CNV study in Nelore with PennCNV software: the control files

Vinicius H. Silva * [1], Luciana C. A. Regitano[2], Poliana F. Giachetto[3], Fábio Pértille[1], Luiz L. Coutinho[1]
* Master's student, ESALQ/USP; 11 Pádua Dias Ave.; Piracicaba, São Paulo, Brazil;
[1]ESALQ/USP, Piracicaba, São Paulo; [2]EMBRAPA Pecuária Sudeste, São Carlos, São Paulo;
[3]EMBRAPA Informática Agropecuária, Campinas, São Paulo
* viniciushs@usp.br

Nelore is a breed of extreme importance to beef production in the Brazilian agribusiness. The systematic study of Nelore's genome can contribute in a unique way for the development of zootechnical traits of interest. These traits, when associated with specific polymorphisms in genome, can be improved with higher accuracy in breeding programs. There are several types of polymorphisms, and one that has been receiving increasing attention is Copy Number Variation (CNV). CNVs are structural variations in genome, which are represented by deletions, duplications, and translocations inter or intra-chromosomal, comprising segments larger than one kb. One way to analyze these genomic events is through SNP-chips. From SNP-chip data, CNV inference uses two input parameters obtained from genotyping reports: log R Ratio (LRR) and B Allele Frequency (BAF) of each sample. The software most frequently used for CNV analysis is the PennCNV, that uses two control files for CNV inference: gcwaveBovineHD (percentage of guanine and cytosine (GC), 500 bp upstream and 500 bp downstream of each SNP) and NelorepopBovineHD.pfb (BAF for each SNP, calculated from population under study). The first objective of this study was to generate, these two specific controls for future studies of CNVs in cattle (which will use Illumina BovineHD® SNP-chip and PennCNV software). One second objective was to compare, by qui-square test, the number of low quality samples (LQS) to waviness factor values in two in silico assays (CNV call step): one with LRR adjusted samples (Gwadjusted) and another without LRR adjusted samples (Gwstandard). The waviness factor depends on the LRR values, and these values can be adjusted when GC percentage, in the 1 Mb window around the SNP, is known. Were analyzed data from 671 male Nelores genotyped with Illumina BovineHD® SNP-chip (approximately 770 thousand SNPs per animal). The first file generated was gcwaveBovineHD, based on bovine reference genome, assembly UMD_3.1. The GC percentage of each SNP was estimated by bedtools software, with the function " bedtools nuc ". The second file generated was the NelorepopBovineHD.pfb control, using the same reference genome and chip that was used in the first, taking into account our population of 671 animals. This file was produced by the " compile_pfb.pl " function from PennCNV. To adjust LRR values in Gwadjusted assay gcwaveBovineHD control file was applied to each sample of our population (" genomic_wave.pl -adjust " function, also from PennCNV). In CNV call step, these adjusted samples were applied to Gwadjusted assay, and samples not adjusted in Gwstandard assay. The CNV call step was run in PennCNV by " detect_cnv.pl -test " function and in " -pfb " argument was employed the NelorepopBovineHD.pfb control file, to both assays described above. The number of LQS in the Gwadjusted and Gwstandard was 24 and 110 (from a total of 671 samples), respectively. This frequency of LQS between assays had a significant difference in qui-square test (α " = " 1%), demonstrating the importance of sample adjust prior to analysis.

Keywords: CNA, genomic, DNA, cattle, standards, bioinformatics