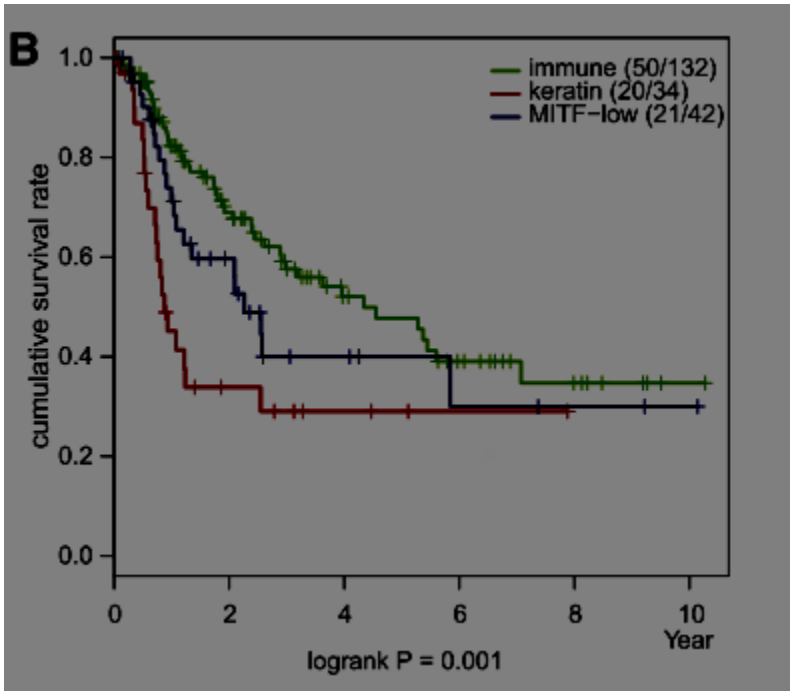
BostonGene
Transforming cancer treatment

Тестовое задание по биоинформатике №4

1. С сайта cbioportal.org получить данные TCGA-SKCM по когорте больных меланомой ссылка (кнопка download).

В архиве будет много информации по разным экспериментам на одних и тех же образцах: RNAseq (data_RNA_Seq_v2_expression_median.txt), клинические аннотации (data_bcr_clinical_data_sample.txt – по образцам, data_bcr_clinical_data_patient.txt – по пациентам. У пациента ID длиной в 12 символов, у образца – 16 символов. Нужно будет сопоставить 2 таблички и вытащить из них только самые интересные столбцы), …

2. Воспроизвести классификацию на 3 экспрессионных типа (MITF-low, keratin, Immune) из статьи, описывающей этот датасет (doi:10.1016/j.cell.2015.05.044).

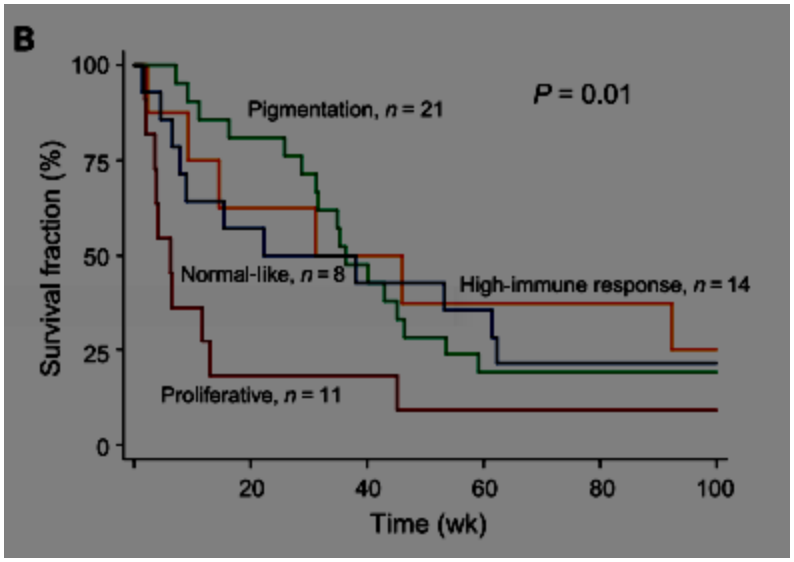


На момент написания статьи было доступно ~330 образцов. К какому классу они относятся можно найти в аннотации из статьи – первый supplementary файл, таблица S1D (лист в excel файле, колонка RNASEQ-CLUSTER_CONSENHIER). Можно использовать эту классификацию, а остальных пациентов доклассифицировать или проделать процедуру из статьи заново.

BostonGene
Transforming cancer treatment

3. На этих же данных (TCGA) сделать кластеризацию, описанную в статье http://clincancerres.aacrjournals.org/content/16/13/3356.full-text.pdf, на 4 класса Pigmentation, Proliferative, Normal-like, High-immune response.

В supplementary data на последней табличке есть список генов, по которым это можно сделать. Важно учесть, что платформы экспериментов разные – второй датасет на DNA microarrays – поэтому нельзя переносить абсолютные значения центроидов на NGS данные (TCGA).



4. Сравнить полученные классификации. Описать полученный результат.

Комментарий:

Не обязательно делать каждый пункт "идеально". Достаточно сделать простым и быстрым способом, кратко обосновать почему он выбран и описать последовательность шагов с поясняющими картинками.

В результате ожидается отчет в формате doc/pdf или Ipython/R ноутбук – как удобно.