Контрольная работа 1

1. Условия проведения

- Время проведения – полтора часа (дедлайн 17:30)
- Результаты контрольной ожидаются от вас на вашем гитхабе в формате Rmd + html
- Если не успеваете – залейте те результаты, которые у вас есть (они будут оценены)
- Можно пользоваться своими домашками и своим гитхабом
- Можно задавать вопросы мне и гуглу
- Нельзя задавать вопросы соседям и как-либо их отвлекать

2. Задание

2.1. Описание датасета

Вам будет предложен датасет RNA-seq 2017 года из статьи "*Three distinct cell populations express extracellular matrix proteins and increase in number during skeletal muscle fibrosis*". Статья о том, что важно понимать какие клетки производят компоненты внеклеточного матрикса; в поперечнополосатой мышце нахождение таких клеток важно в контексте заболеваний, где изменения внеклеточного матрикса могут вызывать фиброз и последующую повышенную жесткость тканей и их дисфункцию.

Примечательного в статье то, что ребята сначала нашли те клетки, которые производят компоненты внеклеточного матрикса, а затем научились эти клетки изолировать и показали, что их три разных подтипа: fibroblasts (FB), fibro / adipogenic progenitors (FAP), skeletal muscle progenitor (SMP). После чего они взяли два вида мышей: the nesprin-desmin double knockout (DKO) mouse и wild-type (WT) mouse, отсортировали из них эти клетки и сделали RNA-seq.

Описание этих образцов находится в файле *GSE89633_ conditions.tsv*, а сами данные экспрессии после выравнивания ридов на индексный геном находятся в файле *GSE89633_ counts.tsv*

2.2. Что нужно сделать

2.2.1. Visual quality control and self-consistance

Код:

- Построить PCA-plot на всех генах данного датасета
- Взять 8000 самых экспрессированных генов (по средней экспрессии во всех сэмплах), кластеризовать эти гены используя функцию *Kmeans* из пакета *amap* (возьмите число кластеров, которое вам нравится от 8 до 12), и построить heatmap, как в домашней работе по кластеризации.

Необходимо также ответить на вопросы (просто написать ответ вне кодового блока):

- Можем ли мы судя по PCA plot и heatmap сказать, что в этом датасете есть явные аутлаеры?
- Можем ли мы предположить по PCA plot, чем объяснена большая часть вариации в наших данных?

2.2.2. Differential expression

Код:

- Необходимо произвести три сравнения (для каждого типа клеток) DKO vs WT:
  1. FB WT vs FB DKO
  2. FAP WT vs FAP DKO
  3. SMP WT vs SMP DKO
- Для результатов этой дифф.экспрессии построить три volcano plot (рядом друг с другом через facet_grid)
- Взять из всёх трех сравнений дифф.экспрессированные гены (с p.adj $<$ 0.01) и построить тройную диаграмму Вена

Необходимо также ответить на вопросы (просто написать ответ вне кодового блока):

- Можем ли мы по volcano plot предположить, транксрипционное состояние каких типов клеток изменилось сильнее/слабее после двойного нокаута?