Proposta de Trabalho Final

Plano A

Objetivo: Montar uma função no R que retorne análises básicas de um estudo genético de associação.

O estudo genético de associação por caso-controle é uma das abordagens mais utilizadas para testar se uma variante genética está associada à ocorrência de um fenótipo. Nessa abordagem, pacientes e controles são genotipados para polimorfismos (variantes de um mesmo locus gênico comuns na população) de interesse. Frequências estatisticamente diferentes dos polimorfismos entre pacientes e controles indicam que o polimorfismo testado está em uma região do genoma associada com o fenótipo estudado.

Plano de execução: a partir de um arquivo de entrada dos dados (arquivo em formato csv, com indivíduos em linhas e genótipos dos polimorfismos em colunas), as funções deverão retornar:

a) Frequências alélicas (2 classes esperadas, ex: A, a) de cada polimorfismo.

b) Frequências genotípicas (3 classes esperadas, ex: AA, Aa, aa) para cada polimorfismo.

c) Construção de histograma com a distribuição das frequências genotípicas observadas.

d) Teste qui-quadrado para verificar se a distribuição dos genótipos está de acordo com o equilíbrio de Hardy-Weinberg (a partir das frequências genotípicas obtidas)

e) Teste qui-quadrado para diferenças de frequências dos polimorfismos entre casos e controles

f) Cálculo de odds ratio para cada polimorfismo (no caso, mede o tamanho do efeito de cada genótipo na determinação do fenótipo)

g) Construção de Q-Q plot para representar os resultados do item e para todos os polimorfismos testados.

Plano B

Outra abordagem bastante utilizada para investigar a relação genótipo-fenótipo é a análise de correlação entre genótipos de um polimorfismo e níveis de expressão gênica (medidos pela quantidade de RNA transcrita pelo gene de interesse). Em outras palavras, investiga-se se variantes de um polimorfismo podem ter relação com níveis maiores ou menores dos produtos dos genes.

Proposta: A partir de uma tabela contendo genótipos de diferentes polimorfismos (3 classes de genótipo esperadas para cada polimorfismo, ex: AA, Aa, aa), e níveis de expressão gênica (em unidades relativas) de diferentes genes, montar um função que retorne o coeficiente de correlação de Spearman entre polimorfismos e níveis de expressão gênica. A função deve retornar também um scatter plot com linha de tendência, com níveis de expressão nas ordenadas e classes genotípicas nas

abscissas.

Comentários

PI

Me parecem boas propostas, com generalidade e factíveis. Vc teve o cuidado de definir claramente entradas e saídas, que é mesmo o primeiro passo para construir funções, ótimo! Como não sou da área, vou pedir ao nosso monitor Diogo que tb dê uma olhada.

Diogo

Também achei tranquilo. Se der tempo, pense na eficiencia das contas. é facil fazer isso ai com um monte de for, mas dependendo do tamanho do seu conjunto de dados pode ser sacal. Tente pensar vetorialmente para aumentar a eficiencia da função.

O plano B tb é interessante, principalmente se vc pensar em formas de obter matrizes de correlação robustas ao pequeno numero amostral e grande numero de parametros estimados. Veja esse artigo: Schäfer, Juliane, and Korbinian Strimmer. 2005. “A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics.” Statistical Applications in Genetics and Molecular Biology 4 (1): 32.

Apresentação do Trabalho Final

Página de ajuda

```
alleLu              package:nenhum              R Documentation

Teste genético de associação por caso-controle

Description:

     Calcula frequências alélicas e genotípicas, e testa as variantes para
equilíbrio de Hardy-Weinberg e associação alélica com as classes dos
individuos. Produz um quantile-quantile plot para o teste de associação.

Usage:
```

```
     alleLu(x, ...)

Arguments:

    x: input file (.csv)
    allele.table: data frame produzido a partir do arquivo input.
    classes.col: número de classes da coluna "disease.status". Default=2


Details:

    A partir de um input .csv contendo genótipos de n loci para diferentes
classes especificadas (ex: pacientes e controles), produz uma tabela
contendo frequências alélicas e genotípicas, p-valor para teste de
equilíbrio de Hardy-Weinberg e p-valor para teste alélico de associação.


Value:

     É criada uma matriz na worktable detalhando cada parâmetro
separadamente para cada combinação de variante analisada e classe
disease.status definida.


Warning: O arquivo input deve estar em formato csv.


Note:

     A opção "Construção de histograma com a distribuição das frequências
genotípicas observadas" foi descontinuada da função, pois mostrou-se sem
utilidade alguma para a análise e interpretação dos dados.
    A função ainda necessita de correção para múltiplos testes (para os
testes de associação).

Author(s):

     Luciano Abreu Brito (luciano.brito@usp.br)


References:

    Hatterslet AT, McCarthy MI. What makes a good genetic association study?
Lancet. 2005 Oct 8;366(9493):1315-23. Review.




Example of input file:
```

```
    project<-read.csv("projteste1.csv", header=TRUE, na.strings="NA",
as.is=TRUE, sep=";")
    alleLu<-project
```

Código da função

```
#######################################
########## Trabalho final ############
#######################################

#Chamando o arquivo
project<-read.csv("projteste1.csv", header=TRUE, na.strings="NA",
as.is=TRUE, sep=";")

#Explorando alguns dados:
head(project)
is.na(project[,-c(1:2)]) #para ver se tem NA nos genótipos
genotipos=project[,-c(1:2)] #indexando: só ficam as colunas dos genotipos
genotipos

#############função#############

alleLu<-function(allele.table, classes.col=2)
{
{
genotypes.freq<-NULL
classes<-levels(as.factor(allele.table[,classes.col]))

for (j in 1:length(classes)){


allele.subset=allele.table[allele.table[,classes.col]==classes[j],]

tab.freqs<-NULL
for(i in 3:length(allele.subset)){
freq.abs<-table(allele.subset[,i])
freq.rel<-freq.abs/sum(freq.abs)
tab.freqs<-rbind(tab.freqs,c(freq.abs,freq.rel))
}
rownames(tab.freqs)<-
paste(classes[j],colnames(allele.table[,3:length(allele.table)]),sep=".")


genotypes.freq<-rbind(genotypes.freq,tab.freqs)


}
```

```
freqs<-
cbind(genotypes.freq,genotypes.freq[,4]+genotypes.freq[,5]/2,genotypes.freq[
,6]+genotypes.freq[,5]/2)
colnames(freqs)[1:3]<-paste("Abs",colnames(freqs)[1:3],sep=".")
colnames(freqs)[4:6]<-paste("Rel",colnames(freqs)[4:6],sep=".")
colnames(freqs)[7:8]<-c(1,2)
colnames(freqs)[7:8]<-paste("Rel",colnames(freqs)[7:8],sep=".")

loci.number<-length(freqs[,1])/2 # Cria objeto com o nÃºmero de loci
analisados
control.set<-1:loci.number # Controls subset
test.set<-loci.number+control.set # Tests subset


Exp11=freqs[,7]^2*sum(freqs[1:3])
Exp12=2*freqs[,7]*freqs[,8]*sum(freqs[1:3])
Exp22=freqs[,8]^2*sum(freqs[1:3])
Genexp=cbind(Exp11, Exp12, Exp22)

Chisq_HW<-rowSums(((freqs[,1:3]-Genexp[,1:3])^2)/Genexp[,1:3])
Pvalue_HW<-1-pchisq(Chisq_HW,1)
locus<-as.factor(colnames(allele.table[,3:length(allele.table)]))
Obs1=freqs[,1]*2+freqs[,2]
Obs2=freqs[,3]*2+freqs[,2]

Obs=cbind(Obs1,Obs2)

freqs<-cbind(freqs,Pvalue_HW,locus,Obs)


Exp1<-(freqs[control.set,11]+freqs[test.set,11])/2
Exp2<-(freqs[control.set,12]+freqs[test.set,12])/2

freqs<-cbind(freqs,Exp1,Exp2)




# Teste de associação e QQplot

Chisq_assoc<-rowSums(((freqs[control.set,11:12]-
freqs[control.set,13:14])^2)/freqs[control.set,13:14]+((freqs[test.set,11:12
]-freqs[test.set,13:14])^2)/freqs[test.set,13:14])
Pvalue_assoc<-1-pchisq(Chisq_assoc,1)

freqs<-cbind(freqs,Genexp,Chisq_assoc,Pvalue_assoc)

qqnorm(freqs[,19], col="red", main="QQ plot",xlab="Esp",ylab="Obs")
QQ=qqline(freqs[,19])

print(freqs)
```

```
}
}

alleLu(project)
```

Modelo input file

projteste1.csv

Código da Função

trabalho_final_luciano.r