Construção de um *pipeline* para identificação e análise de CNVs utilizando dados de *chips* de genotipagem de SNPs

Fernanda Cristina de Paiva Pereira[1]
Poliana Fernanda Giachetto[2]

Variações no número de cópias *Copy Number Variations* (CNVs) podem ser definidas como regiões genômicas onde o número de cópias do DNA difere entre 2 ou mais indivíduos de uma população. Em humanos, observa-se que algumas dessas CNVs são responsáveis pela variabilidade fenotípica, incluindo a susceptibilidade a doenças (BECKMANN et al., 2007). Várias publicações têm relatado o efeito das CNVs na expressão gênica e na associação com doenças complexas (ROVELET-LECRUX et al., 2006). Em animais de produção, a caracterização dessa variação genética é um passo importante na identificação de genes ou regiões genômicas ligadas a características fenotípicas, particularmente as de importância econômica. A Empresa Brasileira de Pesquisa Agropecuária (Embrapa), tem adotado amplamente a tecnologia dos *chips* de genotipagem de SNPs em rebanhos bovinos para a utilização dos dados obtidos em estudos de associação genética em larga escala. Esses estudos têm sido utilizados como ferramentas em programas de melhoramento animal; cujo foco é a melhoria da qualidade de carne e a obtenção de animais mais resistentes a endo e ectoparasitas, entre outros; para a identificação de marcadores moleculares para essas características. Recentemente, com o emprego dos chips de alta densidade de SNPs, metodologias que permitem a sua utilização na identificação de *Copy Number Variation* (CNVs) foram desenvolvidas (HENRICHSEN et al., 2009). Assim, o objetivo desse estudo foi construir um pipeline de bioinformática para a identificação e análise de CNVs a par-

[1] Sistemas de Informação/PUC-Campinas, fernandapaiva.pucc@gmail.com
[2] Embrapa Informática Agropecuária, poliana.giachetto@embrapa.br

tir dos dados gerados pelos chips de genotipagem de SNPs da plataforma Illumina, o qual tem sido largamente utilizado na genotipagem de rebanhos bovinos da Embrapa. Para a construção do *pipeline*, foram utilizados dados de 400 animais (Canchim), participantes de um programa de melhoramento da Embrapa Pecuária Sudeste, genotipados com o BovineHD BeadChip (Illumina). O *pipeline* foi baseado na utilização da ferramenta PennCNV (WANG et al., 2007) para a identificação das CNVs a partir dos dados brutos de intensidade de sinal gerados pela metodologia de genotipagem; pela ferramenta ANNOVAR (WANG et al., 2010), para anotação das CNVs identificadas; e programas e *scripts* desenvolvidos em Perl, para a conversão dos arquivos de entrada e visualização dos arquivos de saída. Analisando os dados acima citados no *pipeline* proposto, um total de 5.684 CNVs foram detectadas em 192 amostras (que restaram após a utilização de filtros de qualidade) de DNA de gado Canchim, com um tamanho médio de 578.159bp. Nós estamos, agora, identificando as *Copy Number Variariation Regions* (CNVRs), por meio da junção de CNVs que se sobrepõem entre as amostras e identificando os genes e regiões regulatórias presentes nessas regiões, por meio da ferramenta ANNOVAR. Para cada gene identificado será atribuído um termo de Ontologia Gênica (GO), para a identificação daqueles enriquecidos entre as CNVs. Os próximos passos incluem a visualização das CNVs identificadas em um *browser* visualizador de genomas (Gbrowse) e a inclusão do presente *pipeline* na plataforma Web Galaxy, para ampla utilização pelos técnicos da Embrapa envolvidos nos projetos que visam a identificação e análise de CNVs e também de toda a comunidade científica da área.

Agradecimentos

Embrapa, CNPq (PIBIC)

Referências

BECKMANN, J. S.; ESTIVILL, X.; ANTONARAKIS, S. E. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. **Nature Reviews Genetics**, v. 8, p. 639-646, 2007.

HENRICHSEN, C. N.; CHAIGNAT, E.; REYMOND, A. Copy number variants, diseases and gene expression. **Human Molecular Genetics**, v. 18, p. R1-8, 2009.

ROVELET-LECRUX, A.; HANNEQUIN, D.; RAUX, G.; et al. APP locus duplication causes autosomal dominant early-onset alzheimer disease with cerebral amyloid angiopathy. **Nature Genetics**, v. 38, p. 24–26, 2006.

WANG, K.; LI, M.; HAKONARSON, H. Annovar: functional annotation of genetic variants from high-throughput sequencing data. **Nucleic Acids Research**, v. 38, n. 16, Sept. 2010. e164. Doi:10.1093/nar/gkq603.

WANG, K.; LI, M.; HADLEY, D.; LIU, R.; GLESSNER, J.; GRANT, S.; HAKONARSON, H.; BUCAN, M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. **Genome Research**, v. 17, p. 1665-1674, 2007.