Gisele Antonaizzi Cardoso

Sou aluna de doutorado do IB (Genética) orientada da professora Tatiana Teixeira Torres e venho trabalhando desde a iniciação científica com um grupo de moscas que apresentam hábitos alimentares contrastantes. No momento, nós queremos saber se há uma correlação entre a expressão de genes candidatos com os diferentes hábitos alimentares.

O título da minha tese é: Evolução da expressão gênica na família Calliphoridae: um modelo para o estudo do hábito alimentar.

Meus exercícios

Aqui estarão meus exercícios da disciplina: Exercícios Gisele

Proposta de trabalho final

Proposta A

A função compara.rnaseq tem como objetivo avaliar se replicatas biológicas provenientes de RNA-seq tem uma forte correlação (que é a condição ideal para análise comparativa da expressão gênica). O usuário deve apenas fornecer dois data frames contendo os valores de expressão de cada replicata. A partir disso a função faz uma análise exploratória dos dados inicial que consiste na geração de um boxplot para observar a distribuição dos dados, sendo possível nesse primeiro momento identificar diferenças exacerbadas entre os dados. Posteriormente, só temos que testar se há uma correlação positiva e significativa entre as duas replicatas com o teste de correlação de Pearson. E, por fim nós podemos fazer um teste estatístico se há variações significativas entre as replicatas usando o teste de ANOVA. Porém, o usuário terá que avaliar o qqplot dos dados (fornecido pela função) para determinar se os dados seguem uma distribuição normal.

Comentário Melina Leite

Olá Gisele, desculpe se não entendo bem os termos genéticos, você poderia explicar melhor o que é “replicata biológica”, “valores de expressão”? Por conta disso, não entendi muito bem o que são os dados iniciais, e portanto como um boxplot vai servir para observar a distribuição dos dados. O que seriam consideradas “diferenças exacerbadas” entre os dados? Enfim, acho que se você conseguir explicar melhor o que são os dados de entrada, conseguiremos avaliar a função e os dados de saída.

Resposta Gisele

Oi Melina, desculpa se não fui muito clara quanto as minhas propostas, tentarei explicar melhor! Começarei respondendo suas perguntas: Replicatas biológicas são repetições de um experimento em amostras independentes. E os valores que mencionei são os de expressão gênica que eu citei na

terceira linha da proposta. Quanto a diferença exacerbada, o que quis dizer é que olhando a distribuição dos dados em um boxplot é possível observar logo de primeira, por exemplo, se os dados apresentam ou não sobreposição, caso não haja sobreposição significa que a distribuição dos dados entre as replicatas é muito diferente (cenário não ideal). Agora vamos ver se consigo explicar a proposta de uma forma mais clara: o usuário fornece dois data frames (cada um equivale a uma replicata biológica) contendo os dados de expressão gênica de cada transcrito. Primeiramente a função, mescla as duas tabelas (merge) e logo em seguida gera um boxplot para comparar a distribuição dos dados de cada replicata (acho essa etapa essencial para entender os dados, porque muitas vezes só olhando a tabela fica difícil de compreendé-los). O importante para o usuário é que as replicatas não sejam muito diferentes, o que significaria que há um viés em seu experimento. Para testar isso podemos fazer uma correlação (eu sugiro a de Pearson) sendo que o resultado ideal é que as duas replicatas tenham uma correlação forte e significativa ($p > 0.05$ e eu consideraria um r2 a partir de 0.6, porém isso pode variar de acordo com o conhecimento que o usuário tem de seus dados). Além disso, a correlação deve ser positiva, não faria sentido se um valor aumentasse em uma relicatada enquanto diminuisse na outra. E por último, eu pensei em fazer um ANOVA (fiquei na dúvida se seria necessário) que serviria para testar se há variações significativas entre as replicatas. Mas para essa parte, é necessário saber se os dados seguem uma distribuição normal e por isso, pensei em incluir um qqplot, assim o usuário fica apto a decidir se deve considerar o resultado do ANOVA. O que esqueci de citar na proposta é que a saída da função seria em tela, pois a quantidade de dados é pequena, seria somente os dados da correlação (r2 e p-valor) e a tabela de ANOVA. E os gráficos seriam o boxplot, a correlação e o qqplot. Não sei, se esclareci todas as suas dúvidas Melina (espero que sim!).

Proposta B

A segunda proposta é bem mais simples e útil para laboratórios que trabalham com sequenciamento de nova geração, servindo somente de controle interno para o usuário comparar corridas de sequenciamento. A função info.seqs recebe como input um data frame contendo o tamanho das sequências. Usando intervalos propostos pelo usuário, a função retorna quantas sequencias existem em cada intervalo (valores reais e porcentagem) e um histograma para representar os dados.

Comentário Melina Leite

Olá Gisele, o data frame de input seriam valores de tamanho das sequências ou as próprias sequências em si? O que seria “um intervalo”? Novamente eu não consigo avaliar sua função pois me faltam informações que para você são obvias, mas para mim não. Se vc puder explicar melhor do que se tratam os dados a serem analisados, facilitará o entendimento. Obrigada.

Resposta Gisele

Oi Melina! Como eu disse na proposta, o input é um data frame somente com o tamanho das sequências. Mas se você achar que o melhor seria a entrada das sequências, posso tentar fazer desse jeito. E o intervalo fornecido pelo usuário, é o tamanho das sequências que ele deseja saber a frequência, por exemplo: sequencias maiores que 200 pares de base (pb), entre 200 e 1000 pb e maiores que 1000 pb.

> Gisele, pelo que eu entendo sua proposta A é uma função de conferência de dados. A função avaliaria as diferentes réplicas de um mesmo organismo (ou tecido, possivelmente) para verificar se elas estão parecidas o suficiente entre si. Me parece interessante e super útil.
>
> Então, pelo que eu entendi, cada replica/replicata é composta por um conjunto de valores contínuos/quantitativos. Com isso,sua função faz as seguintes coisas 1) verifica se as distribuições se sobrepõe 2) faz uma correlação 3) compara as médias de alguma forma.
>
> Me parece OK. Mas tenho sugestões: 1)eu acho que dá pra pensar em uma medida contínua de sobreposição, ao invés de sim ou não. 2) para comparar as médias, ao invés de uma ANOVA, eu sugiro calcular alguma dessas métricas aqui.
>
> Eu não acho boa ideia fazer uma função que só mostra os resultados na tela, poder salvar os resultados é sempre útil! A resposta vai ser pequena quando apenas duas réplicas de uma mesma amostra estiverem sendo comparadas. Mas acho que a função deveria aceitar múltiplas amostras, cada uma com suas réplicas. E aí a saída da função subitamente ficou grande. Então eu sugiro que você pense em qual a melhor maneira de armazenar esse tipo de informação em um formato que o usuário possa salvar.
>
> E finalmente, eu não acho que vale a pena calcular testes de significância, minha sugestão é apenas reportar índices de sobreposição/correlação/diferença (tamanhos de efeito).
>
> —-*Danilo Muniz*

Resposta Gi

Oi Danilo! Muito obrigada pelas sugestões, gostei bastante, mas acho que tenho que estudar um pouco para entender melhor o tamanho de efeito. Os índices de sobreposição são os quantis ou são outras medidas? (preciso pesquisar melhor sobre disso também!). Eu preciso reescrever a proposta ou já posso colocar a mão na massa?

Trabalho final

Encontre neste link a função da proposta um (porém com nome diferente!) e seu help Função compara.rep