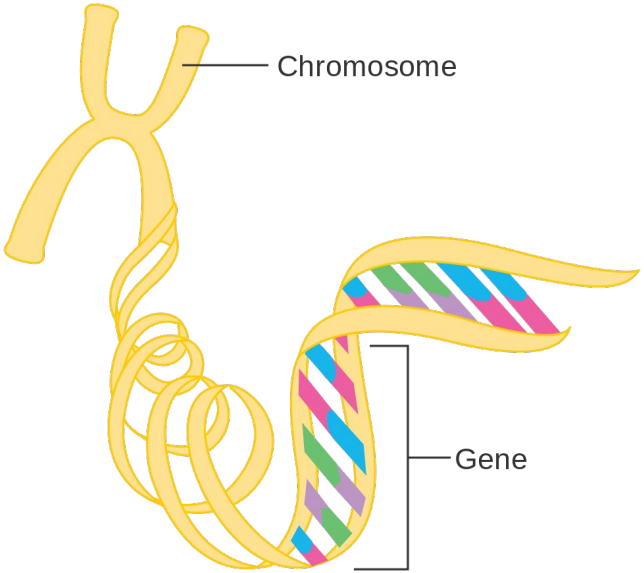
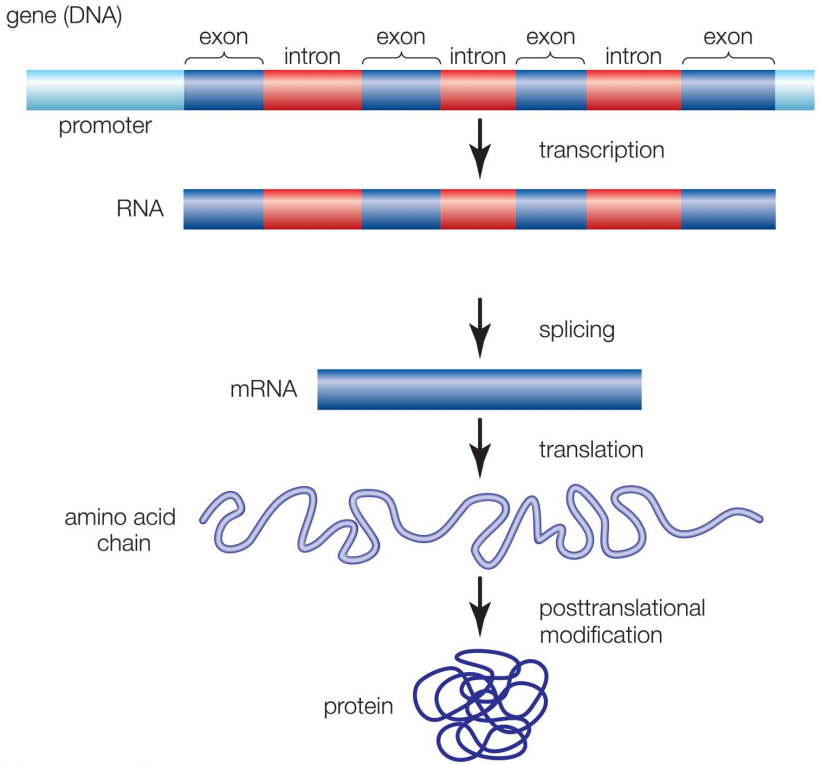
Gene set enrichment analysis

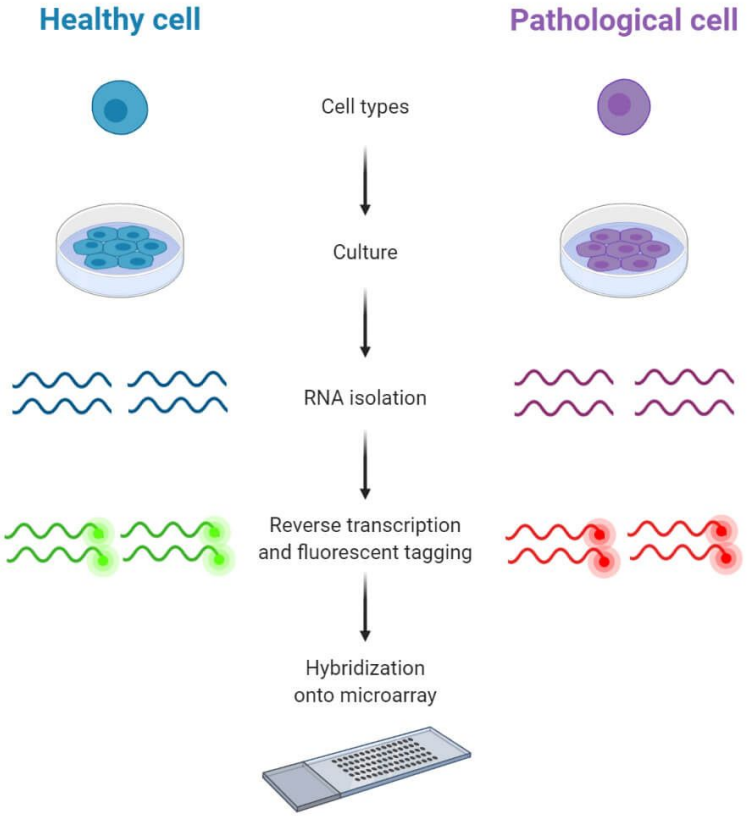
План

- Анализ данных экспрессии микрочипов
- Что такое Gene Set Enrichment Analysis
- Как работает GSEA
- Вариации методов GSEA
- Инструменты и фреймворки
- Примеры

Структура и механизм экспрессии генов







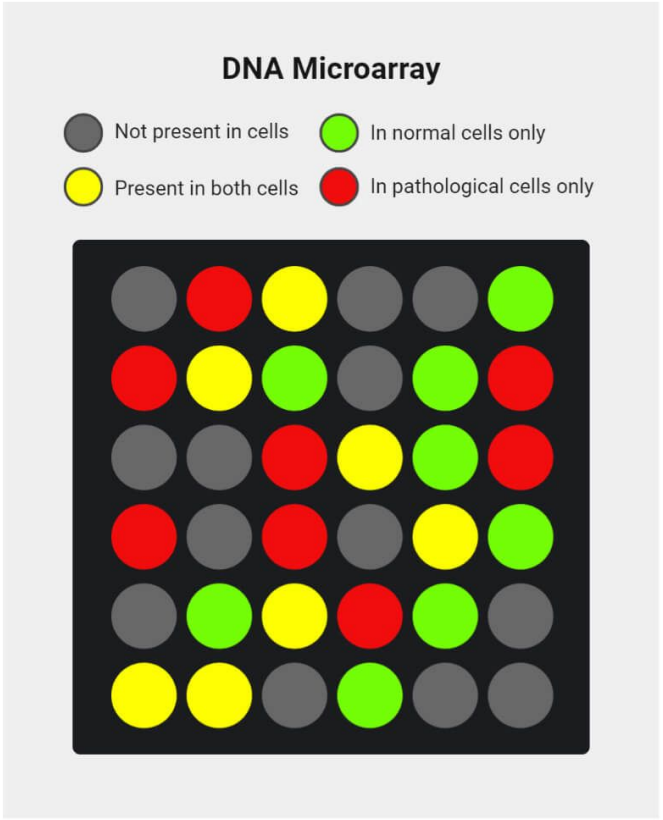
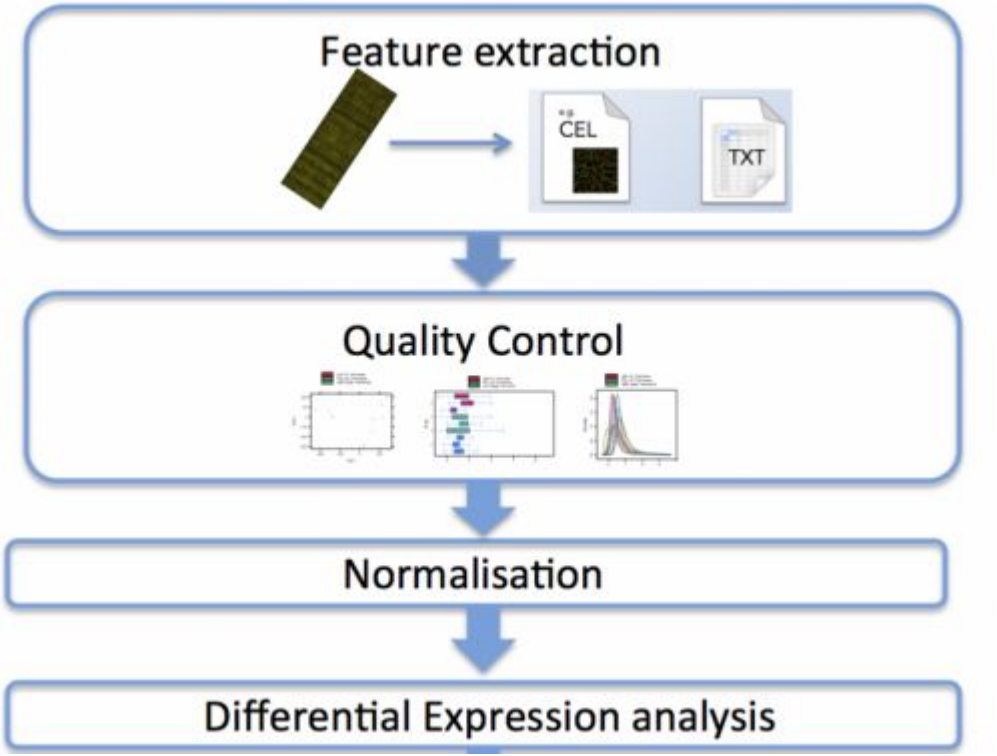
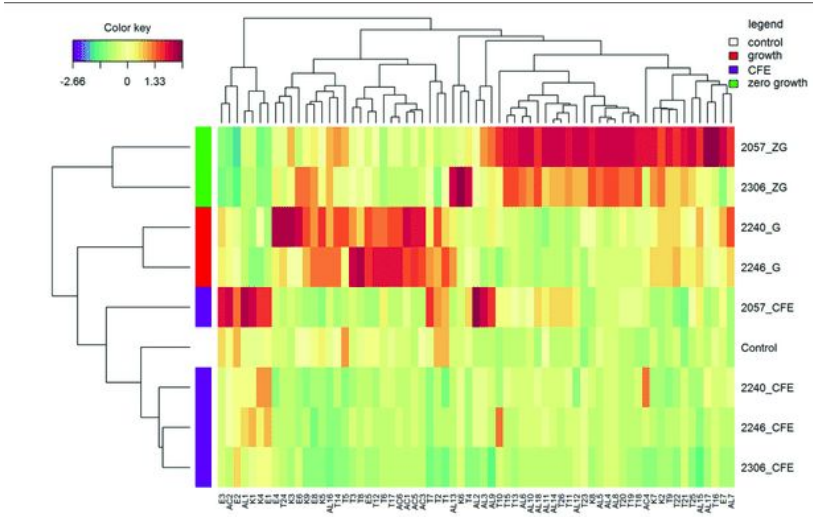


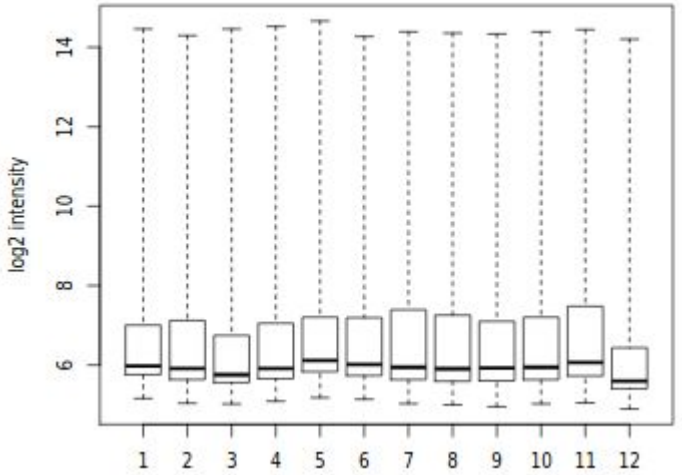
Image By Sagar Aryal, created using biorender.com

Как устроен общий подход к анализу экспрессии генов?



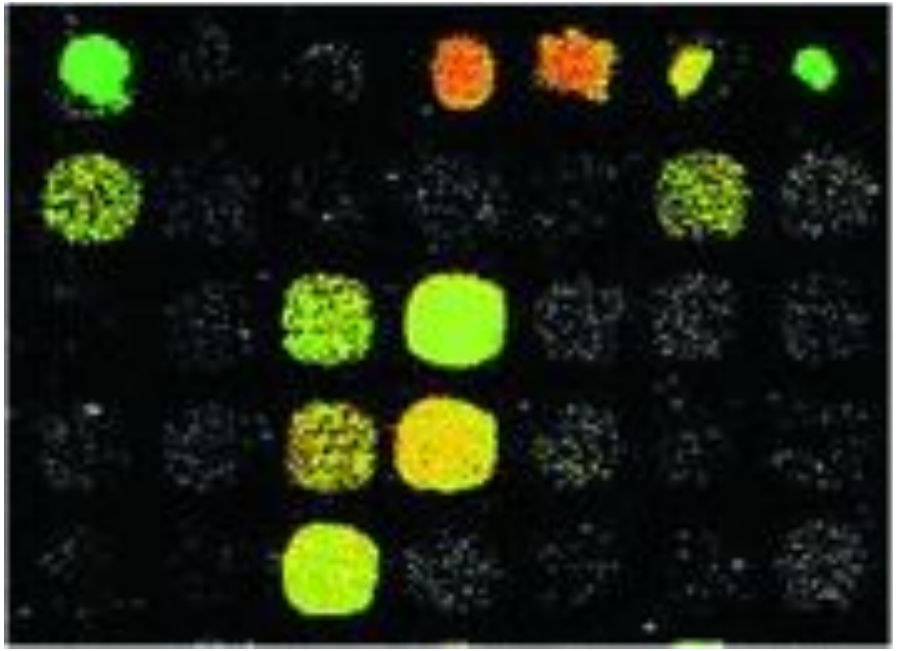
Как устроен общий подход к анализу экспрессии генов?

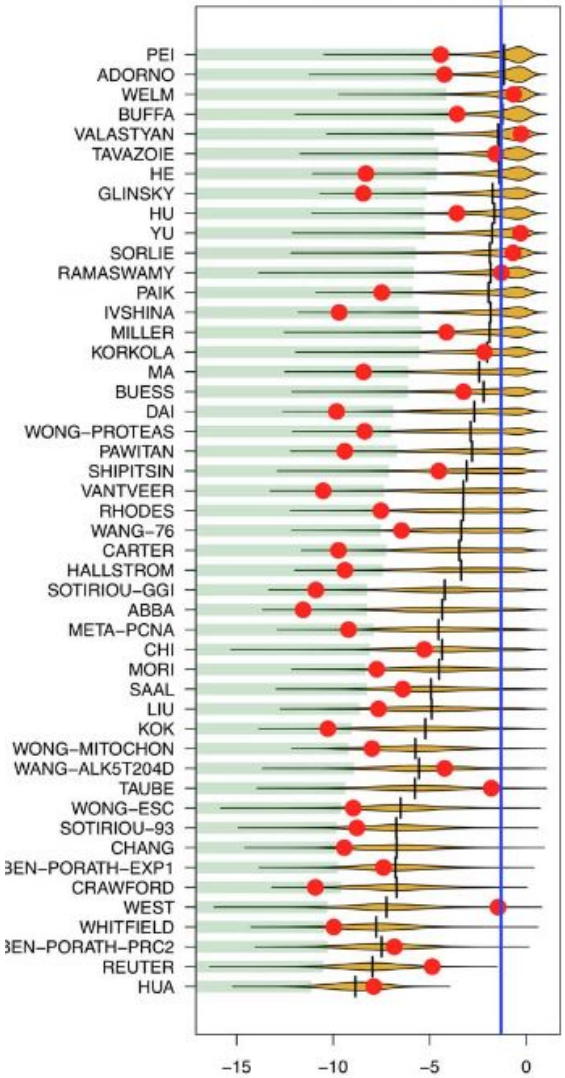




Какие же могут возникнуть проблемы?

Зашумленность данных с микрочипов





Высокая статистическая значимость случайных генов

> “Most published signatures are not significantly better outcome predictors than random signatures of identical size”

Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome
David Venet 2011

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on

Что такое Gene Set Enrichment Analysis?

- Это метод, для интерпретации экспрессии генов, находящий группы генов, которые имеют **сходную биологическую функцию**

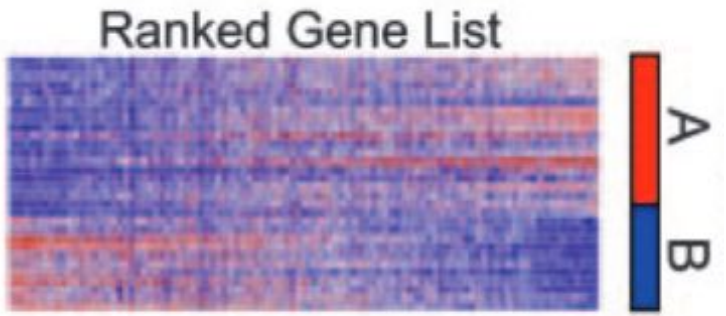
Как работает GSEA

Входные данные

- Генсет с экспрессией ***D*** с ***N*** генами и ***k*** сэмплами
- Априорный генсет ***C*** с функциональной аннотацией (фенотип, pathway)
- Подмножество ***S*** генсета ***D*** по определенной функциональной аннотацией, |S| = ***Nh***

Расчет Enrichment Score

- ***L*** <- sort(***D***) на основе корреляции gj $\in$ $\boldsymbol{D}$: cor(gj) = rj с генами из априорного датасета ***C***



Расчет Enrichment Score

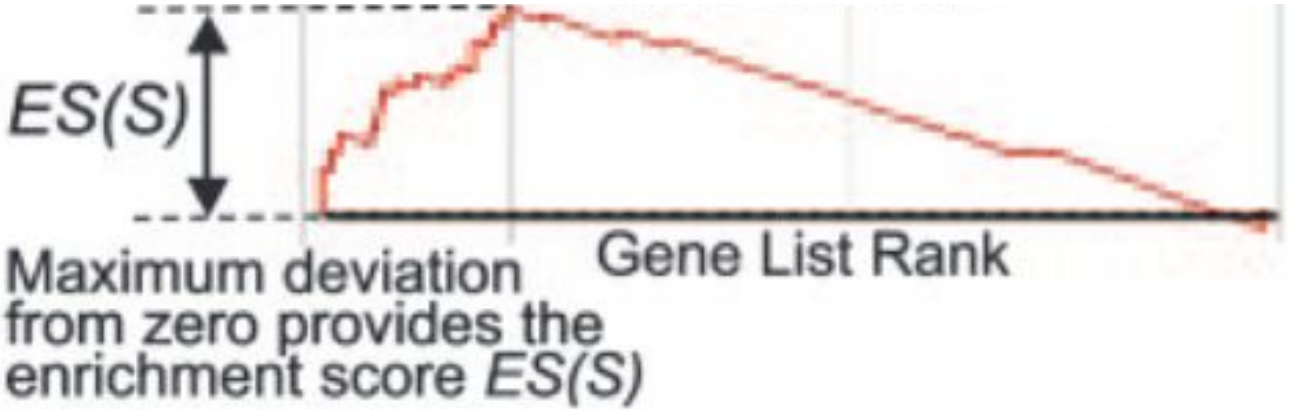
ES рассчитывается путем прохода по всему списку ***L***, применяется функция $P\hbar it$, если g$i \in$ ***S***, или $Pmiss$ в противном случае. (p >= 0)

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{\text{miss}}(S, i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)}.$$

Расчет Enrichment Score

- ***ES*** <- max(|$Phit - Pmiss$|)



Свойства ES

- Если ***S*** случайно распределен, то ***ES*** относительно маленький, Если ***S*** не случайно распределен, то и ***ES*** будет тоже относительно большим

Оценка уровня значимости ES

Статистическая значимость ***ES*** (***P-value***) оценивается следующим образом:

1. Метки классов переставляются и для каждого генного набора рассчитывается новое значение ***ES***, генерирующее нулевое распределение

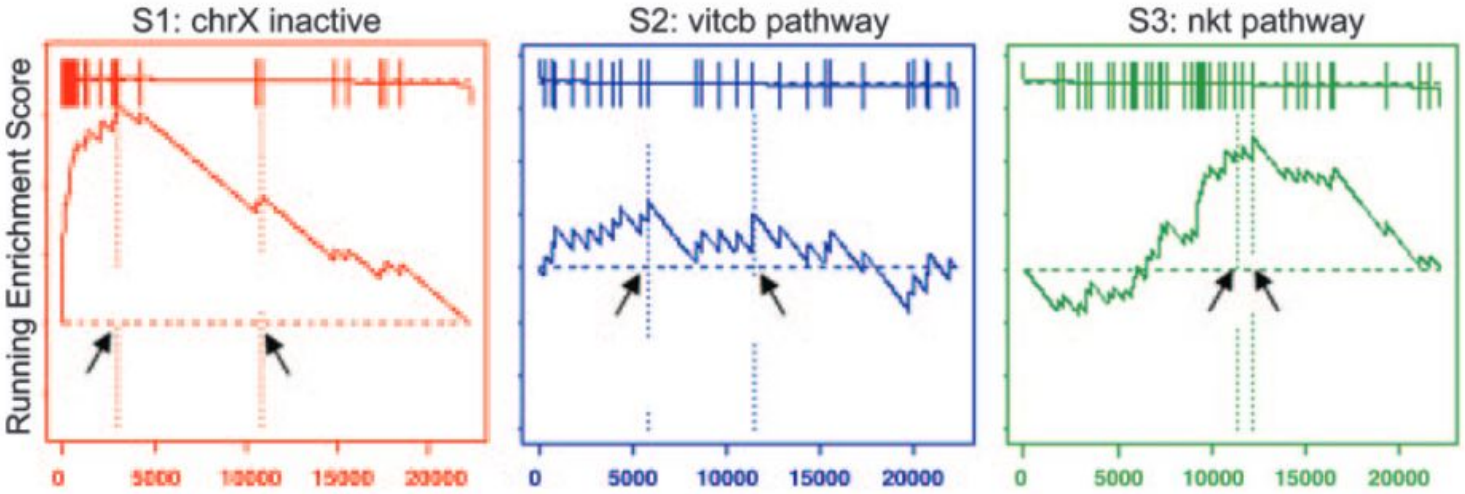
2. Относительно этого нулевого распределения вычисляется ***P-value***

Поправка для проверки множественных гипотез

- ***NES*** <- Нормализация ***ES*** для каждого набора генов с учетом размера набора
- Контролируем долю пропорцию ***false positive rate*** для каждого *NES*, сравнивая хвосты в нулевом распределении и полученном распределении ***NES***

Выделение самых значимых генов

- Данное подмножество располагается в окрестности точки, в которой текущая сумма достигает наибольшего отклонения от нуля



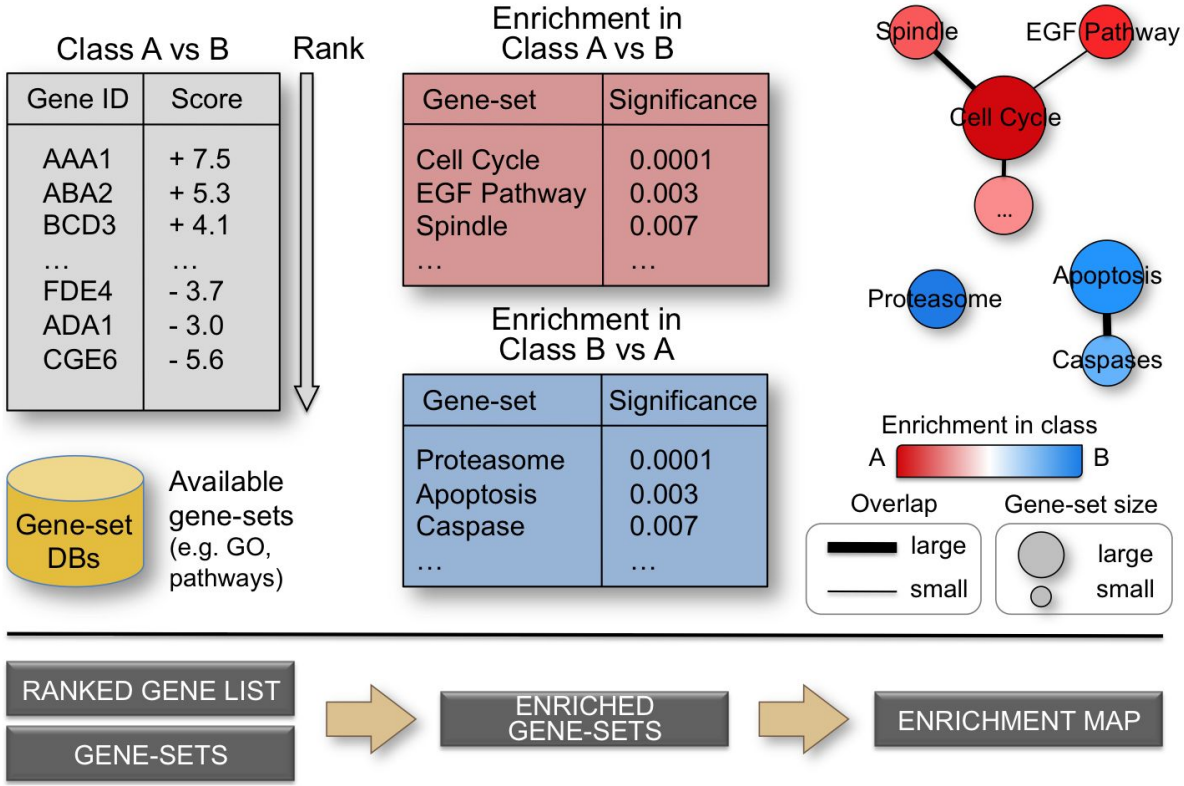
Вариации методов GSEA

- Если мало данных

В этих случаях ***P-value*** оценивается путем перестановки генов

- Если хотим сравнить 1 ген против всех

Построение enrichment map

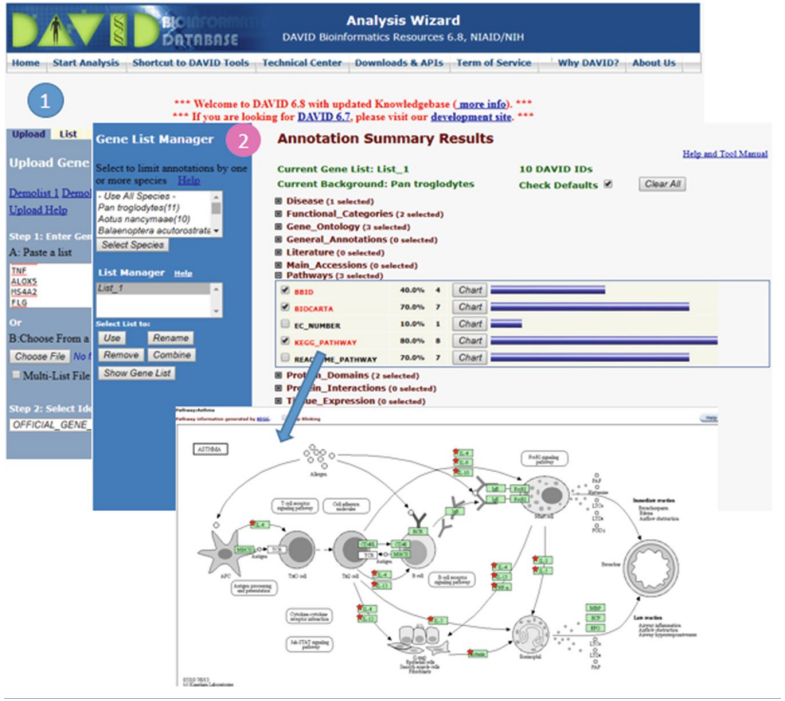


Инструменты и фреймворки

DAVID

- Онлайн ресурс для визуализации и аннотаций генетической информации

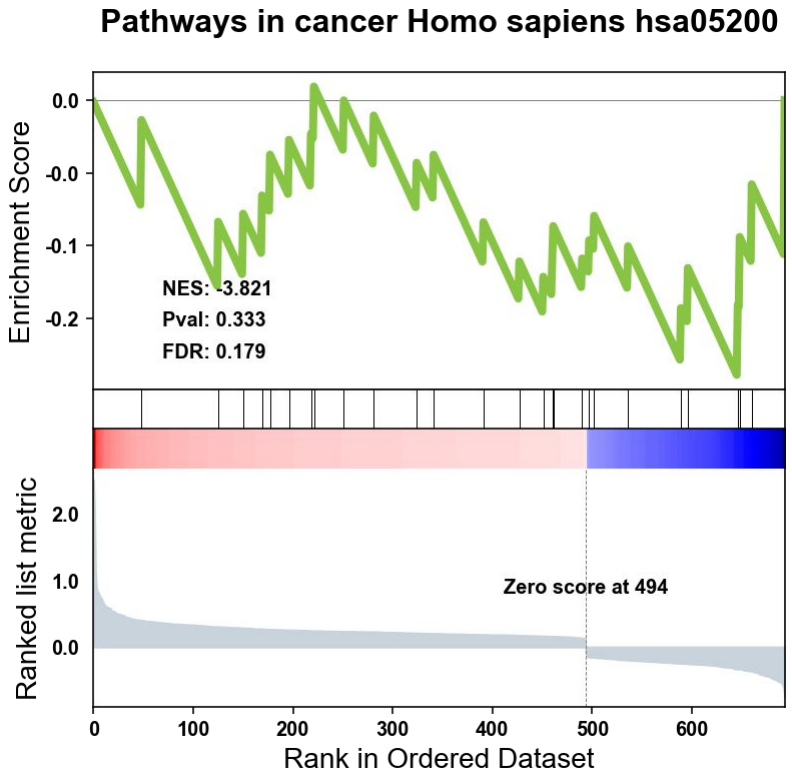
- Помимо GSEA анализа включает: построение KEGG pathways, поиск связей между генами и заболеваниями



GSEAPY

- Простая библиотека, которая в основном реализуем алгоритм из оригинальной статьи

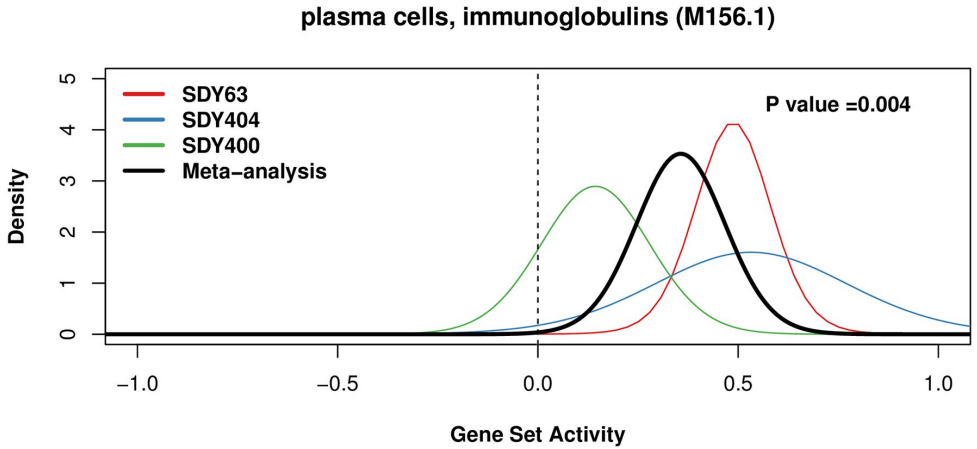
- Доступен в виде библиотеки для Python



QuSAGE

- Метод для осуществления GSEA, способный проводить межгенные корреляции и количественно определять активность генного набора

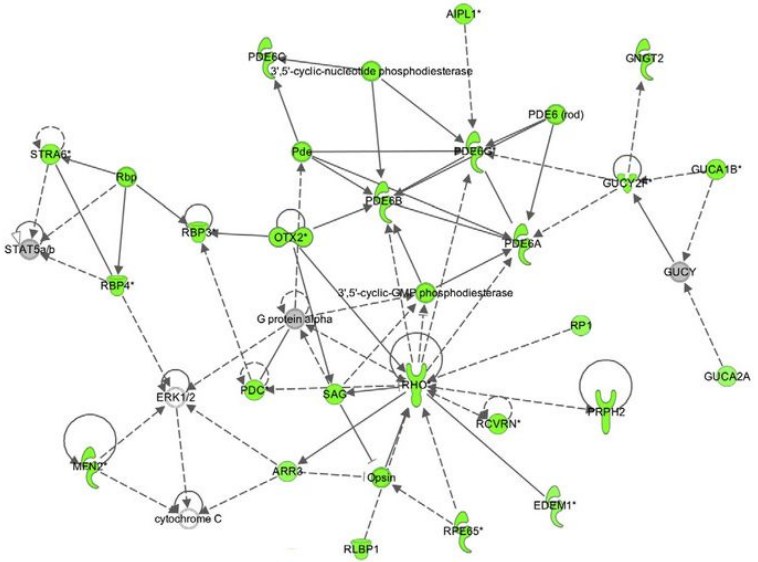
- Доступен в виде пакета для R с возможностью визуализации



Ingenuity

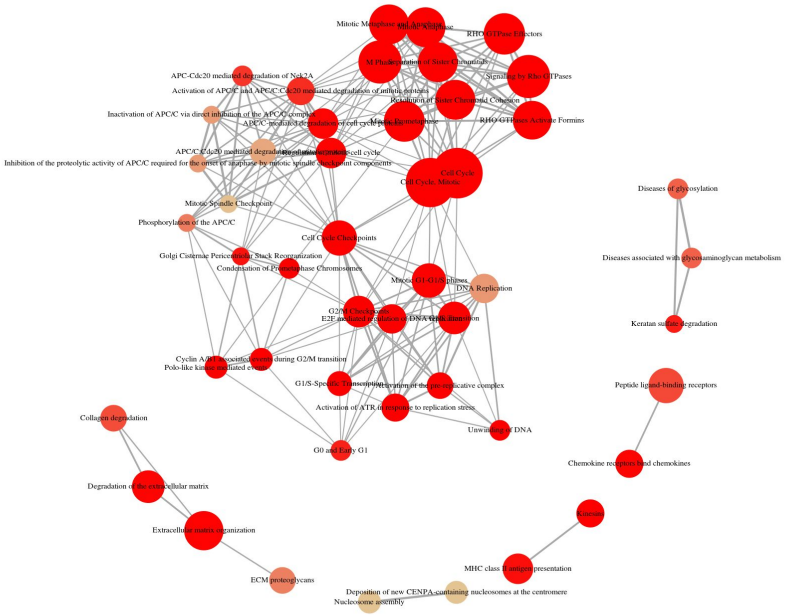
- Веб-приложение, анализирующее данные экспрессии генов / микроРНК / SNP микрочипов

- Содержит функцию анализа генетических сетей (pathways analysis)



Reactome

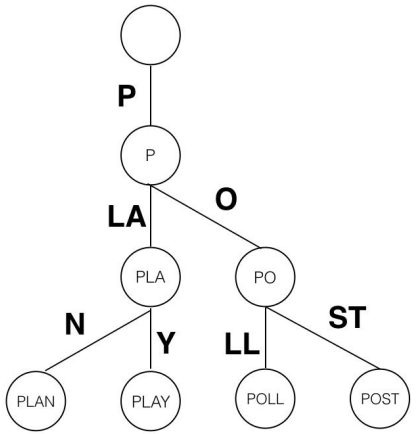
- Открытая база данных по генетическим путям (genetic pathways)
- Включает инструменты для анализа генома, моделирования, визуализации, интерпретации и анализа genetic pathways
- Доступна для R



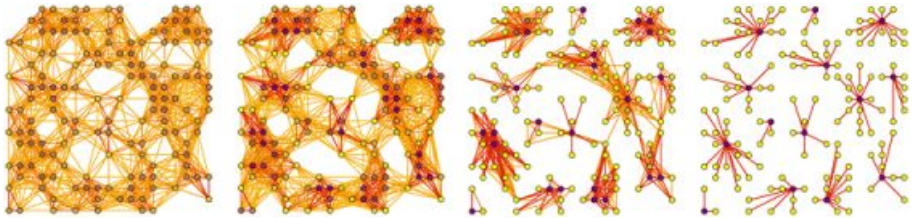
Некоторые методы в Reactome

- Для построения pathways:
  - Radix tree
  - Double-linked tree

Radix Tree

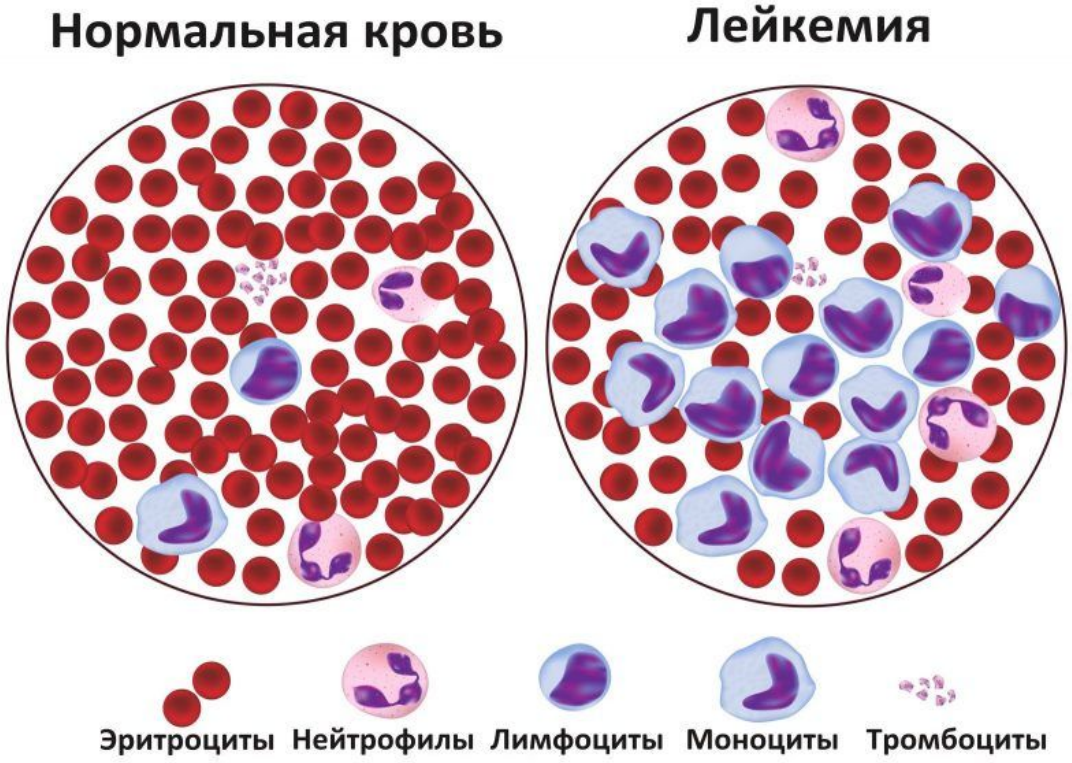


- При анализе microarray data:
  - Markov graph clustering



Примеры использования

Acute Leukemias (острый лейкоз)

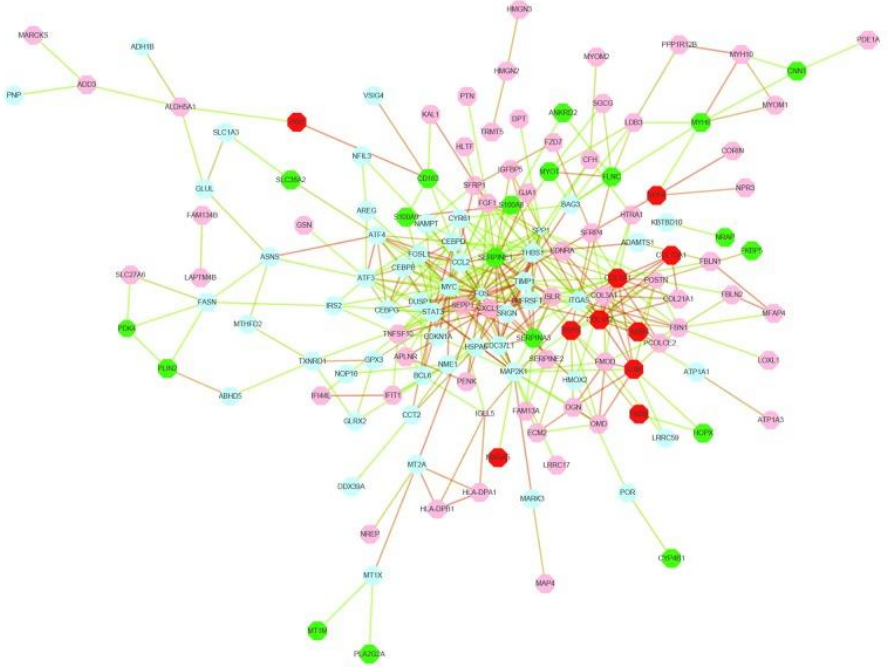


Acute Leukemias (острый лейкоз)

- Взяли датасет из 48 пациентов: 24 с острым лимфолейкозом и 24 с острым миелоидным лейкозом
- Выделенные гены соответствуют современным знаниям об остром лейкозе

Data set: Acute leukemias	
Enriched in ALL	
chr6q21	0.011
chr5q31	0.046
chr13q14	0.057
chr14q32	0.082
chr17q23	0.071

Идентификация потенциальных взаимодействий генов при сердечной недостаточности



- Датасет: 102 образца экспрессии генов
- Была построена и проанализирована сеть белковых взаимодействий для выявления потенциальных регуляторных белков с помощью GSEA

Спасибо за внимание!

Ссылки на статьи:

- https://www.pnas.org/content/102/43/15545 (GSEA)
- https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002240 (Random gene expression)
- https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6216482/

Авторы:

- Аверченко Марк
- Демичева Екатерина
- Прядко Анастасия