0.1 Задача первая

Есть наборы полиморфизмов и вставок/делеций для 20 штаммов *M. tuberculosis* относительно одного референтного генома *M. tuberculosis* H37Rv. (данные по ссылке). Для референтного генома также имеется аннотация (файл **.ptt**). Каждый **.snps** файл представляет из себя таблицу с полиморфизмами (и вставками/делециями) для одного из образцов. Задача: загрузить данные в R, оставить только единичные нуклеотидные замены (убрать вставки/делеции), подгрузить аннотацию *M. tuberculosis* H37Rv и оставить только те замены у каждого из 20 образцов, которые находятся в генах. Предложить метрику расстояния между геномами, получить матрицу расстояний и построить филогенетическое дерево.

Исходить из того, что образцы отличаются только указанными полиморфизмами. Выравнивание было произведено глобальное. В таблицах .snps в первых четырех столбцах указаны: позиция в референсе, нуклеотид в референсе (или делеция «.»), нуклеотид во втором образце (или делеция «.»), позиция во втором образце. Остальная информация не нужна.

0.2 Задача вторая

Есть таблица следущего вида:

species	feature	value
Streptococcus mitis	feature54	35.3
Neisseria macaccae	feature14	98.3
Streptococcus mitis	feature8	71.2
Neisseria macaccae	feature17	30.1
Streptococcus mitis	feature42	99.5
Streptococcus mitis	feature12	24.2
Neisseria macaccae	feature92	53.2

Взять таблицу (10^6 строк) или уже .RData объект можно по ссылке.

Для каждого признака (feature) проставить порядковый номер (rank) его значения (value) в порядке возрастания значений. Причем сделать это для каждого вида (species) отдельно. Т.е. получить таблицу следующего вида:

species	feature	value	rank
Neisseria macaccae	feature17	30.1	1
Neisseria macaccae	feature92	53.2	2
Neisseria macaccae	feature14	98.3	3
Streptococcus mitis	feature12	24.2	1
Streptococcus mitis	feature54	35.3	2
Streptococcus mitis	feature8	71.2	3
Streptococcus mitis	feature42	99.5	4

Предложить оптимальный алгоритм.

0.3 Задача третья

Есть таблица ($2 * 10^6$ строк) следующего вида:

```
element   type      value

elem1      control    14.580546
elem2      decoy      1.863077
elem3      control    15.595858
elem4      control    14.822892
elem5      decoy      8.922175
elem6      control    17.484545
```

Нужно найти такое пороговое значение T для value, при котором кол-во элементов с $value \geq T$ типа decoy будет составлять 5% от кол-ва элементов с $value \geq T$ типа control. Другими словами, если представить, что decoy — это ложные элементы, а control — действительные: «нужно определить пороговое значение value при котором $FDR \leq 0.05$» (см. рисунок).

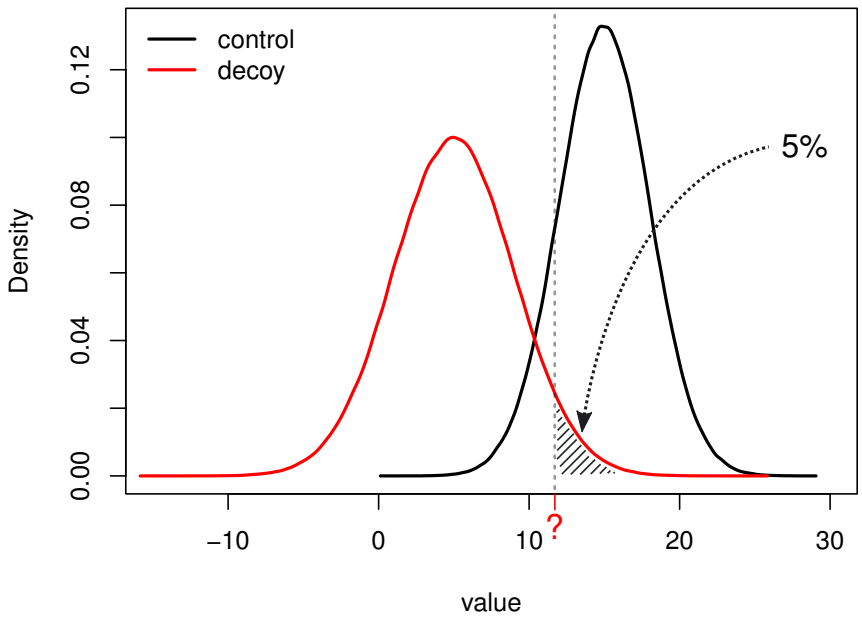

Таблица в .RData или .txt: по ссылке

Предложить оптимальный алгоритм.