Mestrado em Bioinformática
Extração de conhecimento de bases de dados biológicas

Sessão "Team Based Learning" – *Individual readiness assessment*
Março 2024

Nas questões seguintes, considere um dos *datasets* que têm vindo a ser usados nas aulas:

- *Cachexia* – carregue os ficheiros conforme indicado nos slides da aula, juntando os dados com os metadados; deverá ter disponível o objeto *cachexia* com os dados e metadados (última coluna)

1. Considere o seguinte código:

```
> pv.orig = sapply(cachexia[,1:63], function(x) shapiro.test(x)$p.value)
> sum(pv.orig < 0.001)
> pv.log= sapply(cachexia.log[,1:63], function(x) shapiro.test(x)$p.value)
> sum(pv.log < 0.01)
> colnames(cachexia.log)[which(pv.log < 0.01)]
```

a) Qual o objetivo das instruções anteriores?

b) O que conclui do resultado da sua execução?

2. Considere o seguinte código:

```
testCompound = function(ds, index)
{
  g1 = ds[ds$Muscle.loss=="cachexic",index]
  g2 = ds[ds$Muscle.loss=="control",index]
  restt = t.test(g1, g2)
  restt$p.value
}

> pvs = c()
> for (i in 1:63) pvs[i] = testCompound(cachexia.log, i)
> pvs
> rank=order(pvs)[1:10]
> names(cachexia)[rank]
```

a) Qual será o objetivo das instruções anteriores?

b) Quais os principais problemas que se poderão colocar ao código anterior? Sugira alternativas para abordar estes problemas (indique a metodologia estatística e as funções R que usaria)

c) O que teria que fazer se quisesse realizar a mesma tarefa, mas a variável "Muscle.Loss" tivesse um 3º valor distinto? Indique a metodologia estatística e as funções R que usaria.