

Disponibilização de análise de controle de qualidade em GWAS e seleção de SNPs para exclusão de paternidade na plataforma Galaxy

Tuani Pietrobon¹
Roberto Hiroshi Higa²

A computação tem se tornado uma ferramenta essencial no campo de pesquisa científica. A plataforma Galaxy é uma dessas ferramentas, que permite integração de dados biológicos e meios para realizar análises computacionais dos mesmos (GOECKS et al., 2010). O Galaxy é um software aberto implementado em linguagem Python, e em sua comunidade estão inclusos desenvolvedores da ferramenta, usuários, e organizações que dispõem seus próprios programas e exemplos. Os usuários podem submeter programas de diversas linguagens computacionais como o R, por exemplo, que é o programa utilizado neste trabalho.

Uma das características mais importantes do Galaxy é que seus usuários não precisam saber programar e nem aprender detalhes da implementação dos programas disponíveis, o que torna a biologia computacional mais acessível e auxilia pesquisadores que não têm experiência com programação. O que facilita seu uso é o fato de ser um sistema *workflow* em que devem ser construídas análises computacionais com múltiplos passos, como se fosse uma receita, em que é disponibilizada uma página explicando como manipular os dados e que passos realizar ordenadamente. Na Embrapa Informática Agropecuária, o Laboratório Multiusuário de Bioinformática da Embrapa (LMB) (EMBRAPA, 2013a) utiliza o Galaxy para prover a seus usuários acesso a diferentes ferramentas de análise de dados genômicos.

¹ Universidade Estadual de Campinas – tuani.pietrobon@colaborador.embrapa.br

² Embrapa Informática Agropecuária - roberto.higa@embrapa.br

O Galaxy também permite que novos procedimentos de análise de dados, implementados em diferentes linguagens, sejam incorporados e disponibilizados para seus usuários como uma nova ferramenta passível de ser utilizada na construção de *workflows*. O objetivo desse trabalho é explorar essa característica do Galaxy para tornar disponíveis duas análises, escritas em linguagem R, desenvolvidas em projetos financiados pela Embrapa. Um deles é um conjunto de procedimentos utilizados na análise de controle de qualidade em estudos de associação genômica ampla – GWAS (ZIEGLER et al., 2008) e o outro um procedimento para seleção de SNPs para teste de exclusão de paternidade. Esse trabalho será realizado em duas etapas, primeiro organizando-os em pacotes R e, então, construindo os arquivos XML que descrevem os procedimentos da análise passo a passo, conforme exigido pelo Galaxy. Esse procedimento ficará disponível na implementação Galaxy do LMB (EMBRAPA, 2013b).

Referências

- EMBRAPA. Laboratório Multiusuário de Bioinformática. **Sobre o LMB**: Laboratório Multiusuário de Bioinformática. Campinas: Embrapa Informática Agropecuária, [2013a]. Disponível em: <<https://www.agropediabrasilis.cnptia.embrapa.br/web/lmb/home>>. Acesso em: 12 dez. 2013.
- EMBRAPA. Laboratório Multiusuário de Bioinformática. **Galaxy, instalação do LMB**. [2013b]. Disponível em: <<https://www.lmb.cnptia.embrapa.br/galaxy/>>. Acesso em: 27 set. 2013.
- GOECKS, J.; NEKRUTENKO, A.; TAYLOR, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. **Genome Biology**, v. 11, n. 8, Aug. 2010. 11:R86. Doi:10.1186/gb-2010-11-8-r86.
- ZIEGLER, A.; KÖNIG, I. R.; THOMPSON, J. R. Biostatistical aspects of genome-wide association studies. **Biometrical Journal**, Berlin, v. 50, n. 1, p. 8-28, Feb. 2008. Doi:10.1002/bimj.200710398.