Design of a low-density SNP chip for *Bos indicus*: GGP indicus technical characterization and imputation accuracy to higher density SNP genotypes

J.B.S. Ferraz¹, X. Wu²,³, H. Li²,³, J. Xu²,⁴, R. Ferretti², B. Simpson², J. Walker², L.R. Silva², J.F. Garcia⁵, R.G. Tait Jr.², S. Bauck²

¹Dept of Veterinary Medicine, College of Animal Science and Food Engineering, University of São Paulo, Pirassununga, Rua Duque de Caxias Norte, 225, 13635-970, Pirassununga, SP, Brazil jbferraz@usp.br (corresponding author)
²Biostatistics and bioinformatics, Neogen GeneSeek, Lincoln, NE 68504, USA
³Department of Animal Sciences, University of Wisconsin, Madison, WI 53706, USA
⁴Department of Statistics, University of Nebraska, Lincoln, NE 68583, USA
⁵Dept. of Animal Production and Health, College of Veterinary Medicine, UNESP, 16050-680, Araçatuba, SP, Brazil

Summary

The majority of commercial bovine SNP chips were developed based on information from *Bos taurus* animals, which results in a considerable percentage of the SNP being monomorphic in *Bos indicus* populations. Although costs of genotyping are decreasing over time, low-cost SNP chips need to be developed for broader use in *Bos indicus*, a species or subspecies which account for a large proportion of beef production in tropical and subtropical regions, and beef trade around the world. A new low-density (LD) GeneSeek Genomic Profiler SNP chip was developed specifically for *Bos indicus* (GGP indicus), which consists of 34,000 SNPs optimally selected by a multiple-objective, local-optimization (MOLO) algorithm, plus 1,000 SNPs reserved for those on Y chromosome, “causative” mutations for a variety of economically relevant traits, genetic health conditions, and international society for animal genetics (ISAG) globally recognized parentage markers for those breeds of cattle. SNPs on this new LD *Bos indicus* SNP chip were selected from a pool of 788,879 SNPs genotyped on 40,092 Zebu cattle (and crossbreds) of more than seven breeds, which included Australian Brahman, Droughtmaster, Guzerath, Gyr, Nellore (including polled Nellore), Santa Gertrudis, and Tropical Composite. Of them, a majority (69.32%) of the reference animals were from the Nellore breed, the most important breed in the Brazilian beef industry. The results showed that the new LD SNP chip was highly informative and the selected SNPs were well positioned on each chromosome. Relative to Illumina BovineHD SNP Chip, weighted minor allele frequency (WMAF) increased from 0.1829 to 0.4293 for the selected 35K SNPs and average locus-wise Shannon Entropy, which measured the average locus-wise information of SNPs, increased from 0.6019 to 0.9674. Average maximum gap by chromosome on the new 35K SNP chip was 89.44% as much as the gap on the existing GeneSeek GGP high-density *Bos indicus* (HDi) SNP...
chip. The overall imputation accuracy was 96.9% and 98.7%, respectively, when imputed to 80K SNPs on the GeneSeek GGP HDi chip and 777K SNPs on the Illumina Bovine HD SNP chip. In conclusion, this new LD Bos indicus (LDi) SNP chip represented a successful effort to leverage existing knowledge and genotype resources toward the public release of a cost-effective LDi SNP chip with good imputation accuracy to HD genotyping platforms. This new LDi SNP chip replaces the existing Z chip and it supplements GGP HDi (80K) SNP chip and will be available in late 2017.

**Keywords:** Bos indicus, cattle, genotype imputation, low-density SNP chip

**Introduction**

SNPs are useful genetic markers to test their association with diseases or quantitative traits of economic relevance, and SNP chips are useful for large scale genotyping to facilitate genetic improvement of farm animals and plants (Meuwissen et al., 2001). In Bos indicus cattle, they have been widely used for quantitative trait loci (QTL) mapping, genome-wide association studies and genomic selection (e.g. Santana et al., 2014; Magalhães et al., 2016, Magnabosco et al., 2016). Bos indicus SNP chips are available in the market, such as GeneSeek Genomic Profiler high-density Bos indicus (GGP HDi) SNP chip (80K SNPs) and GGP low-density Bos indicus SNP chip, also referred to as Z chip, but they were designed with both Bos taurus and Bos indicus content, which were not optimally designed for Bos indicus cattle, because of limited genotype data available from Bos indicus cattle.

Bos indicus cattle are well adapted to withstanding high temperatures, and raised throughout the tropical and subtropical countries, both as pure indicus or crossbred with taurine cattle. In Brazil, Zebu cattle (which are mostly of the Nellore breed) comprise more than 80% of the beef cattle (Magnabosco et al., 2016), because of their tolerance to the tropical climate and high resistance to ectoparasites. Although Zebu cattle have an important role in beef production in Latin America, their success depends on a number of factors, of which a critical one is the availability of SNP chips that are appropriately designed for Bos indicus cattle (Montaldo et al., 2012). Nevertheless, genotype data for Zebu animals have increased substantially in the past few years, which has provided sufficient genotype data to re-design Bos indicus SNP chips.

The aim of this study was to use existing databases of genotyped of Bos indicus cattle to develop a specific low-density Bos indicus SNP chip, and to evaluate its imputation accuracy to the content of existing higher-density SNP chips.

**Materials and methods**

**Genotype data**

Genotype data came from 40,092 animals from Bos Indicus and Bos Indicus x Bos Taurus cross cattle breeds, which included mainly Brahman, Guzerath, Gyr, Nellore (69% of data), and crossbreds like Santa Gertrudis, Droughtmaster and Tropical composites. These animals were genotyped on the commercial SNP chips such as Illumina Bovine HD SNP chip, GeneSeek GGP HDi SNP chip, GGP LD V4 SNP chip, Illumina SNP50 SNP Chip, Z-Chip and GGP-F250 SNP
After data cleaning and quality control, 788,879 polymorphic SNPs on chromosomes 1-29 and X remained for the subsequent selection of LD SNPs. Allele frequencies were computed for each cattle breed and breeds were arbitrarily weighted for selection. Weighted minor allele frequency – WMAF - was computed across breeds with the weights defined as follows: Brahman (0.15), Guzera (0.15), Gyr (0.15), Nellore (0.35), DroughtMaster (0.05), Santa Gertrudis (0.05), Tropical Composite (0.05), and other Indicus (0.05). GGP indicus average minor allele frequency (MAF) varied from 0.1566 (Guzerath) to 0.2643 (Droughtmaster). WMAF across breeds ranged between 0 and 0.4758, with the mean being 0.1815. WMAF had the largest correlation with that of Nellore cattle (0.944-0.952) and the least correlation with Santa Gertrudis (0.371-0.412).

Selection of LD SNPs

LD SNPs were selected using the MOLO algorithm in the “selectSNP” package (Wu et al., 2016). The objective function, \( I \), maximized the adjusted SNP information (Shannon entropy) and minimized gap lengths. The constrains included weighted minor allele frequency (WMAF), location distribution of SNPs, obligatory SNPs, number and size of gaps on the map of each chromosome.

SNP information were measured locus-wise and adjusted for the uniformness of their distribution. Putative map distribution of SNPs was initialized uniformly and then realized by optimally selected SNPs. Gaps were minimized given the number of SNPs on each chromosome. The SNP quality and fidelity criteria, such as historical call rate and Mendelian inconsistency rates, were resolved prior to optimization and hence SNP that did not meet the quality criteria were not included as candidates by the MOLO algorithm. Because the objective function in the equation was highly non-linear and solutions to this optimization problem were not analytically available a heuristic search algorithm was used to find local optima which approximated the global optimum.

Assessment of imputation accuracy

Imputation of SNP genotypes on the new GGP indicus (35K) SNP chip to those on the GGP HDi (80K, \( N = 1,858 \)) SNP chip and Illumina BovineHD (777K, \( N = 2,358 \)) SNP chip, respectively, were conducted in a Nellore cattle population using the FImpute package (Sargolzaei et al., 2014). Imputation accuracy was computed as the percentage of correctly imputed genotypes for all imputed SNPs. In the same way, imputation error rate is the percentage of incorrectly imputed genotypes for all SNPs.

Results

Characterization of the new LDi SNP chip

There were 34,000 SNPs which were optimally selected from the 788,879 candidate SNPs, in addition to 1,000 slots reserved for obligatory SNPs, which are mostly related to functional mutations or used for parentage testing, as recommended by ISAG (2016). SNPs on this new GGP indicus SNP chip were approximately uniformly distributed and had higher MAF than the GGP HDi SNP chip. In the animals used for the SNP chip development, average WMAF was
from 0.1649 for SNPs on the GGP HDi chip and it was 0.4318 for SNPs in this new GGP indicus chip.

**Imputation accuracy by breeds**

GGP HDi has 34,090 SNPs in common with the new GGP-LDi SNP chip. The fold increase of number of SNP imputed was 6.1 times better when 80K SNP genotypes were imputed and 22.8 times when 777K SNP genotypes were imputed, both from their respective common set of SNP with this new GGP indicus SNP chip. Across all the SNPs, average accuracy was 96.9% with imputation to 80K SNPs and more than 98% of SNPs had >90% imputation accuracies. For imputation to 777K SNPs, average accuracy was 98.7% with more than 99% of SNPs having >90% imputation accuracy and more than 98% of SNPs having >96% accuracy. Our results of imputation accuracies were within a comparable range of those obtained by Boison et al. (2015), who reported accuracies of between 0.888 and 0.993 when SNP genotypes were imputed from 3K, 7K, 20K, 50K, and 80K, respectively, to 777K SNPs in Gyr (*Bos indicus*) dairy cattle. Their results showed that imputation accuracy increased with the SNP density of the LD chip but the gain in accuracy became relatively smaller when LD chip size reached 15K SNPs and beyond because imputation accuracy was already high at this density (Carvalheiro et al., 2014).

**Conclusion**

The new GGP indicus SNP chip represents a successful effort to leverage existing knowledge and genotype resources toward the public release of a cost-effective LD SNP chip specifically for *Bos indicus* cattle with good imputation accuracy to HD SNP genotypes.

**List of References**


Montaldo HH, Casas E, Ferraz JBS, Vega-Murillo VE, and Román-Ponce SI. 2012.

