Juliana Emilia Prior Carnavalli

Doutoranda do Programa de Pós-Graduação em Genética do Instituto de Biociências-USP. Seu mestrado teve como foco estudo de associação de genes candidatos com obesidade em populações quilombolas, concluído em 2016. Atualmente, sob orientação da professora Carla Rosenberg, investiga as possíveis causas genéticas da deficiência intelectual em pacientes com algum grau de consanguinidade. Tem interesse pela área de bioinformática, tendo concluído alguns cursos ao longo de sua formação acadêmica, e análise de Big Data.

Curriculum Lattes

Contato

Meus exercícios

Meus exercícios

Proposta de Trabalho

Proposta A

Introdução: Eric S. Lander e David Botstein, em 1987, propuseram uma estratégia para o mapeamento de genes relacionados a traços autossômicos recessivos em famílias consanguíneas. Essa estratégia (mapeamento por homozigose) baseia-se na hipótese de que regiões adjacentes ao loco a ser mapeado devem ser homozigotas por descendência comum nos indivíduos que apresentam a característica autossômica recessiva em análise. Ou seja, indivíduos afetados por doenças autossômicas recessivas raras, em famílias consanguíneas, muito provavelmente serão homozigotos para os marcadores ligados ao loco da doença. Quando dois indivíduos compartilham um ancestral comum relativamente recente, eles apresentam regiões em seus genomas que serão idênticas por descendência. Se esses indivíduos vierem a se casar poderão, portanto, transmitir o mesmo segmento para um filho que, por sua vez, será homozigoto por descendência para o segmento em questão. Para nos referirmos a esses longos trechos de homozigose utilizamos o termo ROH (runs of homozygosity). Espera-se que, quanto mais próximo for o parentesco entre os pais mais longas sejam as ROH encontradas em seus descendentes. Isso ocorre pois houve pouco tempo para que os eventos de recombinação quebrassem esses segmentos idênticos por descendência. Consequentemente, o oposto ocorre quando o parentesco entre os pais é mais distante. Além de variarem quanto ao comprimento, as ROH podem, também, ser encontradas em quantidades diferentes dependendo do grau de consanguinidade dos pais. Globalmente, a união consanguínea mais comum é entre primos de primeiro grau, cuja prole, nesse caso, apresentará 1/16 de lócus em homozigose por origem comum. Até 2006, para o mapeamento dessas ROH eram utilizados microssatélites em grandes famílias consanguíneas com muitos afetados e, consequentemente, poucos genes foram

identificados. O primeiro grande estudo utilizando array de SNPs surgiu no ano seguinte e identificou 8 intervalos de homozigose em 76 famílias iranianas e consanguíneas com dois ou mais filhos afetados por DI não-sindrômica. Muitas dessas ROH eram longas, o que significava que centenas de genes candidatos tiveram de ser averiguados detalhadamente. O mapeamento por homozigose tem se mostrado eficaz para a identificação de novos genes candidatos da DIAR, principalmente com a chegada de novas metodologias de genotipagem como os array genômicos em grande escala e as técnicas de sequenciamento de próxima geração (next generation sequencing – NGS), que permitem a triagem simultânea de mutações em grande número de genes em muitos pacientes.

Objetivo da função: A função irá identificar em cada cromossomo quais são as regiões de homozigose. Verificará quantas são, qual o tamanho das ROHs encontradas e qual a exata posição cromossômica e irá devolver um arquivo de texto separado por tabulação contendo essas informações (qual cromossomo está, posição que começa, posição que termina, tamanho)

Entrada de dados: São dados dois arquivos texto separados por tabulação de entrada (informações estão no Help): ped e map.

Exemplo dos arquivos ped e map

Argumentos: ped= arquivo de texto ped, map= arquivo de texto map, length= tamanho mínimo da ROH (por padrão serão considerados 10Mb), het= quantidade de heterozigotos permitidos dentro da ROH (por padrão serão aceitos 2)

Saída dos dados: a função irá devolver um arquivo de texto separado por tabulação contando as seguintes colunas: IDsample= identificação da amostra, cromossomo= qual cromossomo está a ROH, posição genômica que começa a ROH, posição genômica que termina a ROH, tamanho da ROH.

> Oi Juliana, tudo bem?
>
> Sua proposta A está muito bem explicada, parabéns e obrigado! Ela talvez seja pouco genérica, mas pelo visto ela fará uma tarefa importante para geneticistas e te trará desafios interessantes (prevejo muitos FORs e IFs). Mas, como eu não entendo tanto assim da área, preciso esclarecer alguns pontos com você para imaginar quais os caminhos que você pode seguir no R, ok?
>
> Primeiro, me ajude a entender o que é o objeto PED. Cada linha é uma amostra, certo? Cada amostra é o material genético total de uma pessoa, com todos os cromossomos? O número de colunas de PED vai depender do tamanho total do material genético da pessoa?
>
> Segundo, a função comparará cada amostra e vai dizer quais são mais parecidas, em qual região está o ROH, e em cima disso você fará os outros cálculos para ter o objeto de saída que promete. Mas eu não entendi qual a relação com o objeto MAP. Existe um objeto MAP para cada pessoa? O MAP é gerado a partir do PED e é só o intermediário para

construir o objeto final que será a saída da sua função? Estou entendendo corretamente?

Veja também meus comentários na outra proposta.

Abraços! — *Rodolfo Liporoni Dias* *2017/06/02 16:21*

Oi Rodolfo, tudo bem?

Primeiro lugar obrigada pelos comentários! Creio que, por mais específica que a função vá ser, terá uma boa aplicabilidade para aqueles que trabalham com o assunto (como você mesmo pontuou). Mas vamos às suas dúvidas! Espero que eu possa resolvê-las.

No arquivo PED cada linha corresponde à uma amostra (exatamente como você supos) contendo todas as informações das genotipagens obtidas, ou seja, se eu usei um painel com 30 marcadores cada linha vai conter o genótipo do indivíduo obtido para cada um desses 30 marcadores. Cada duas colunas correspondem a UM genótipo. Por exemplo, para um marcador o indivíduo é A/A, essa informação não estará em uma única coluna, mas sim dividida em duas. Uma coluna com um A e a outra com o outro A. Então, para o exemplo dos 30 marcadores, haveria 60 colunas correspondentes aos dados genotípicos. Fui clara? Espero que sim!

O objeto MAP é um arquivo contendo a informação da localização genômica para os marcadores utilizados. O que isso quer dizer? Para o meu exemplo citado acima, eu teria um arquivo MAP me dando o endereço (cromossomo e posição genômica) para cada um dos marcadores. Esse arquivo é importante para eu poder relacionar as informações contidas no arquivo PED, pois o MAP funciona como um “catálogo” do endereço de cada marcador. É através do MAP que eu consigo o endereço genômico dos genótipos obtidos no arquivo PED.

Veja se o exemplo nesse Link fica mais claro agora com essas informações.

Beijos! Juliana Carnavalli 04/06/2017

Olá Juliana, tudo bem?

Muito obrigado pelos esclarecimentos, agora já está bem mais claro o que são os objetos PED e MAP! Eu até tinha visto este link que mandou antes, mas não tinha entendido que cada par de colunas com letras correspondia a um marcador, vulgo locus neste caso né? (embora estivesse escrito lá; você desenhou e eu entendi rsrs).

Bom, agora acho que posso opinar. Sua proposta A me parece mais interessante e desafiadora para você. Pelo o que entendi, os objetos de entrada são fixos em estrutura e o código da função será meio rígido. Que tal deixar o usuário brincar com quais partes desses objetos ele quer usar? Assim o código teria alguns desafios interessantes para você. E se a função tivesse um argumento que deixasse o usuário escolher quais marcadores dos objetos de entrada ele quer analisar? Imagino que em alguns casos o usuário pode usar PED e MAP de outras pessoas e estar interessado em apenas alguns dos marcadores coletados. Faz sentido ou estou viajando? Se fizer sentido, acho bom incluir essa opção, porque uma função é tão poderosa quanto a liberdade que ela dá ao usuário! 😀

Outra sugestão: uma vez tendo o objeto de saída que você promete dar, o que o usuário faria com eles? Tem algum teste estatístico interessante posterior? Talvez vale a pena incluir algum teste (ou mais de um) dentro da função e deixar o usuário escolher qual usar (ou se usar ou não). Outra opção é incluir um gráfico que ajudasse a visualizar melhor os resultados da função (e que pudesse ser uma saída opcional também, a critério do usuário). Tudo isso você faz criando argumentos de entrada lógica. O que acha?

Como disse, são sugestões, fica a seu critério incluir ou não. Mas recomendo muito pensar em incluir pelo menos algum outro objeto de saída além do data.frame (seja um gráfico opcional, ou o resultado de um teste opcional, ou o que você puder pensar!) para deixar a função mais turbinada.

Qualquer dúvida, não hesite em perguntar!

Abraços! — *Rodolfo Liporoni Dias 2017/06/05 15:45*

Oi de novo, Rodolfo. Tudo bem?

Mais uma vez obrigada pelas sugestões, tentarei explicar a limitação de algumas e propor alternativas para elas, ok?

Quanto à rigidez da função ela é proposital por conta da

aplicabilidade da função. O que eu quero dizer com isso? Quando estamos em busca dessas ROHs (regiões de homozigose) o interesse é global, ou seja, buscamos essas informações em todos os cromossomos e todas as posições que tiverem sido genotipadas no arquivo PED. Não é comum escolher os marcadores porque buscamos longos trechos em que esses marcadores sejam iguais em ambos os alelos (homozigotos). Minha sugestão para deixar a função menos "durona" é permitir ao usuário escolher qual cromossomo ele quer analisar. Ou seja, por padrão a função olharia tudo, mas o usuário poderia optar por verificar um único cromossomo (e a função devolveria as informações apenas do cromossomo pedido). Faz sentido? Eu apenas precisaria ver como é possível realizar isso! hahaha

Quanto aos outros aprimoramentos:

1- A ideia do gráfico é boa sim, eu tinha pensado sobre isso na ideia inicial e não sei porque acabei desistindo. Obrigada por reforçar a necessidade disso!! O que pode ser feito é um gráfico de barras por indivíduo mostrando o número de ROHs ou a média das ROHs encontradas (por exemplo, se em um individuo foram encontradas 8 ROHs, seria feito um cálculo da média desses resultados e plotado em um gráfico).

2- Quanto a testes estatísticos, não é possível realizar nenhum pois não há hipótese nula a ser testada ou dados que sirvam de comparação. A interpretação dos resultados caberia ao usuário, dependendo do que ele deseja verificar. ROHs podem ser importantes para o descobrimento de novos genes associados à doenças (como comentei na introdução), além de indicarem parentesco. Quanto maior a quantidade e o tamanho - principalmente o tamanho - dessas ROHs, maior será a chance que o indivíduo em questão seja filho de pais consanguíneos. Existe um parâmetro que mede o grau de consanguinidade dos pais a partir das ROHs do filho, mas não é algo tão simples quanto parece e creio que demandaria mais tempo para ser executado (não sou geneticista de populações, então fica um pouco mais nebuloso para mim). Então eu descartaria essa opção!

Fui clara o suficiente, Rodolfo? Se algo estiver confuso pode dizer que eu arrumo outro jeito de me fazer entender!

Novamente obrigada pela ajuda desde já!

Beijos! Juliana Carnavalli 05/06/2017

Olá Ju, tudo bem?

Vou responder só aqui desta vez. Entendi os pontos que você colocou, tudo bem não ter um teste estatístico e não poder variar os marcadores de interesse, mas gostei da sua ideia de deixar o usuário escolher o cromossomo e se quer um gráfico! Isso deixaria a função bem mais interessante.

Quanto a escolher o cromossomo que a função vai retornar, isso não será tão difícil quanto parece. Minha dica é fazer o código pensando em todos os cromossomos, e no final pensar em como dar só os resultados do cromossomo que o usuário quer. Você precisará usar algum controle de fluxo para isso (IF, WHILE, etc.).

Já quanto ao gráfico, gostei da sua ideia de gráfico de barras, mas se é possível calcular médias, que tal ter outro tipo de gráfico também, tipo um bloxplot? Seria interessante? Se tiver mais de um gráfico possível, acho legal colocar também (fazer gráfico não dá tanto trabalho assim né).

E quando sugeri a ideia do gráfico, eu pensei num desenho de mapa cromossômico, como este ou este. Você poderia desenhar os cromossomos e pintar as regiões de ROH (isso é possível né?). Assim, daria para ter uma ideia visual dos resultados do data.frame que a função vai retornar, como a distribuição dos ROH entre os cromossomos, não seria interessante??? Não sei quanto geneticistas usam esses mapinhas, mas pelo menos eu, que não sou da área, acho informativos e mais legais!

Enfim, novamente são só sugestões. Escolha qual(is) achar melhor(es) para deixar a função menos rígida e com mais opções para o usuário brincar com ela. Assim, você vai aprender mais!

Já pode mandar bala na proposta A! Qualquer problema, só me procurar por e-mail ou pessoalmente.

Abraços! — *Rodolfo Liporoni Dias 2017/06/06 15:04*

Proposta B

Introdução: O estudo de ligação baseia-se no conceito de ligação genética, que se descreve como a tendência de alelos em dois lócus distintos localizados muito próximos numa mesma região cromossômica não sofrerem recombinação entre si e, portanto, serem transmitidos em conjunto na meiose. Em geral, são utilizados para estudar a transmissão de um segmento cromossômico de uma

geração para a outra em uma genealogia e compará-la com o padrão de transmissão de uma doença. Desse modo, pode-se definir uma região cromossômica candidata a conter o gene responsável pela doença. Essa abordagem é feita sem pré conhecimento sobre a função dos genes e facilita a identificação, nas regiões cromossômicas mapeadas, tanto dos já descritos como aqueles ainda não conhecidos. Os dados de ligação são analisados utilizando um dos softwares disponíveis para isso (por exemplo, MERLIN), mas para primeiramente é necessário preparar os arquivos de entrada, um trabalho que demanda atenção e tempo.

Objetivo da função: Preparar os arquivos para a análise de ligação a partir de um arquivo de texto fornecido pelo usuário. A função também irá realizar algumas filtragens: retirar marcadores em duplicatas ou sem identificação e retirar regiões teloméricas (não informativas para esse tipo de análise). Além disso será impresso na tela quantos marcadores foram removidos por falta de informação (NAs).

Entrada de dados: Arquivo de texto (a informação da separação das colunas será informada pelo usuário). O help da função irá informar qual a ordem de colunas e quais informações devem estar contidas em cada uma delas.

Argumentos: dados=arquivo texto, sep= qual tipo de separação do dados (por padrão será considerado tabulação), mark= quantos marcadores (para que a função saiba de quantas colunas serão tiradas as informações), ped= nome que deseja salvar o arquivo, map= nome que deseja salvar o arquivo, dat= nome que deseja salvar o arquivo. Nos três últimos argumentos, se não for especificado nada, por padrão a função irá devolver os arquivos com os nomes: output.ped, output.map, output.dat

Saída dos dados: A função irá devolver três arquivos de textos separados por tabulação: ped, map e dat. Além disso irá imprimir na tela quantas duplicatas de marcadores foram removidos e quantos NAs foram encontrados.

Site de referência dos outputs da função

> Oi de novo!
>
> Proposta também clara, mas me parece mais simples que a anterior. Se entendi bem, a função pegaria os dados no formato que o usuário geralmente coleta e devolveria reorganizado do jeito necessário para entrar no programa MERLIN, é isso mesmo?
>
> Você poderia explicar melhor como seria o objeto de entrada da função para eu entender o nível de complexidade envolvido, por favor? 😃
>
> Abraços! — *Rodolfo Liporoni Dias 2017/06/02 16:50*

> Oie de novo, Rodolfo!
>
> Mais uma vez obrigada desde já pela ajuda!

A ideia da função é exatamente essa: a partir de um único arquivo fornecido pelo usuário gerar três arquivo texto separados por tabulação que serão utilizados para análise de ligação no software MERLIN.

O objeto de entrada seria uma planilha (data frame) contendo todas as informações necessárias para se gerar os arquivos PED, MAP e DAT utilizados pelo MERLIN. O usuário precisaria apenas preparar esse arquivo contendo as informações necessárias em uma ordem específica, ou seja, as informações de cada coluna seriam fixas. O que a função faria seria reorganizar essas informações, retirando os NAs e alterando os genótipos de A/A para 1/1, pois o MERLIN não reconhece as “letras”.

Eu estive pensando sobre essa função depois de elaborar a proposta e acho que gerar o arquivo MAP pode ser meio “desnecessário” porque, como é um arquivo com o endereço genômico dos marcadores, ele já estaria pronto (normalmente o arquivo MAP é fornecido pela própria empresa da plataforma de genotipagem). Então a função geraria apenas o PED e DAT a partir de um data frame fornecido pelo usuário.

Deu para entender? Qualquer coisa só gritar que eu corro aqui para responder!

Beijos! Juliana Carnavalli 04/06/2017

Oizis!

Obrigado pelos esclarecimentos novamente, Ju!

Então, acho que a proposta B está boa (precisaria ser um pouco mais complexa), mas a proposta A está muito melhor (ainda mais se você incluir alguma das sugestões que dei ali em cima). Você tem preferência por alguma?

A proposta A me parece mais desafiadora e com potencial de te trazer maior aprendizado e sedimentar melhor os conceitos discutidos ao longo do curso. Se aceitar minha sugestão, já pode começar a trabalho no código do plano A.

Qualquer dúvida, estou aqui!

Abraços! — *Rodolfo Liporoni Dias* *2017/06/05 16:05*

Oies!

Então, minha preferência também é pela proposta A, ainda mais se eu conseguir implementar algumas das suas sugestões!

Beijos! Juliana Carnavalli 05/06/2017

Trabalho Final

Página de Ajuda

```
Colocar o Help
```

Código da Função

```
Código da função
```

Arquivo da Função

Formato .R:

Formato .txt:

From:
http://ecor.ib.usp.br/ - **ecoR**

Permanent link:
http://ecor.ib.usp.br/doku.php?id=05_curso_antigo:r2017:alunos:trabalho_final:carnavalli.jep:start

Last update: **2020/08/12 06:04**