

Optimizing Bovine SNP discovery efforts

Lakshmi K. Matukumalli¹, John J. Grefenstette², Tad S. Sonstegard³,
Curtis P. VanTassell⁴.

Keywords: SNP, Bovine Genome, Machine Learning, Polymorphisms, QTL, QTN

1 Introduction

Cow genome sequencing is underway at Baylor College of Medicine (BCM) sequencing center and will be completed in the next few months. A recent white paper indicated a goal to identify 100,000 SNP for use in identification and mapping of QTL regions [1]. In September 2004, the preliminary assembly of the cow (Btau_1.0), Hereford breed, using whole genome shotgun (WGS) reads with an average of three fold coverage was released. BCM predicted 15,000 *in silico* SNP by comparing the Hereford preliminary assembly with traces from other breeds. Those SNP are currently being validated at the U.S. Meat Animal Research Center (MARC), ARS-USDA. The SNP discovery efforts by the consortia are currently randomized in the assembly and are optimized only for maximizing the experimental SNP confirmation.

Currently there are a total of 4400 bovine SNP at dbSNP with only a 60 of them validated as compared to 5 million validated SNP in humans. Other SNP discovery efforts in bovine genome were from the interactive bovine in silico SNP (IBISS) database that predicted about 17,000 SNP by clustering publicly available EST and mRNA sequences. Following the human genome sequencing efforts, SNP discovery and HapMap generation are likely to be the next important focus areas for the bovine genome sequencing consortium. SNP discovery efforts will facilitate LD estimation across the genome, HapMap generation, population structure determination and ability to perform whole genome association studies.

2 SNP Discovery Optimization strategy

As compared to the human SNP consortium [2] that performed large-scale high density coverage, the funding resources for the bovine genome will be limited and hence the SNP discovery efforts have to be optimized to maximize the benefits. In this abstract we summarize our SNP discovery optimization schema to identify among the high quality SNP in the population that has possible effects on regulating the gene function or altering the protein structure.

Initially, a genome-wide scan of bovine SNP is performed by an approach similar to Guryev et al [3]. SNP validation efforts will be focused in the QTL regions and only for the SNP that can affect either gene function or protein structure. All the bovine sequences along with the Phred quality scores from WGS, BAC ends and many EST are available in the trace archive. The preliminary

¹ Bioinformatics and Computational Biology, George Mason University, 10900, University Blvd, MSN 5B3, Manassas VA 20110, USA, E-mail: lmatukum@gmu.edu

² Bioinformatics and Computational Biology, George Mason University, 10900, University Blvd, MSN 5B3, Manassas VA 20110, USA, E-mail: jgrefens@gmu.edu

³ Bovine Functional Genomics Laboratory, US Department of Agriculture, ARS, Beltsville Agricultural Research Center, Beltsville, MD 20705, USA E-mail: tads@anri.barc.usda.gov

⁴ Bovine Functional Genomics Laboratory, US Department of Agriculture, ARS, Beltsville Agricultural Research Center, Beltsville, MD 20705, USA. E-mail: curtvt@anri.barc.usda.gov

bovine genome assembly is used as an anchor to build an assembly from the individual traces that have significant match to the assembly. Similar to the NQS method [3], all the variations observed at poor quality bases or in poor quality neighborhoods are eliminated. To account for the SNP observed due to paralogs and related gene families, all the SNP occurring within a breed are also eliminated in this analysis. The remaining SNP then are categorized based on the assembly annotation as rSNP – Regulatory SNP (promoter / enhancer elements), cSNP – non-synonymous SNP (amino acid change), sSNP – synonymous SNP (no amino acid change), tSNP – creates a stop codon in the reading frame, iSNP – Intron SNP or gSNP – SNP in the genomic region.

The SNP discovery efforts are further focused to identify SNP implicated in the previously known QTL regions to identify the QTN candidates among the rSNP/tSNP/cSNP. These predicted SNP are then experimentally tested in the Beltsville Agricultural Research Center (BARC) Dairy Cattle Diversity Panel to verify the predicted SNP and then can be used to fine map QTL regions. We have applied this methodology so far to identify and confirm SNP in neutrophil genes that were differentially expressed at parturition.

3 Figures and tables.

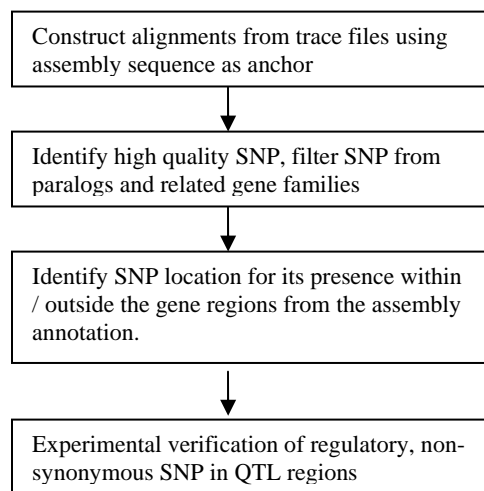


Figure 1: Optimization Schema for Bovine SNP discovery

References

- [1] Gibbs, R., Weinstock, G., Kappes, S., Loren, S., and Womack, J. 2004. White paper on bovine genomic sequencing initiative (<http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/BovineSEQ.pdf>)
- [2] Thorisson G. A., and Stein L.D. 1998. The SNP Consortium website: past, present and future. *Nucleic Acids Res.* 31:124-127.
- [3] Altschuler D., Pollara V. J., Cowles C. R., et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 2000; 407:513-516.
- [4] Guryev V., Berezikov, E., Malik, R., Plasterk R, H., and Cuppen E. 2004. Single nucleotide polymorphisms associated with rat expressed sequences. *Genome Res.* 14:1438-1443.