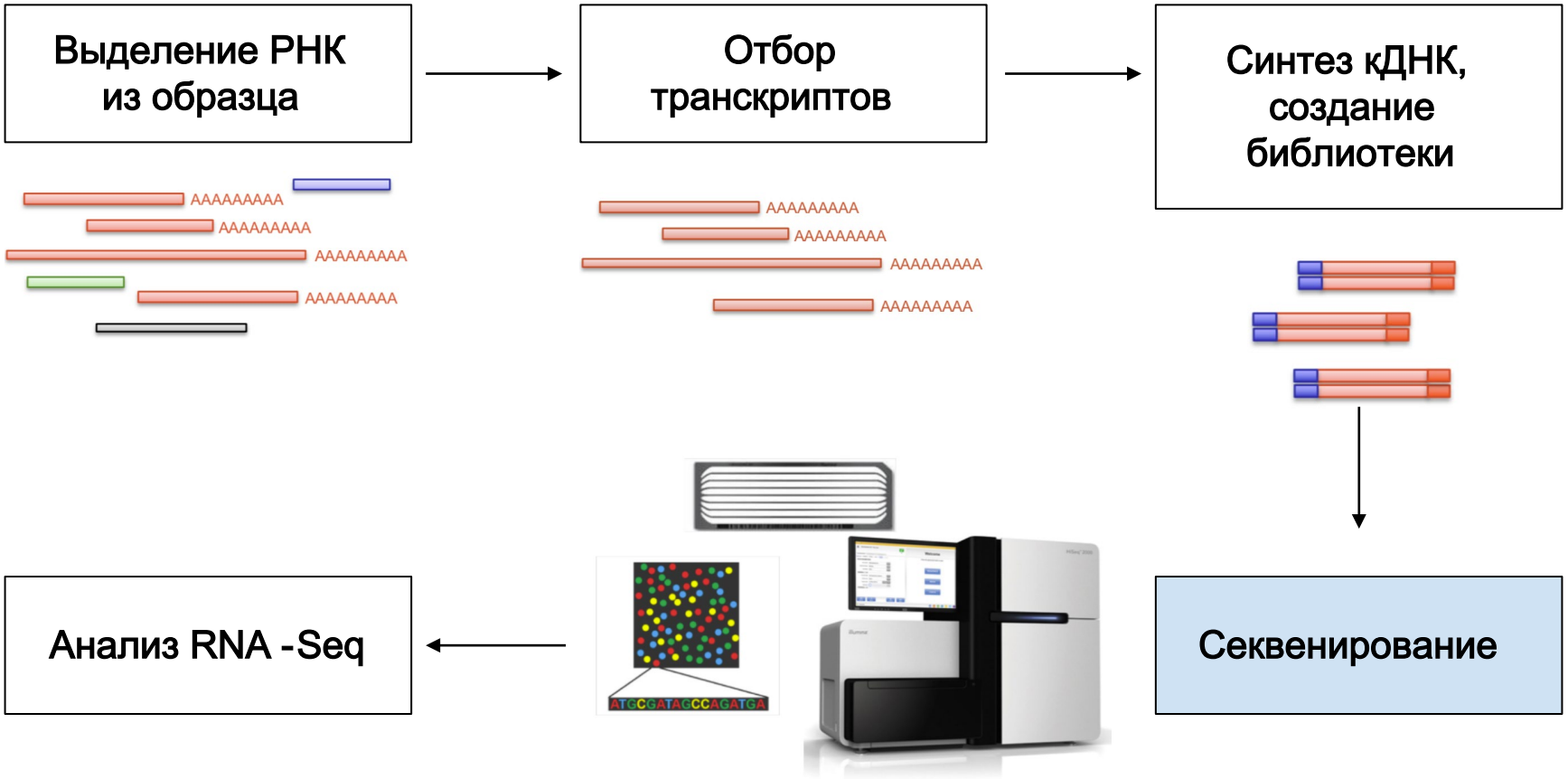
Chen et al., **Single Cell Sequencing and Systems Immunology** , 2015

Метод основан на том, что обратная транскриптаза некоторых вирусов после кэпа всегда добавляет три цитозина на строящейся цепи РНК

К этим трём цитозинам может гибридизоваться TSO

В результате мы можем во время процедуры ПЦР в качестве одного из праймеров использовать последовательность в TSO — в итоге амплифицированными будут только полноразмерные РНК

Дорожная карта подготовки библиотеки



Illumina

Sequencing System	iSeq™	MiniSeq™	MiSeq®	NextSeq®	HiSeq®	HiSeq® X	NovaSeq®
					4000	Five/Ten	6000
Output per run	1.2 Gb	7.5 Gb	15 Gb	120 Gb	1.5 Tb	1.8 Tb	1 Tb - 6 Tb[1]
Instrument price	$19.9K	$49.5K	$99K	$275K	$900K	$6M[2]/$10M[2]	$985K
Installed base[3]	NA	~600	~6,000	~2,400	~2,300[4]		~285

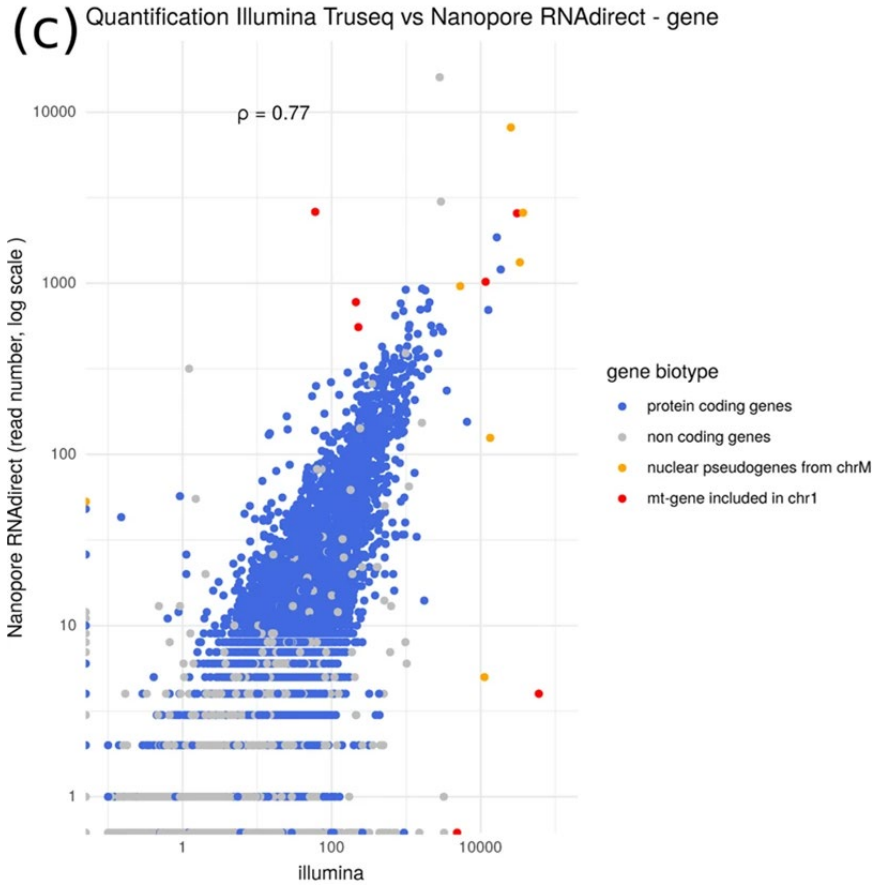
В индустрии обычно используют NovaSeq (2 x 150 bp), себестоимость секвенирования на нём самая низкая

Oxford Nanopore vs. Illumina

В последнее время появляется всё больше работ с секвенированием РНК при помощи ONT

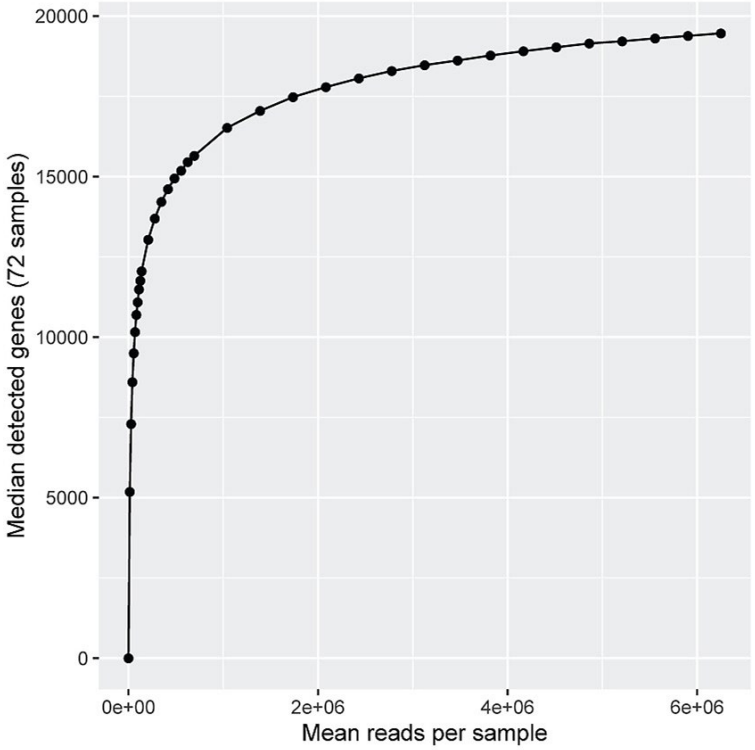
ONT подходит в первую очередь для анализа альтернативного сплайсинга

Также ONT позволяет работать с РНК напрямую (и, соответственно, с модификациями РНК)



Sessegolo et al., **Sci Rep**, 2019

Достаточно ли я отсеквенировал?

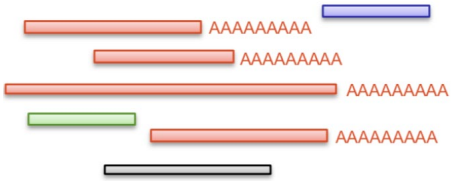


Иногда неясно (особенно при экспериментах scRNA-Seq, но не только), является ли отсутствие экспрессии какого-то гена реальным отсутствием, или же мы просто недостаточно отсеквенировали библиотеку

Ответ на этот вопрос можно найти, построив saturation plot

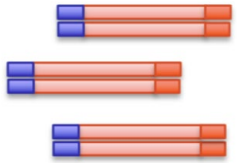
Дорожная карта подготовки библиотеки

Выделение РНК из образца

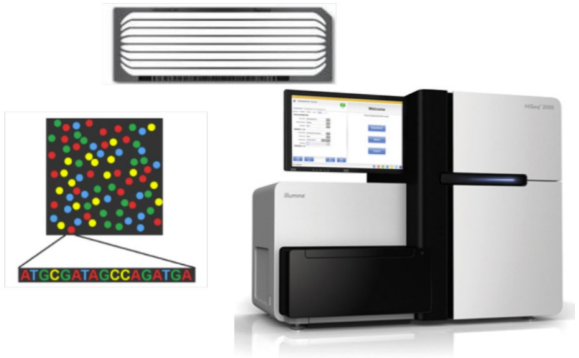


Отбор транскриптов

Синтез кДНК, создание библиотеки



Секвенирование



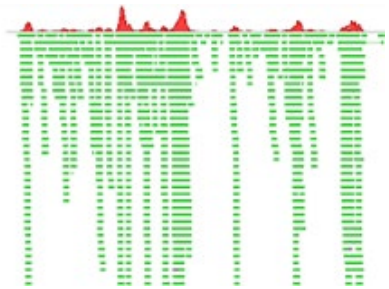
Анализ RNA -Seq

Дорожная карта анализа RNA -Seq

QC прочтений

FastQC Report

Выравнивание

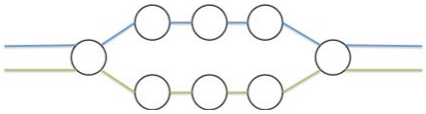


Подсчёт экспрессий

	Wild-type	
	Sample 1	Sample 2
Gene 1	24	31
Gene 2	0	3
Gene 3	1988	1125
Gene 4	5	0
...	...	...
Total	22341961	20739175

Нормализация

Псевдовыравнивание



Дифференциальная экспрессия

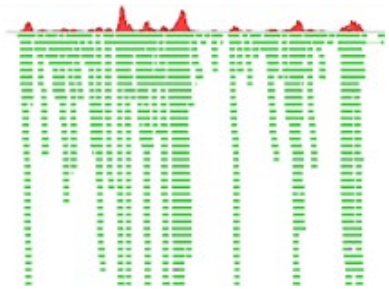
GO, GSEA, ssGSEA и прочее

Дорожная карта анализа RNA -Seq

QC прочтений

FastQC Report

Выравнивание

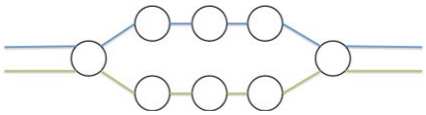


Подсчёт экспрессий

	Wild-type	
	Sample 1	Sample 2
Gene 1	24	31
Gene 2	0	3
Gene 3	1988	1125
Gene 4	5	0
...	...	...
Total	22341961	20739175

Нормализация

Псевдовыравнивание

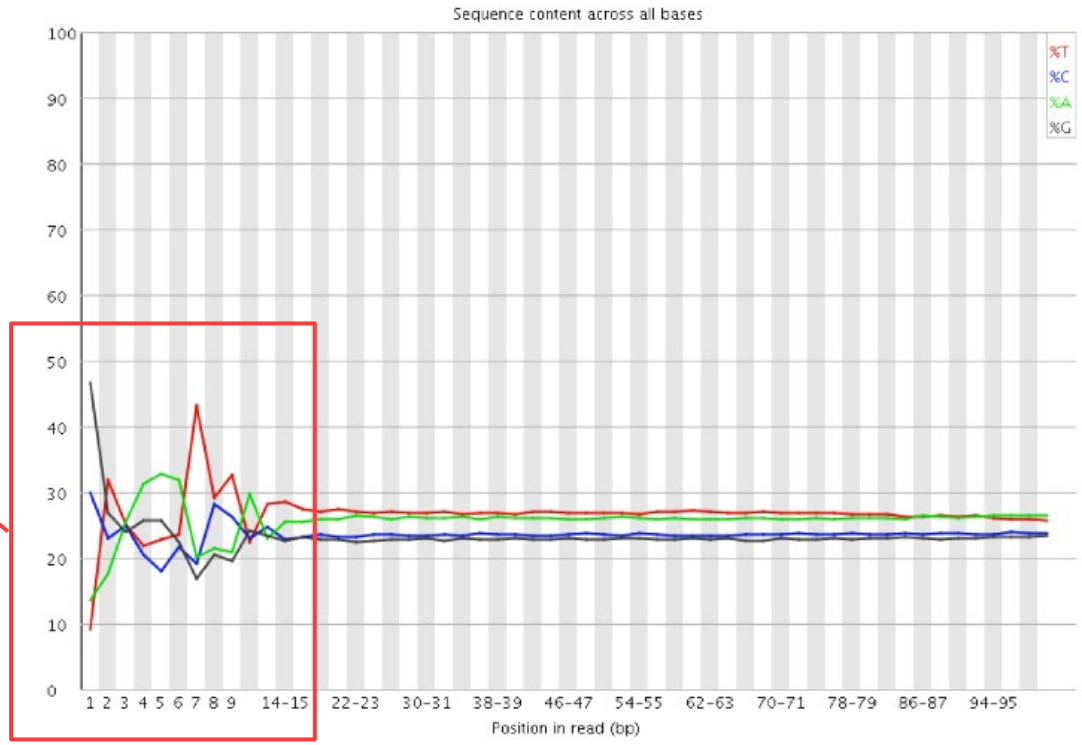


Дифференциальная экспрессия

GO, GSEA, ssGSEA и прочее

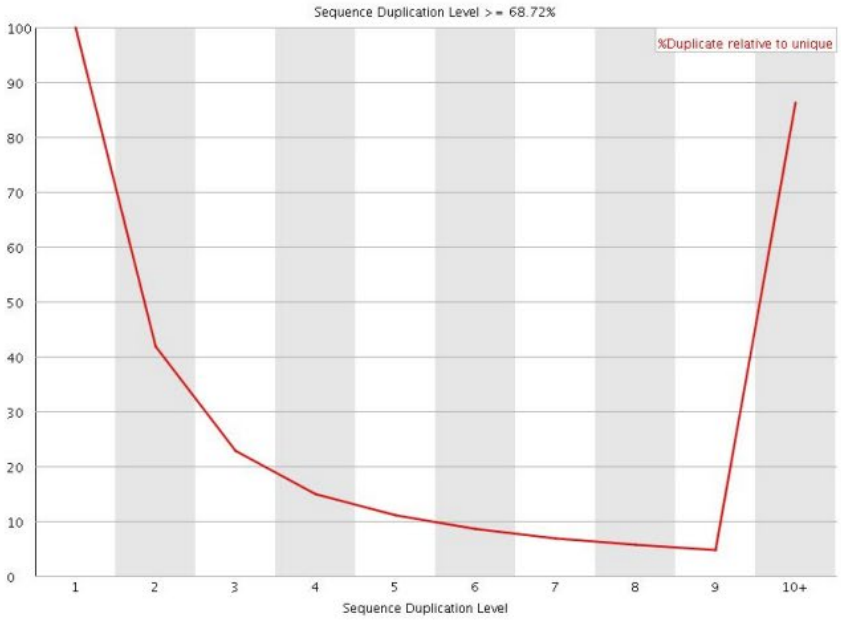
Случайные праймеры на QC

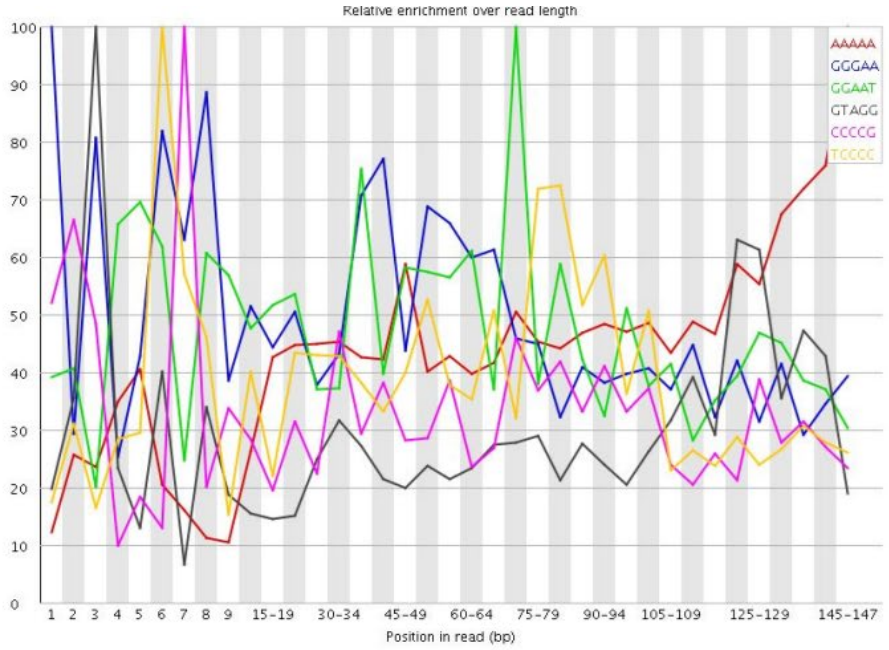
“Случайные” праймеры не такие уж и случайные



рРНК / мтДНК

Наличие рРНК в образце можно увидеть по высокому числу повторяющихся прочтений





Как заметить контаминацию?

Образец, скорее всего, контаминирован, если есть несколько (два и больше) мод на распределении GC-состава прочтений