HSE University
Faculty of Computer science

«Анализ данных NGS»

Лекция #4

Функциональный анализ bulk RNA-Seq

Серёжа Исаев

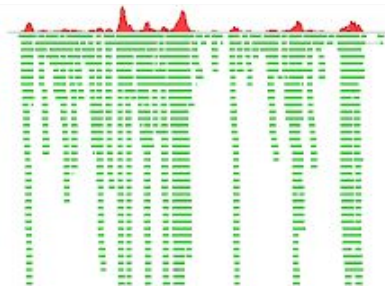
аспирант **MedUni Vienna**

Дорожная карта анализа RNA-Seq

QC прочтений

FastQC Report

Выравнивание



Подсчёт экспрессий

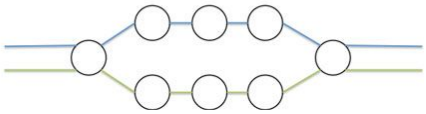
	Wild-type	
	Sample 1	Sample 2
Gene 1	24	31
Gene 2	0	3
Gene 3	1988	1125
Gene 4	5	0
...	...	...
Total	22341961	20739175

Нормализация

Дифференциальная экспрессия

GO, GSEA, ssGSEA и прочее

Псевдовыравнивание



Как работает tximport на аутпуте kallisto?

1. Для образца имеется оценённое (**нецелочисленное**) число каунтов для каждого из транскриптов
2. Каждый из транскриптов является каким-то геном (например, *ген1 = транскрипт1 + транскрипт2 + транскрипт3*)
3. **Сложим все транскрипты**, соответствующие одному гену, и просто **округлим** полученное значение каунтов — это будет число каунтов этого гена
4. Вспомним, что такое оффсет

$$K_{i,j} \sim NB(\mu_{i,j}, \alpha_i)$$
$$\mu_{i,j} \sim s_j p_{i,j}$$
$$\log_2(p_{i,j}) = x_{j,A}\beta_{j,A} + x_{j,B}\beta_{j,B}$$

Как работает tximport на аутпуте kallisto?

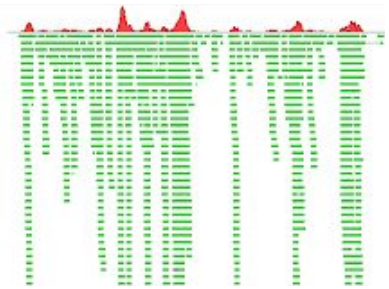
1. Для образца имеется оценённое (**нецелочисленное**) число каунтов для каждого из транскриптов
2. Каждый из транскриптов является каким-то геном (например, *ген1 = транскрипт1 + транскрипт2 + транскрипт3*)
3. **Сложим все транскрипты**, соответствующие одному гену, и просто **округлим** полученное значение каунтов — это будет число каунтов этого гена
4. Кроме поправки на размер библиотеки мы можем ввести дополнительную поправку, прямо пропорциональную средневзвешенной сумме транскриптов (длина1 * каунты_транскрипта1 + …) / (сумма каунтов транскрипта)
5. Таким образом, использование более длинных изоформ будет приводить к тому, что p будет меньше при том же числе каунтов

Дорожная карта анализа RNA-Seq

QC прочтений

FastQC Report

Выравнивание

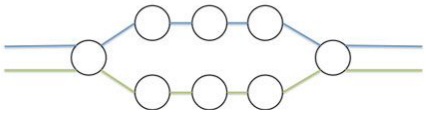


Подсчёт экспрессий

	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



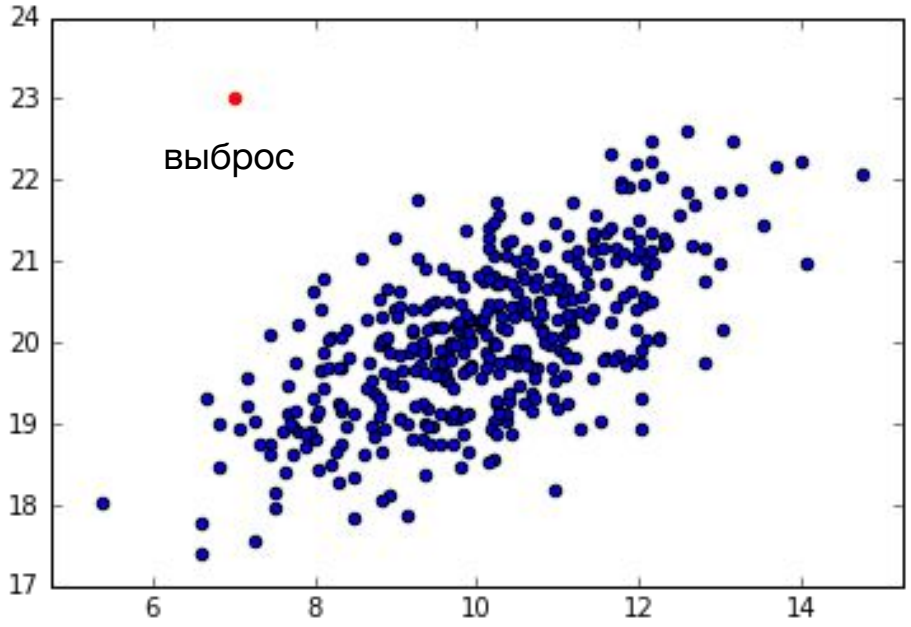
Дифференциальная экспрессия

GO, GSEA, ssGSEA и прочее

Проверка самосогласованности данных

Для того, чтобы проверить гомогенность данных, можно использовать PCA (метод главных компонент)

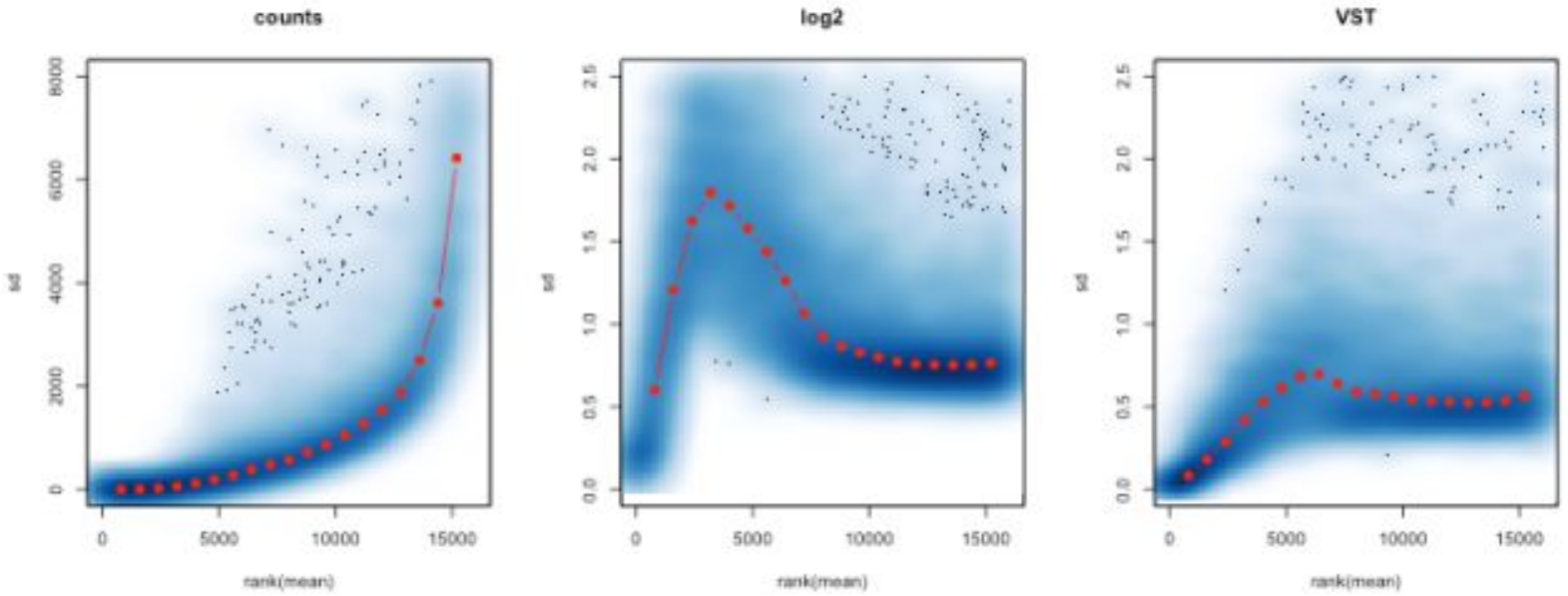
Однако что произойдёт, если мы выполним PCA на сырых каунтах?



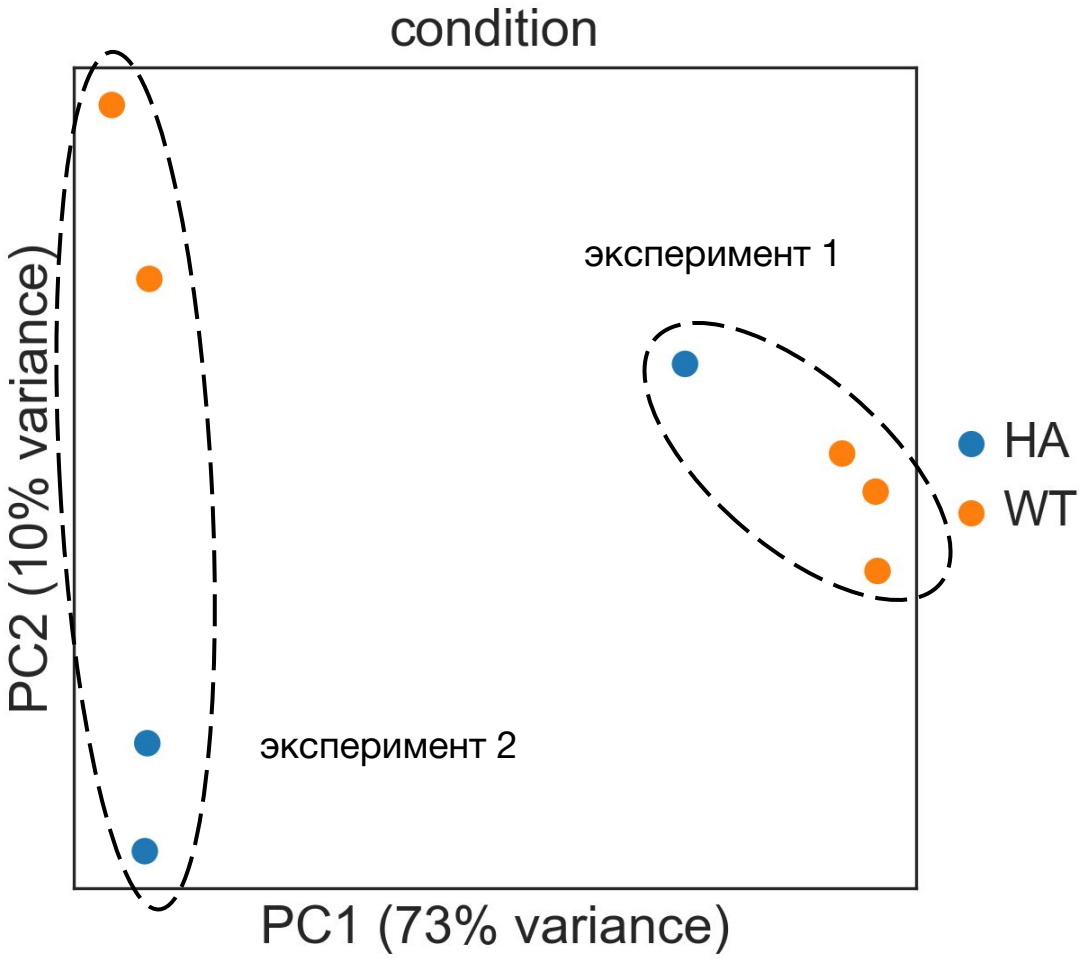
Дисперсия NB распределения

PCA пытается описывать дисперсию в данных, и проблема **овердисперсии** данных вылезает на первый план

Для борьбы с ней можно использовать различные трансформации (в первую очередь VST — variance stabilizing transformation)



Пример “плохого” PCA

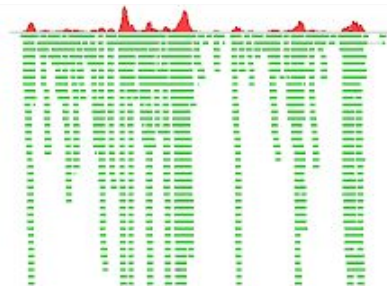


Дорожная карта анализа RNA-Seq

QC прочтений

FastQC Report

Выравнивание



Подсчёт экспрессий

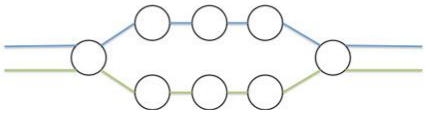
	Wild-type	
	Sample 1	Sample 2
Gene 1	24	31
Gene 2	0	3
Gene 3	1988	1125
Gene 4	5	0
...	...	...
Total	22341961	20739175

Нормализация

Дифференциальная экспрессия

GO, GSEA, ssGSEA и прочее

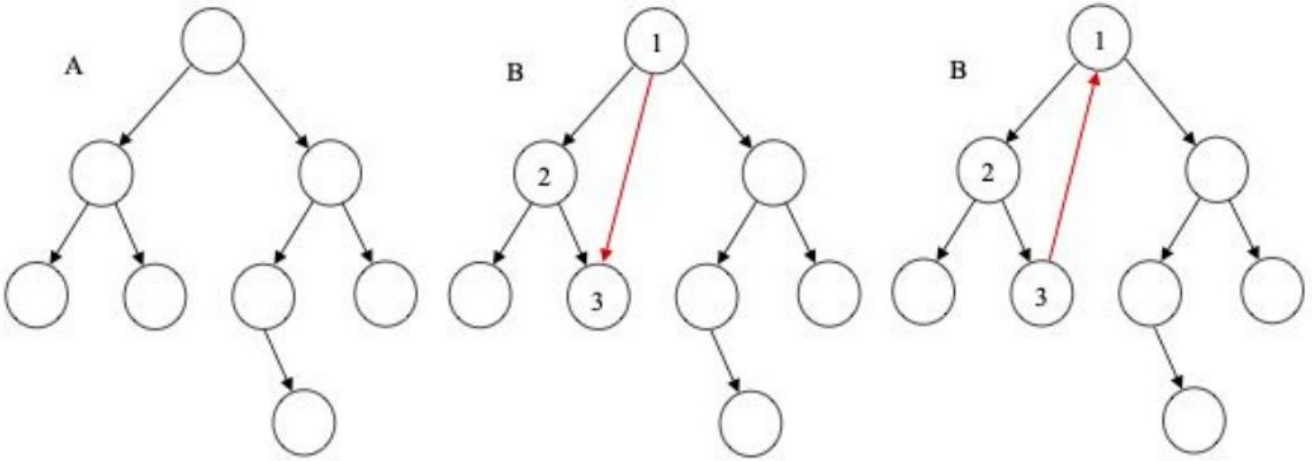
Псевдовыравнивание



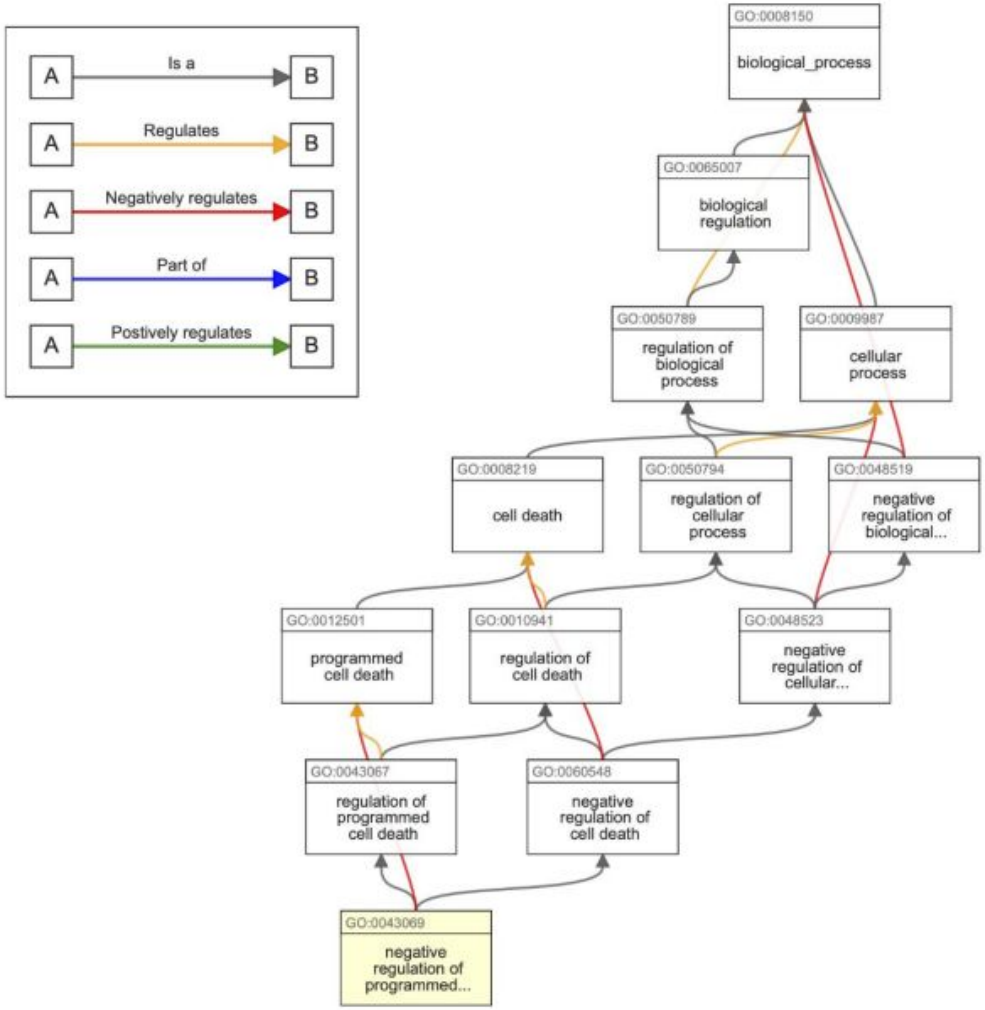
Gene Ontology (GO)

Три онтологии: **Biological Process**, **Molecular Function**, **Cellular Component**

Представляет собой направленный ациклический граф



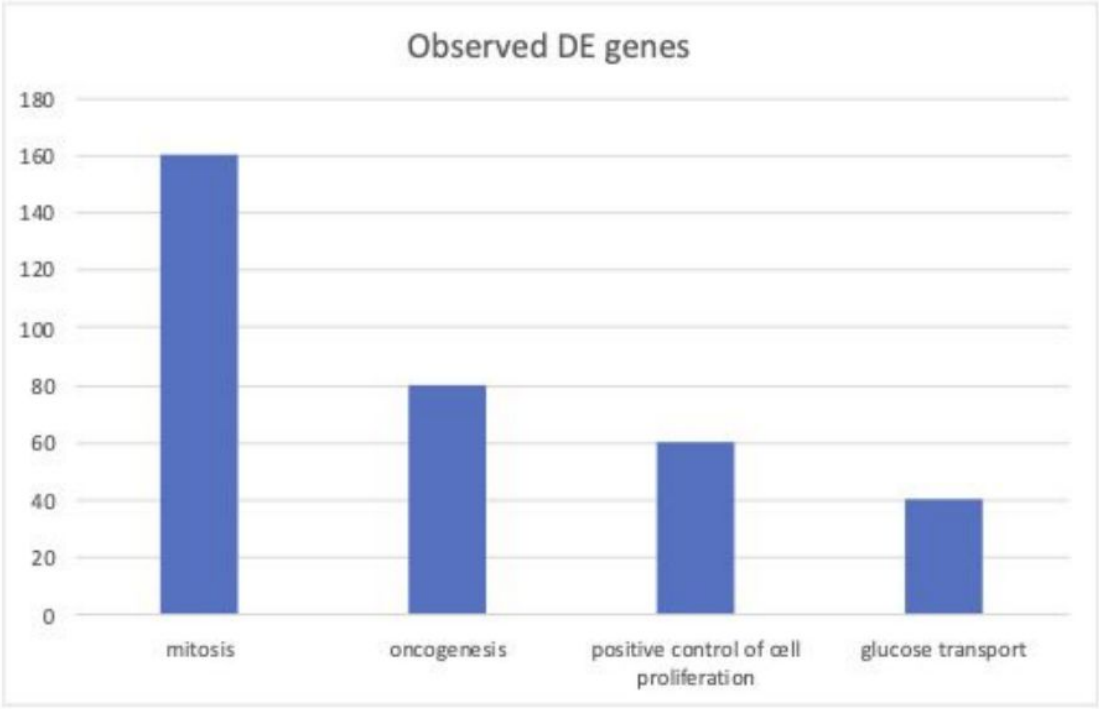
Архитектура GO



Результаты анализа GO

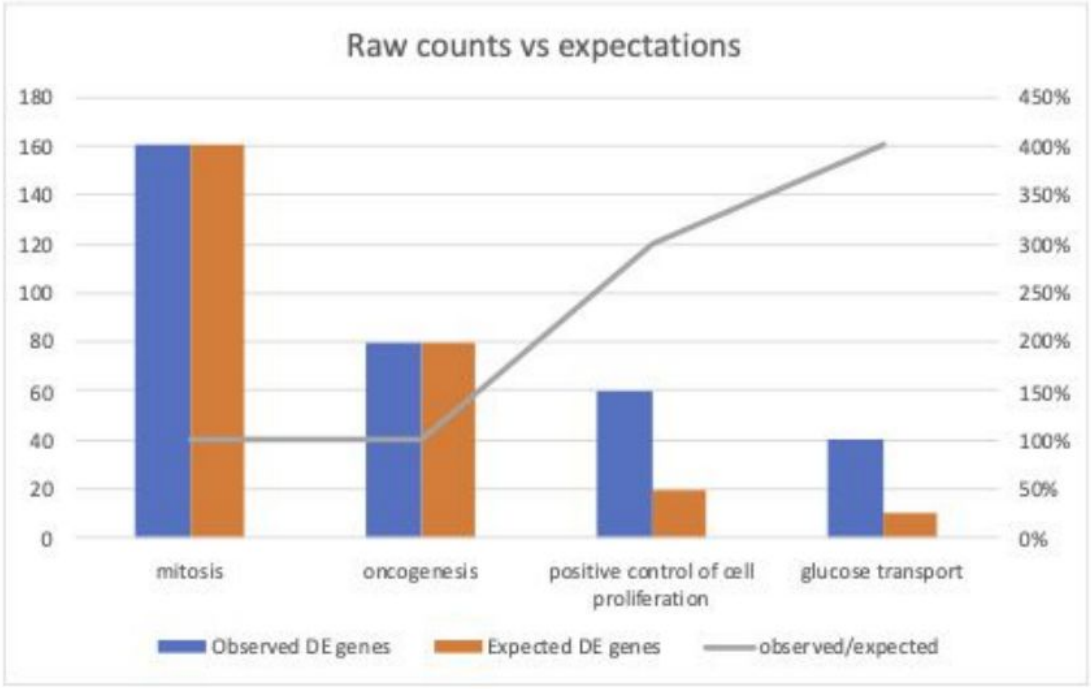
Мы можем определить, какие из дифференциально экспрессированных генов попадают в ту или иную онтологию

Можем ли мы исходя из этого делать какие-либо выводы?



Результаты анализа GO

Для корректной интерпретации результатов необходимо учитывать контекст, то есть сколько в принципе в этой онтологии содержится генов, а сколько у нас считаются дифференциально экспрессированными



Обогащение путей

У нас есть результаты дифференциальной экспрессии

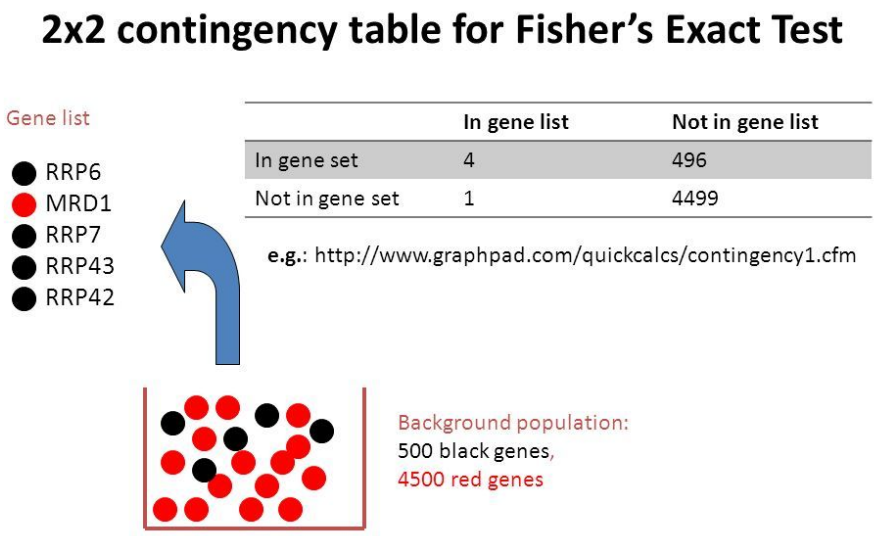
"Изменяется регуляция пути" = "Гены из этого пути изменяются неслучайно"

Мы хотим найти подобное неслучайное изменение в наших результатах

Точный тест Фишера

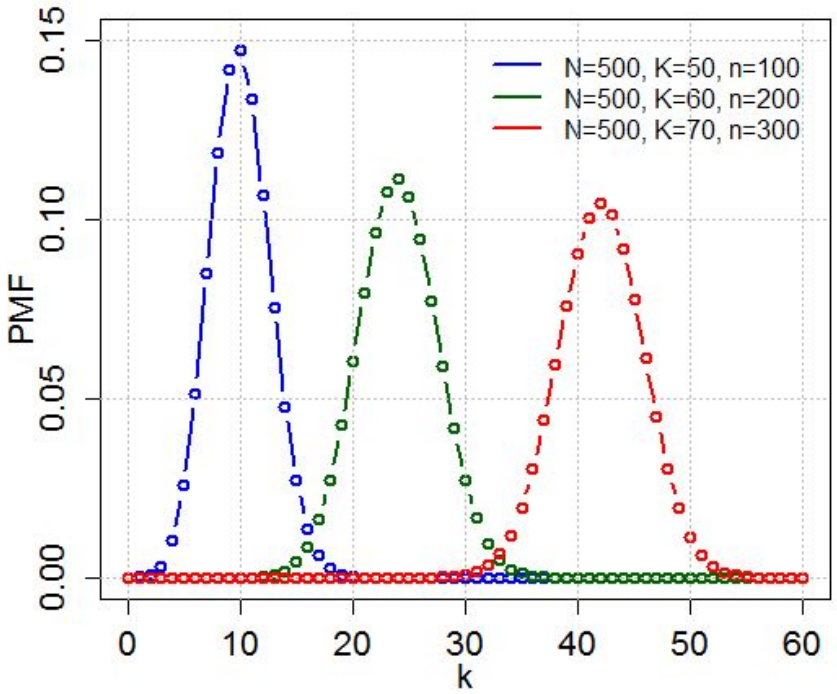
Он же гипергеометрический тест

1. Выбираем только значимо дифференциально экспрессированные гены
2. Проверяем пересечения этих генов с путями
3. Случайны ли пересечения?



Точный тест Фишера

$$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$



N = 12000: общее число генов (TOTAL)

K = 41: число генов в пути (SUCCESSES)

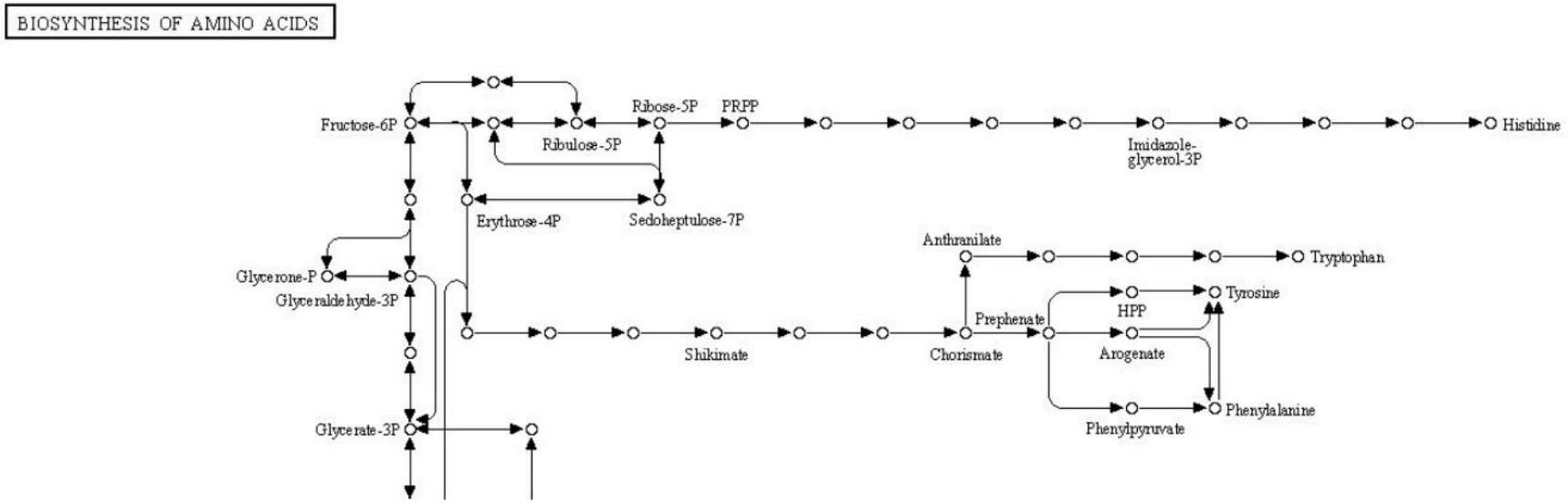
n = 113: число дифференциально экспрессированных генов (DRAWS)

k = 1: пересечение (SUCCESSFUL DRAWS)

Нулевая гипотеза — гены вытаскиваются из 12000 случайно относительно пути

Kyoto Encyclopedia of Genes and Genomes (KEGG)

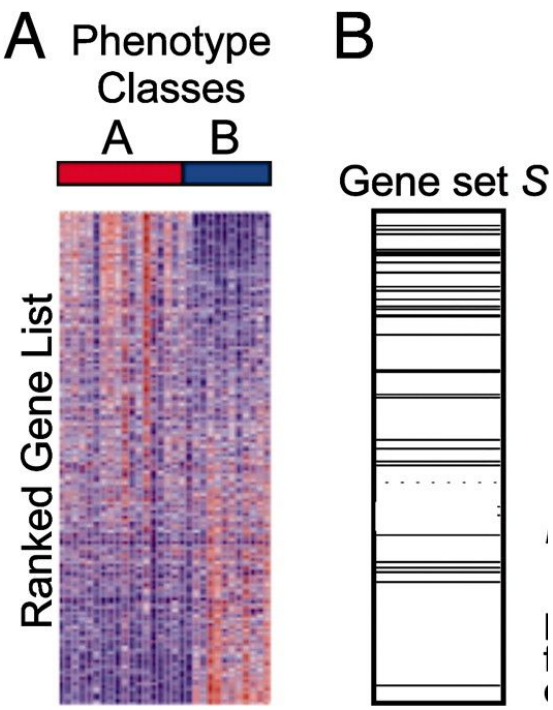
Курируемая база данных с биохимическими путями, имеет смысл использовать, когда есть гипотеза о какой-либо биохимической разнице между разными образцами



GSEA

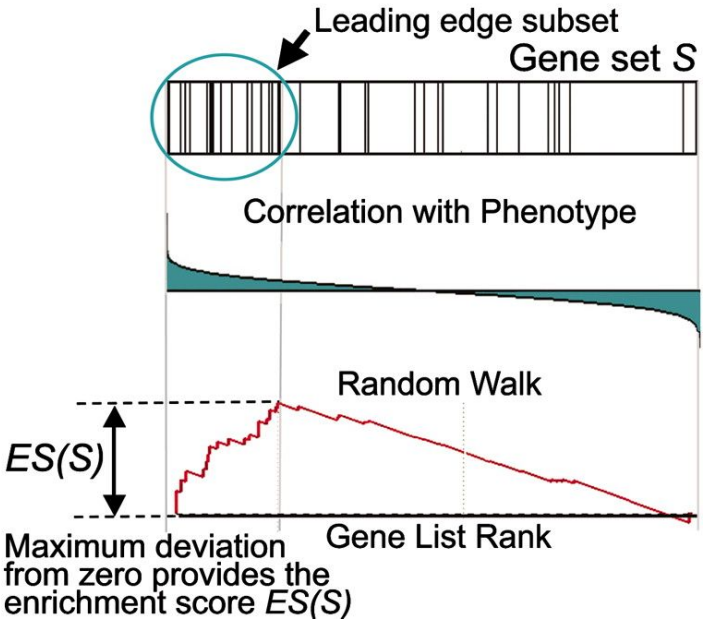
Классический алгоритм состоит из нескольких шагов:

1. У нас есть две группы (два фенотипа) — A и B
2. Считаем коэффициент корреляции Пирсона каждого из генов с таргетным фенотипом (с фенотипом A)
3. Ранжируем все гены по значению коэффициента корреляции
4. Где-то в этом ранге будут гены из нашей сигнатуры



GSEA

- Затем мы проходим по этому списку, высчитывая значение ES: прибавляем отнормированное значение коэффициента корреляции гена, если он из этой сигнатуры, или же вычитаем аналогичный показатель, если ген вне сигнатуры
- Максимальное (или минимальное) значение ES и будет нашим скором
- Перемешиваем лэйблы фенотипов, пермутируем 1000 раз — находим 0 распределение ES-статистики для сигнатуры => p-value



Preranked GSEA

Со временем пайплайн GSEA модифицировался, и теперь для ранжирования генов используют часто не коэффициент корреляции с фенотипом, а другие показатели —

- logFC гена между группами,
- p-value значимости различия генов между группами и т. п.

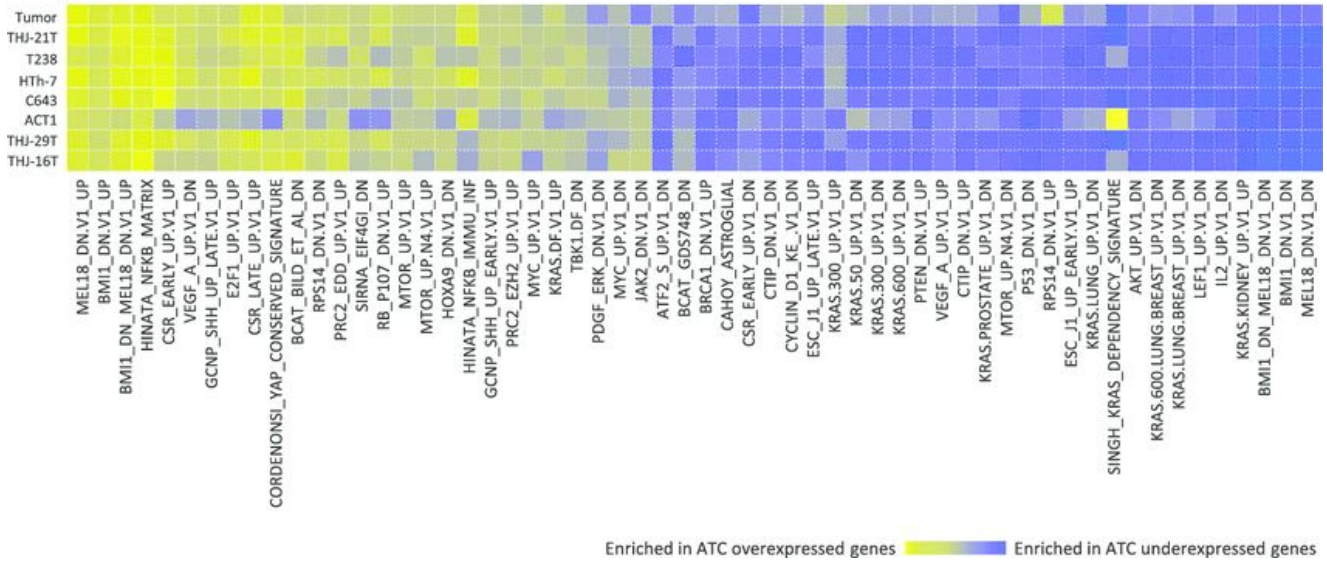
В таком случае p-value Enrichment score оценивается не пермутациями фенотипов, а пермутации генов в сигнатуре

Почему пермутации фенотипов более предпочтительны для оценки значимости представленности?

Single sample GSEA (ssGSEA)

Подобную процедуру можно выполнить и на обычных рангах экспрессии одного образца — только тогда мы получим некоторую оценку представленности данной сигнатуры в образце, без значимости

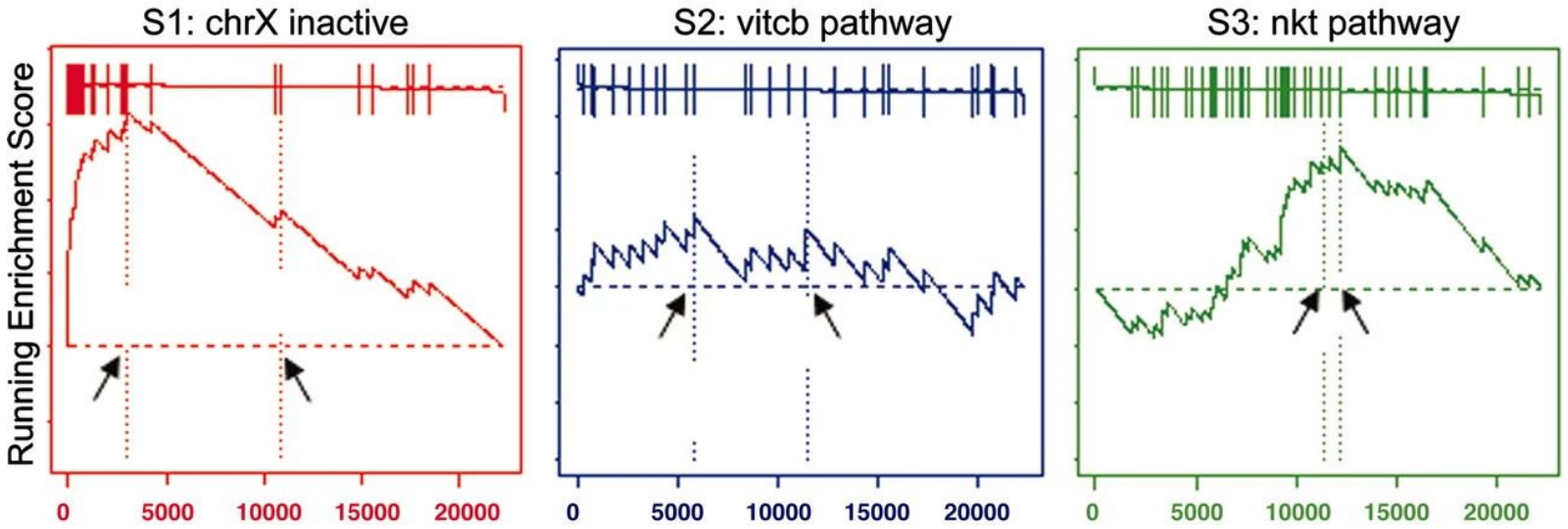
Однако потом мы можем использовать эту оценку для того, чтобы кластеризовать образцы или работать с ними на уровне сигнатур, а не отдельных генов



Normalized ES (NES)

Как легко понять, ES сильно зависит от того, насколько большая используется генная сигнатура

Для этого ES нормализуют на количество генов в сигнатуре



MSigDB

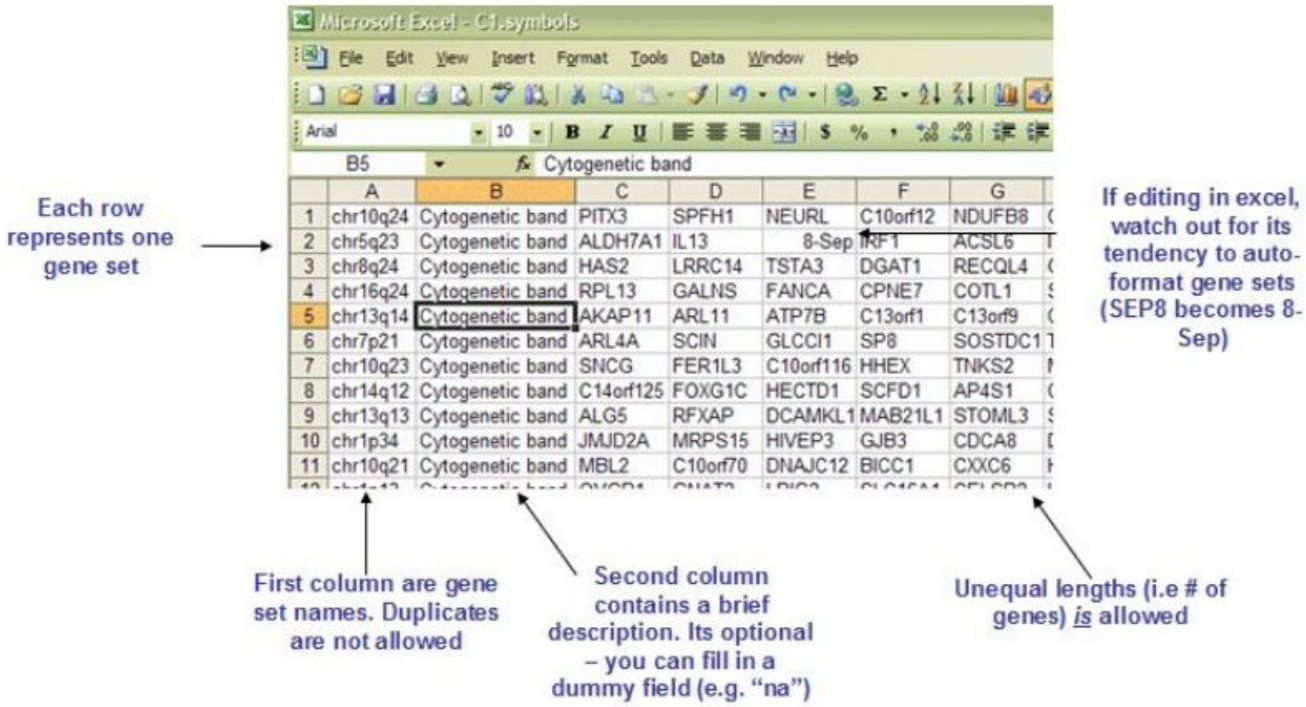
Отдельная база данных, аналогичная GO, но также включающая в себя множество сигнатур, полученных в ходе экспериментов по нокаутам / нокдаунам и проч.

В принципе включает в себя все сигнатуры из GO

Классически используется с GSEA, но подходит также и для обычного анализа обогащения

Формат gmt

GMT = Gene Matrix Transposed



Microsoft Excel - C1.symbols

File Edit View Insert Format Tools Data Window Help

B5 Cytogenetic band

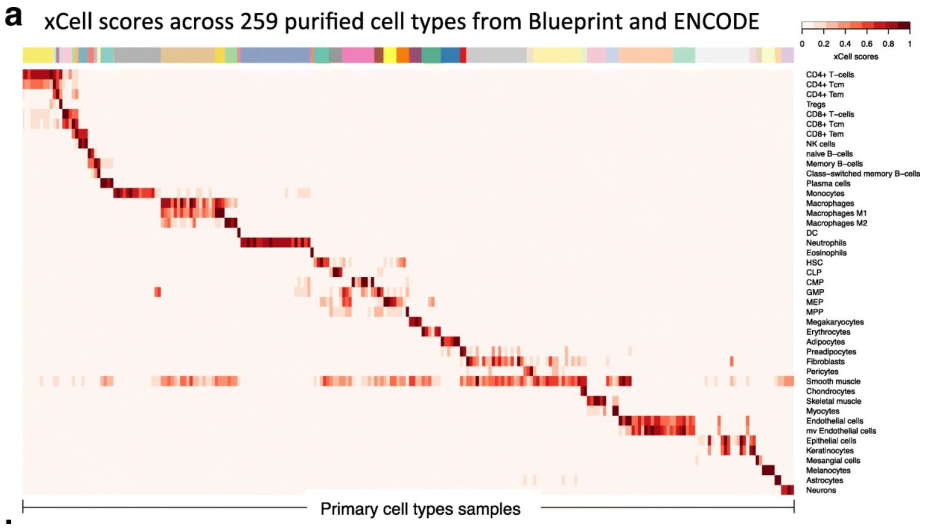
	A	B	C	D	E	F	G
1	chr10q24	Cytogenetic band	PITX3	SPFH1	NEURL	C10orf12	NDUFB8
2	chr5q23	Cytogenetic band	ALDH7A1	IL13	8-Sep	IRF1	ACSL6
3	chr8q24	Cytogenetic band	HAS2	LRRC14	TSTA3	DGAT1	RECQL4
4	chr16q24	Cytogenetic band	RPL13	GALNS	FANCA	CPNE7	COTL1
5	chr13q14	Cytogenetic band	AKAP11	ARL11	ATP7B	C13orf1	C13orf9
6	chr7p21	Cytogenetic band	ARL4A	SCIN	GLCCI1	SP8	SOSTDC1
7	chr10q23	Cytogenetic band	SNCG	FER1L3	C10orf116	HIHEX	TNKS2
8	chr14q12	Cytogenetic band	C14orf125	FOXG1C	HECTD1	SCFD1	AP4S1
9	chr13q13	Cytogenetic band	ALG5	RFXAP	DCAMKL1	MAB21L1	STOML3
10	chr1p34	Cytogenetic band	JMJD2A	MRPS15	HIVEP3	GJB3	CDCA8
11	chr10q21	Cytogenetic band	MBL2	C10orf70	DNAJC12	BICC1	CXXC6

GMT format is convenient to store large databases of gene sets. For a handful of sets (<256) the gmx format offers greater excel-editability

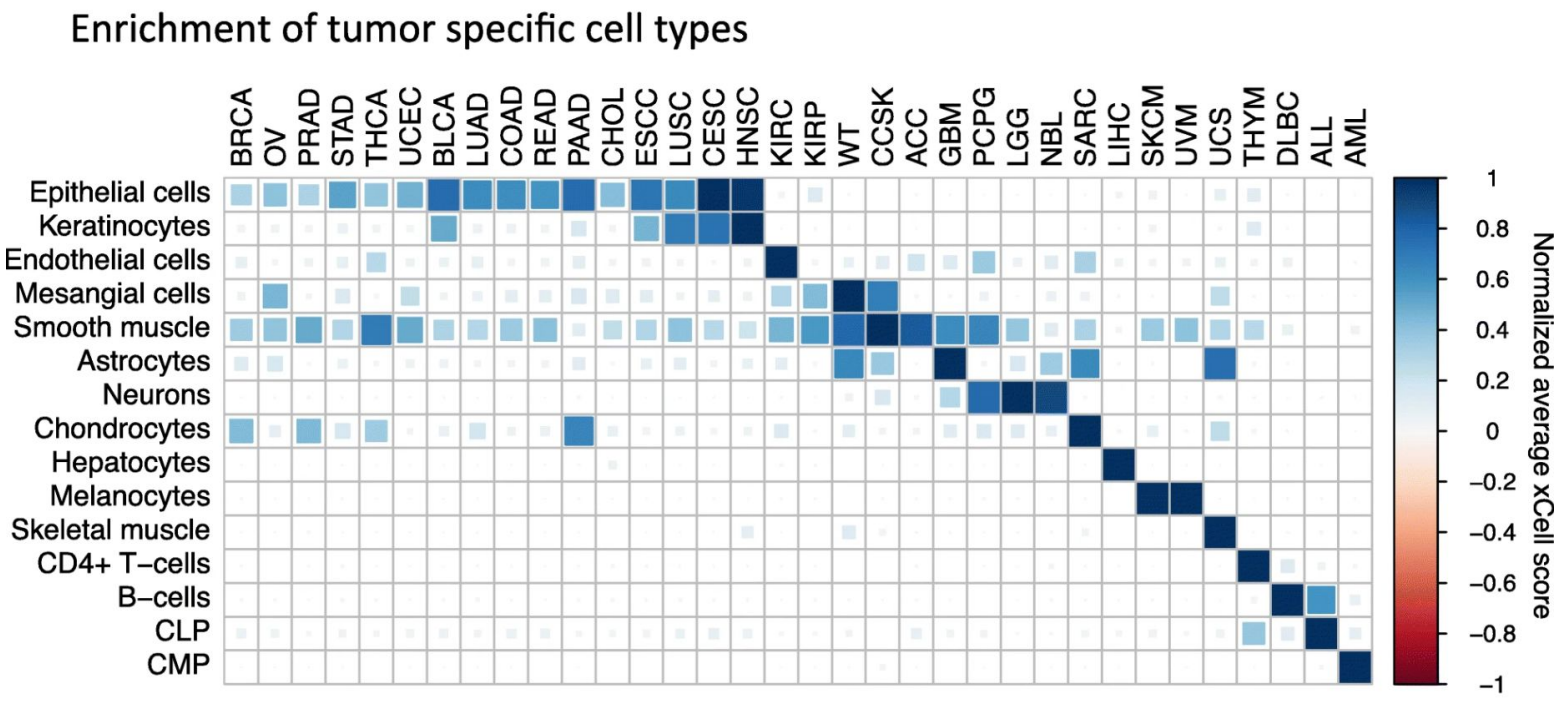
xCell

Инструмент xCell включает в себя базу данных с большим количеством тип клетки-специфичных сигнатур

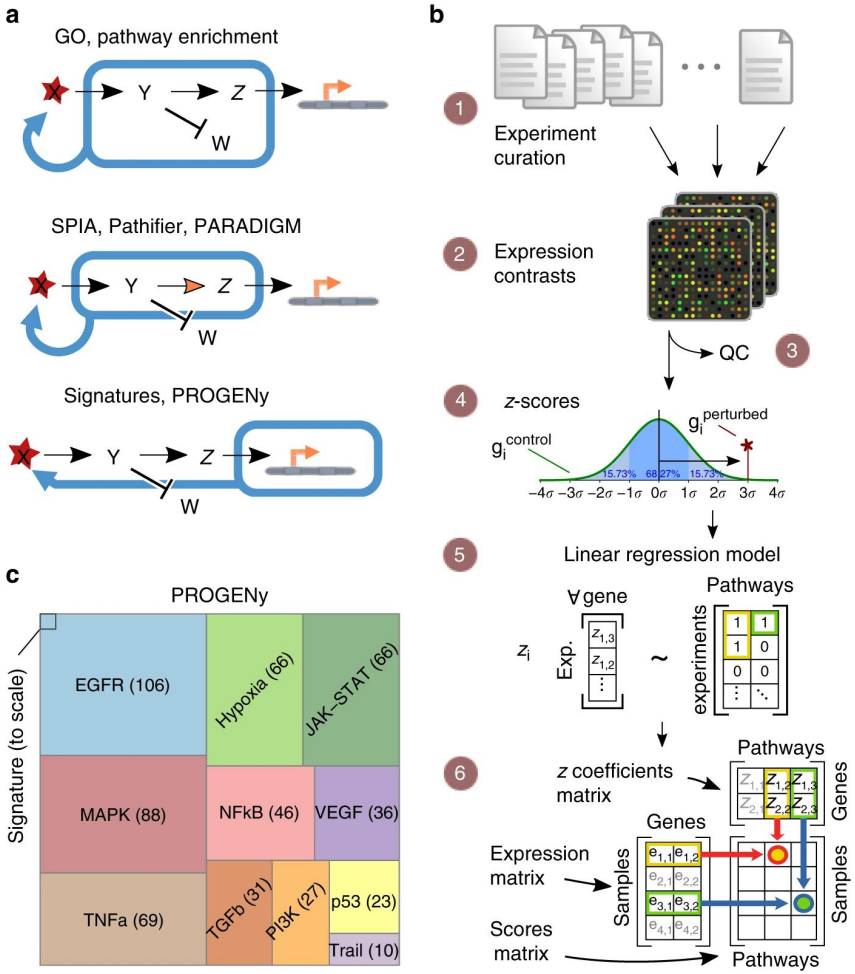
В результате можно для каждого образца оценить сигнатуры типов клеток, после чего работать с bulk RNA-Seq образцами в пространстве сигнатур каждого из типов клеток, то есть в первом приближении сравнивать их состав



Портреты раковых опухолей xCell



Pathway RespOnsive GENes (PROGENy)



База данных, которая содержит биологические пути (ген и его вес), необходимые для реконструкции активности пути

Активность пути — это линейная комбинация экспрессий генов с весами

Веса-вклады каждого гена в каждый из путей оценивались при помощи массового скрининга пертурбационных экспериментов

DoRothEA

База данных с большим количеством регулонов (пар транскрипционный фактор + список таргетов транскрипционного фактора)

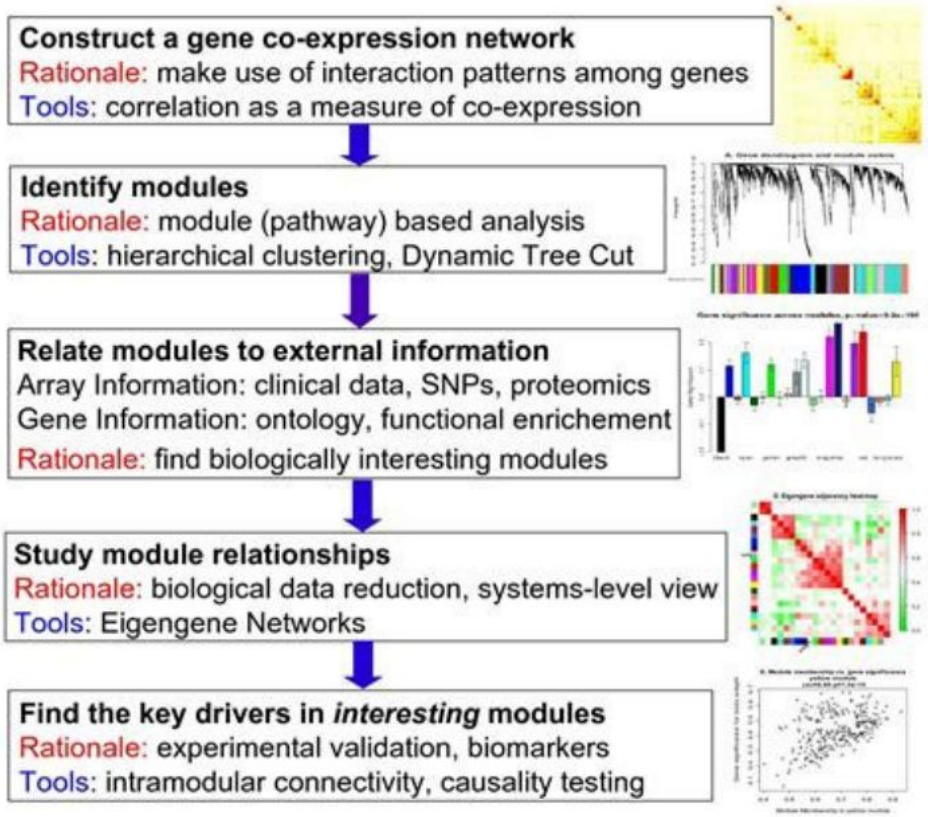
Каждому гену в регулоне соответствует уровень уверенности в том, что он регулируется именно этим транскрипционным фактором (в некоторых случаях этот уровень уверенности интерпретируется как коэффициент линейной модели)

WCGNA

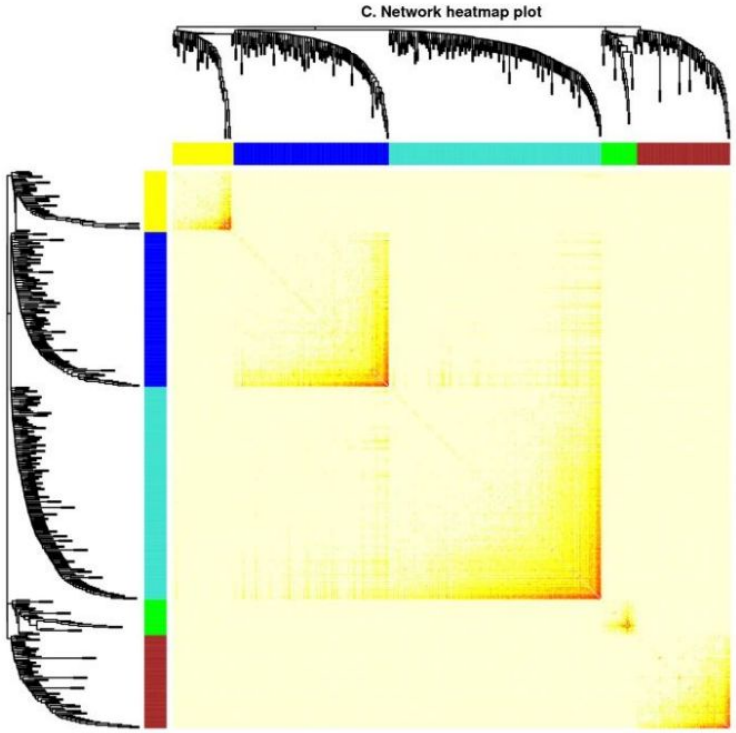
Метод для анализа групп коэкспрессированных генов

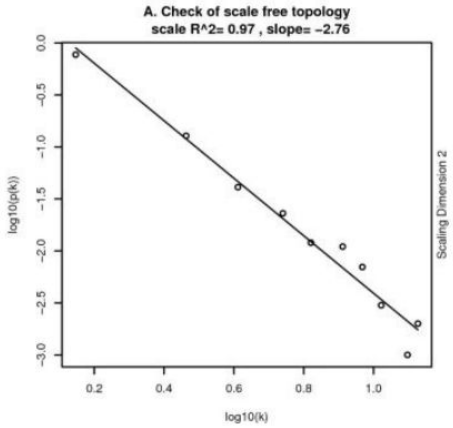
Модули коэкспрессии рассчитываются на основе корреляций экспрессии генов между образцами

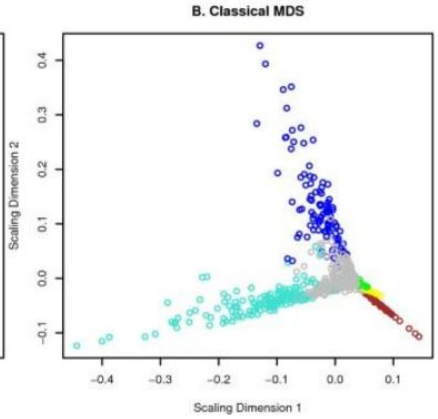
Модули можно анализировать GSEA или другими методами обогащения



Визуализация сетей взаимодействия генов







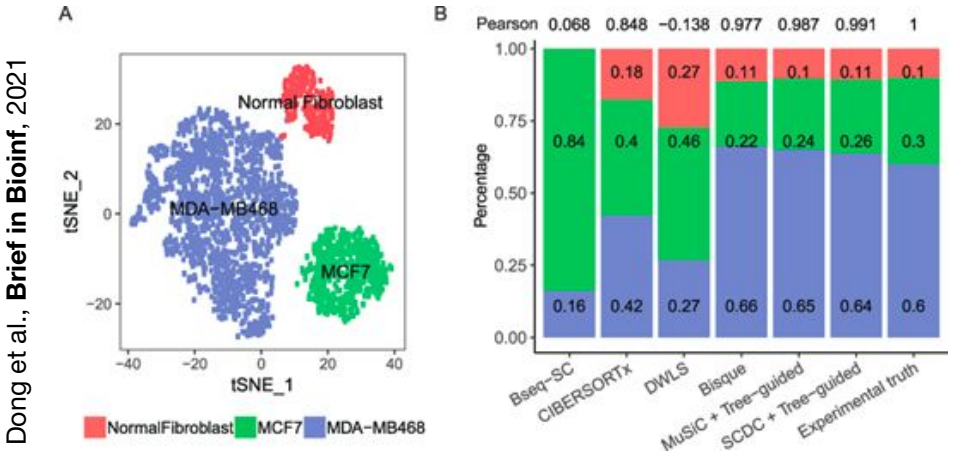
Деконволюция bulk RNA-Seq

Деконволюция bulk RNA-Seq — это процесс определения того, в каких долях какие клеточные типы содержатся в пробе

Основанная на маркерных генах
(BisqueRNA)

gene	cluster	avg_logFC
Gene 1	Neurons	0.82
Gene 2	Neurons	0.59
Gene 3	Astrocytes	0.68
Gene 4	Oligodendrocytes	0.66
Gene 5	Microglia	0.71
Gene 6	Endothelial Cells	0.62

Основанная на референсе
(SCDC, BisqueRNA)



Dong et al., **Brief in Bioinf**, 2021