Utilização da plataforma Galaxy na análise de dados de RNAseq

Luis Augusto Eijy Nagai[1]
Poliana Fernanda Giachetto[2]
Adhemar Zerlotini Neto[2]

Galaxy é uma plataforma baseada em web, open source, utilizada para a criação e execução de uma série de *workflows* em bioinformática (GIARDINE et al., 2005). Este trabalho teve como objetivo avaliar a plataforma Galaxy na análise de dados de RNA-seq, uma metodologia de sequenciamento de transcritos (moléculas de RNAm) que utiliza as novas tecnologias de sequenciamento (NTS). As NTS, responsáveis por uma revolução no campo das ciências genômicas, caracterizam-se pela geração de um grande volume de dados, com custo bastante reduzido quando comparadas à metodologia convencional de sequenciamento. Apesar das inúmeras vantagens, a análise dos dados gerados não é trivial, demandando elevada capacidade computacional e a utilização de ferramentas de bioinformática complexas, ainda em fase de consolidação. O projeto Galaxy integra uma série de ferramentas de bioinformática, incluindo o *pipeline* Tuxedo (TRAPNELL et al., 2012) para a análise de RNA-seq, que inclui os programas Bowtie, Tophat e Cufflinks. Nesse *pipeline*, o TopHat executa o alinhamento das sequencias geradas (*reads*) contra um genoma referência e o Cufflinks utiliza os arquivos do mapeamento para montar as reads em transcritos e para estimar o seu nível de expressão, baseado em um índice denominado FPKM (*reads* por kilobase, por milhão de *reads* mapeadas, do inglês, *Fragments Per Kilobase of transcript per Million fragments mapped*). O Cufflinks é uma ferramenta de análise de expressão diferencial que compara os índices FPKM dos diferentes tratamentos experimentais

[1] Biotecnologia/UFSCar, eijynagai@hotmail.com
[2] Embrapa Informática Agropecuária, {poliana.giachetto, adhemar.zerlotini}@embrapa.br

e identifica alterações significativas no nível de expressão dos transcritos entre os tratamentos, por meio de um rigoroso teste estatístico. A utilização desse *pipeline* tem mostrado grande aceitação pela comunidade científica, o que pode ser observado por meio de uma série de estudos recentes de transcriptomas em humanos, plantas e animais (HUANG et al., 2012; REITZ e tal., 2012; ZHANG et al., 2012). Um vez obtidos os resultados da análise de expressão diferencial, o pacote do R CummeRbund os transforma em gráficos e figuras prontas para a publicação. A plataforma Galaxy e o *pipeline* Tuxedo foram avaliados por meio da análise de 200Gb de dados, gerados pela plataforma Illumina de sequenciamento. Os dados foram obtidos a partir de 2 experimentos de RNA-seq distintos, com o objetivo de identificar genes diferencialmente expressos entre bovinos de diferentes raças, com características distintas de maciez da carne (carne dura x carne macia) e entre caprinos, resistentes ou não, a parasitas gastrointestinais. A utilização do *pipeline* em uma interface web intuitiva, permitiu a análise dos dados por pesquisadores da área biológica, sem conhecimento avançado em computação, em menos de 2 semanas. O projeto Galaxy possui uma comunidade colaborativa e em constante crescimento, com tutoriais disponíveis e fácil instalação em UNIX/Linux. Apesar da análise de RNA-seq gerar arquivos bastante grandes, a plataforma Galaxy permitiu uma fácil visualização dos dados por meio de um *browser* de visualização de dados genômicos e ferramentas de filtro de dados, que possibilitam a seleção e classificação destes. Concluindo, foram obtidos resultados satisfatórios em um curto período de tempo, por não especialistas em computação e com pouco treinamento em bioinformática. O *workflow* simplificado, juntamente com uma reduzida curva de aprendizado, são pontos relevantes, que podem motivar a utilização do Galaxy para a análise de RNA-seq por novos usuários.

Agradecimentos

Embrapa Informática Agropecuária, Embrapa Pecuária Sul, Embrapa Caprinos e Ovinos.

Referências

GIARDINE, B; RIEMER, C; HARDISON, R. C; BURHANS, R.; ELNITSKI, L.; SHAH, P.; ZHANG, Y. BLANKENBERG, D.; ALBERT, I.; TAYLOR, J.; MILLER, W.; KENT, W.J.; NEKRUTENKO, A.. Galaxy: a platform for interactive large-scale genome analysis. **Genome Research**, v. 15, n. 10, p. 1451-1455, Oct. 2005.

HUANG, W.; NADEEM, A.; ZHANG, B.; BARBAR, M.; SOLLER, M.; KNATLB, H. Characterization and comparison of the leukocyte transcriptomes of three cattle breeds. **PLoS ONE**, v. 7, n. 1, 2012. e30244. doi:10.1371/journal.pone.0030244

REITZ, MU; BISSUE, JK; ZOCHER, K.; ATTARD, A.; HÜCKHELHOVEN, R.; BECKER, K.; IMANI, J.; EICHMANN, R.; SCHÄFER, P. The subcellular localization of tubby-like proteins and participation in stress signaling and root colonization by the mutualist Piriformospora indica. **Plant Physiology**, v. 160, n. 1, p. 349-364, Sept. 2012.

TRAPNELL, C.; ROBERTS, A.; GOFF, L.; PERTEA ,G.; KIM, D.; KELLEY, D. R.; PIMENTEL, H.; SALZBERG, S. L.; RINN, J. L.; PACHTER L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. **Nature Protocols**, v. 7, n. 3, p. 562-578, Mar. 2012.

ZHANG, L. Q.; CHERANOVA, D.; GIBSON, M.; DING, S.; HERUTH, D. P.; FANG, D.; SHUI QING YE. RNA-seq reveals novel transcriptome of genes and their isoforms in human pulmonary microvascular endothelial cells treated with thrombin. **PLoS ONE**, v. 7, n. 2, 2012. e31229. DOI:10.1371/journal.pone.0031229.