Murillo Fernando Rodrigues

Sou aluno de Mestrado do Programa de Pós-Graduação em Genética e Biologia Evolutiva do IB/USP, sob orientação do Prof. Dr. Rodrigo Cogni. Desenvolvo o projeto intitulado *Variação clinal em genes do sistema imune de* Drosophila melanogaster.

Exercícios

Exercício 1 Exercício 4 Exercício 5 Exercício 7a Exercício 7b Exercício 8 Exercício 9

Trabalho Final

O resultado do trabalho final pode ser encontrado no seguinte link: funções finalizadas

— *Murillo Fernando Rodrigues* *2016/05/13 10:55*

Proposta A

Motivação

Estudos em escala ômica (genômica, transcriptômica, etc.) muitas vezes analisam como um conjunto de genes se comporta. Existem fatores que podem confundir a interpretação do padrão observado (e.g., desequilíbrio de ligação, heterogeneidade de taxas de mutação, etc). É possível que se queira, então, delimitar um conjunto de genes controle, para poder definir se o efeito observado não é resultado de fatores de confusão.

Obbard et al. (2009), por exemplo, estudaram a seleção a longo prazo em genes do sistema imune de *Drosophila melanogaste*r. Para isso, avaliaram a proporção de substituições adaptativas em genes imunes e genes controle próximos àqueles. Os autores explicitaram os critérios adotados para definir os genes controles, que são:

1. o gene controle esteja a uma distância *x* do gene de interesse
2. o gene controle não pertença a mesma categoria funcional do gene de interesse

No entanto, falharam em explicitar como fizeram exatamente a escolha dos genes controle:

- o gene está há distância *x* à direita ou à esquerda do gene de interesse?

- e se não houver gene a uma distância *x*?
- e se o gene a tal distância já foi colocado como controle de outro gene?

Também não foi disponibilizada uma tabela com os genes de interesse e seus respectivos genes controle.

Uma das características fundamentais da ciência é a reprodutibilidade. No caso do trabalho anterior, a reprodutibilidade poderia ser alcançada se os autores tivessem elaborado um algoritmo que encontre genes controle de acordo com os critérios determinados.

Ideia para implementação

Input: data.frame contendo informações dos gene de interesse (ID, chromosome, start_location, end_location, biological_function), data.frame com informações de todos os genes nos arredores das localizações dos genes de interesse (incluindo os genes de interesse; com o mesmo header do primeiro data.frame) e a distância ótima (*x*) entre o gene de interesse e o controle (unidade em kb). [1)]

Pseudo-código: [2)]

- Pegar o *i*-ésimo gene de interesse do primeiro data.frame.
    - Subsetar o data.frame completo para apenas os genes no cromossomo de *i*. Identificar a posição de *i* nesse data.frame.
    - Enquanto dist(*i*,*i*+k) ou dist(*i*,*i*-k) =< *x*
        - Verificar a distância de *i* até o gene *i*+k e até *i*-k.
        - Salvar (ou substituir caso não seja a primeira volta no *loop*) no data.frame de *matches* uma linha com o ID do gene de interesse e a as informações do gene com distância mais próxima de *x*, desde que a função biológica seja distinta da função do gene *i* e que esse gene já não esteja presente no d.f de *matches*.
        - k=k+1
    - Se nenhum gene cumpriu esses critérios, salvar uma linha com o ID do gene de interesse de NAs nas demais colunas.
- Voltar ao primeiro passo.

Output: data.frame contendo em cada linha o ID do gene de interesse em uma coluna e cada informação do gene controle correspondente nas outras colunas. Também, serão retornados gráficos exploratórios (histograma e/ou boxplot da distribuição de distâncias entre controle e interesse) comparando com o ideal *x*.

Proposta B

Motivação

Entender diferenciação entre populações é uma questão comum em estudos evolutivos. Existem algumas métricas que permitem quantificar diferenciação, sendo **Fst** uma mais usadas para sumarizar diferenciação populacional. Além disso, o **Fst** está fortemente associado com variância intra e inter-populacional e pode revelar a ação de seleção natural – com seleção atuando distintivamente sobre populações, a variância intra-populacional diminui e inter-populacional

aumenta, o que resulta em altos valores de **Fst**.

Uma abordagem para se calcular o **Fst** é o *method-of-moments*. Em linhas gerais, compara-se a heterozigose entre populações e aquela esperada sobre Hardy-Weinberg [3)].

Ideia para implementação

Input: data.frame com estrutura ID (identificação do loco) e as frequências do alelo *p* para cada uma das populações ($freq. pop_1, freq. pop_2, \ldots, freq. pop_n)$), (opcional) data.frame com ID e pesos para cada uma das entradas do df anterior e um argumento caso queira o **Fst** global ou par-a-par (no último caso, é retornado o **Fst** médio).

Descrição da função: serão calculados as Heterozigoses intra-populacionais (Hp), as Heterozigoses médias inter-populacionais (Hm), podendo ou não ser ponderado pelo df de pesos, e a Heterozigose total sob HW (Ht) esperado para cada loco. Caso a opção par-a-par seja selecionada, Hm e Ht serão calculadas para cada par de populações.

Fórmulas:

$$H_p = 2*p_i*(1-p_i)$$

$$H_m = \frac{\sum\limits_{i=1}^{n}peso_i*H^p_i}{n}$$

$$H_t = 2*\bar{p}*(1-\bar{p})$$

$$Fst = \frac{H_t - H_m}{H_t}$$

Output: data.frame contendo o **Fst** global ou **Fst** médio de cada locos. Neste último caso, também deverá ser retornada uma matriz com os **Fst** pair-wise.

Referências:

Genetics in geographically structured populations: defining, estimating and interpreting Fst

Comentários Danilo (gruingas@gmail.com)

Gostei muito da sua primeira proposta, ela tem um objetivo claro e geral (pode ser usada pra qualquer organismo) e você sabe por onde começar. Só queria chamar a atenção para duas coisas:

1) o help tem que deixar muito claro o que cada coisa nos data.frames quer dizer. Especialmente qual a classificação de "função biológica" que você tinha na cabeça. Isso não muda o que a função faz, mas deve interferir na interpretação dos resultados

2) você tem que incluir algum conjunto de dados (real ou fictício) ou linhas de código que criem dados fictícios para que sua função possa ter um exemplo que roda.

Oi Murilo, Muito legal a sua primeira proposta! Outra coisa

> para ela é deixar o gráfico opcional de ser plotado ou não. A segunda proposta não tão estimulante e útil como a primeira. Sugiro ficar com a primeira. — *Sara*

1)

Perceba que para montar esses data.frames já é necessário que informações sejam extraídas e organizadas a partir de fontes externas. Essa será apenas uma das funções a serem construídas para lidar com o problema exposto.

2)

O algoritmo está meio bruto mesmo. Sugestões são bem-vindas!

3)

Note que essas estatísticas são similares às da ANOVA, onde heterozigose é uma medida de variação.