

# Variation Detection in the Bos Taurus Genome using Novel In-Solution Exome Capture

Jon R. Armstrong<sup>1</sup>, Jarret Glasscock<sup>1</sup>, Matthew McClure<sup>2</sup>, Curt Van Tassell<sup>2</sup>, Tad Sonstegard<sup>2</sup> and Lakshmi Matukumalli<sup>3</sup>

1. Cofactor Genomics, St. Louis, MO,
2. USDA, Agricultural Research Science, Bovine Functional Genomics Laboratory, Beltsville, MD,
3. USDA, National Institute of Food and Agriculture, Washington, DC

The ability and capacity to investigate bovine genetic variation is achievable now that the genome is sequenced, and massively parallel next-generation sequencing technologies take us one step closer to assessing the genetic variation between individual animals and diseases. However, considering the complexity, size, and repeat content of the bovine genome, there is an important need for new technologies that target and extract a defined subset of the genome for variation discovery.

Historically polymerase chain reaction (PCR) has been the predominant procedure for the selection and enrichment of relevant genomic regions. However, PCR is labor intensive, expensive at large scale, and failure-prone. More recently, genotyping arrays have been used to assess genetic variation, however, to be effective they require prior knowledge about the location of genetic variants. To address these issues, we have developed a targeted genome capture strategy that specifically targets and isolates, in a parallel fashion, exons and regulatory regions of the bovine genome for resequencing and SNP discovery. The capture set contains probes designed against 222,782 regions, representing approximately 46 million bases, of the bovine genome.

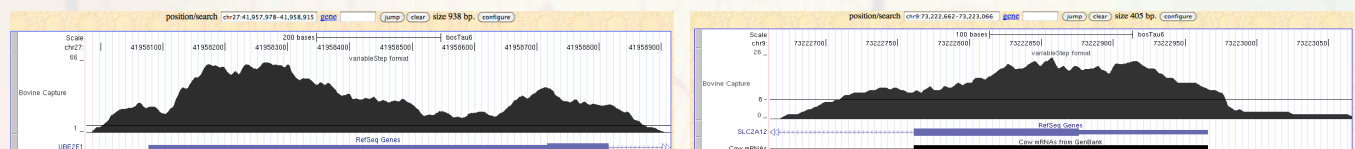
Molecular biases inherent with capture manipulations (locus coverage bias and reduced read specificity; i.e. reads matching “on target”) require oversampling of capture fragments in the sequencing stage to develop the depth of coverage for accurate variant detection. Cofactor Genomics has determined, through our extensive knowledge and expertise in the field of capture sequencing paired with actual data generated from capture experiments, that a target of ~100x or 200x raw coverage is sufficient to cover approximately 75% or 85%, respectively, of the genome at 8x coverage or higher, which is our lower coverage threshold for accurately identifying a heterozygous SNP.

## Results

To determine specificity on the Illumina sequencing platform, we aligned total reads to the bovine reference at the university of Maryland (UMD 3.1) using Novoalign and evaluated the depth and breadth of coverage of target regions. On average, across the three samples at 5 Gb of coverage, 90% of Illumina reads mapped uniquely to the reference genome, of which 75% fell within the target loci. Of the 222,782 genomic targets; 94% had one or more reads aligned, 92% of the target bases were covered by at least one read, 75% of all target bases were covered by at least 8 read, which is considered the minimum coverage to reliably call heterozygous SNPs. Table 1 shows the metrics at each level of sequencing. Homozygous and heterozygous SNPs were called at each level of sequencing depth (Tables 2 and 3).

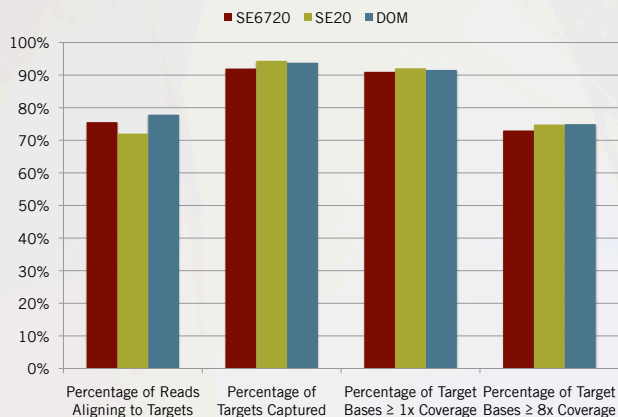
**Table 1. Average Capture Metrics from Bovine HapMap Samples (n=3).**

Sequence Amount (Gb)	Total Number of Reads Analyzed	Raw Coverage of 46 Mb Bovine Exome	Percent of Reads Aligning to Targets	Percentage of Targets Captured	Percentage of Target Bases with $\geq 1x$ Coverage	Percentage of Target Bases with $\geq 8x$ Coverage
2	25,000,000	44	75.67%	90.26%	85.81%	55.30%
3.5	43,750,000	76	75.03%	93.32%	90.32%	69.20%
5	62,500,000	108	74.95%	94.34%	91.83%	74.90%



**Figure 2. Coverage Depth Across Exons from UBE2E1 and SLC2A12 Genes.**

WIG file representation of coverage across 2 target regions of the bovine exome. Grey areas show coverage depth calculated from read alignments and the graphical representation of the exons can be seen in blue below the coverage graph. Numbers at the top of the figure show bases along the bovine genome. Numbers on the left of the figure show minimum and maximum depth of read coverage. The black line represents a depth of 8x coverage.



**Figure 3. Capture Reproducibility**

Three independent bovine samples were captured and sequenced on the Illumina GAIIx. The coverage metrics were computed following sequence alignment of 62.5 million reads per sample.

**Table 2. Homozygous SNPs**

Sequencing Amount	DOM	SE20	SE6720
2 Gb	10,504	47,029	42,408
3.5 Gb	11,695	57,091	52,470
5 Gb	12,141	62,371	57,750

**Table 3. Heterozygous SNPs**

Sequencing Amount	DOM	SE20	SE6720
2 Gb	23,732	40,584	35,893
3.5 Gb	30,290	46,774	42,083
5 Gb	33,911	50,297	45,606

**Tables 2 and 3.** Homozygous and heterozygous SNPs in bovine capture samples. The tables above show the number of homozygous and heterozygous SNPs found in each sample and a certain level of sequence coverage.

### Conclusions

- **75%** of the reads aligned to the target regions.
- **2%** The percent of reads mapped to same targeted regions when sequencing genomic DNA without Bovine Capture.
- **81** The fold enrichment for the targeted gene regions when compared to sequencing the genome.
- Capture method shows high reproducibility between samples.
- **210,004** SNPs, not included on the Illumina BovineHD chip, were discovered.
- Sequence coverage obtained proximal to the target loci that would be missed by PCR-based resequencing
- **12:1** The number of Bovine Captures that can be performed (12) in the same amount of sequence you needed to cover gene regions to an equal depth in (1) whole genome sequence project.
- The bovine capture method provides a robust and highly scalable approach to efficiently capture and sequence nearly every annotated exon from the bovine genome.

### Contact

Quote = <http://cofactorgenomics.com/request-quote>  
 Email = [sales@cofactorgenomics.com](mailto:sales@cofactorgenomics.com)  
 Phone = 314-531-4647  
 Tollfree = 1-888-826-3228